Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by
Dipl.-Phys. Michael Hißmann
born in Paderborn

Oral examination: July 6, 2005

Bayesian Estimation for White Light Interferometry

Referees: Prof. Dr. Fred A. Hamprecht
Prof. Dr. Heinz Horner

Abstract

In this thesis, a new approach for the reconstruction of height maps from scanning white light interferometry is presented. This method unifies the conventional steps of pre- and postprocessing within Bayesian inference. An adept formulation of the prior allows for the exact computation of the height estimate, obviating the need for stochastic sampling or simulation methods.
In conventional surface estimation for white light interferometry, a primary height map is calculated pixel-wise from the raw data, followed by a postprocessing step where outliers and other measurement artifacts are removed. Established and novel algorithms for both steps are discussed. The techniques of Bayesian inference for 2-D image processing, on which the novel surface estimation approach bases, are presented afterwards. For this new method, the localization of the fringe pattern is represented by the likelihood function, while the knowledge about the general surface properties goes into the prior probability of local height configurations. Both the 3-D data set and this prior are considered simultaneously in the estimation procedure, which analytically yields the optimum surface reconstruction as a mode of the marginal posterior probability. A method for quantitative comparison of height maps is developed and used to assess the performance of different postprocessing algorithms.

Zusammenfassung

In dieser Dissertation wird ein neues Verfahren zur Rekonstruktion von Höhenkarten aus der scannenden Weißlicht-Interferometrie vorgestellt, in dem die konventionell nötigen Schritte – Vor- und Nachverarbeitung – in einem Bayes'schen Ansatz verbunden werden. Die Höhenkarte kann hier bei einer geschickten Wahl des Priors direkt berechnet werden, so daß die üblicherweise nötigen Monte Carlo-Methoden entfallen können.
Bei den bekannten Verfahren zur Bestimmung der Oberfläche eines Objekts mithilfe der Weißlicht-Interferometrie wird zunächst pixelweise eine erste Höhenkarte bestimmt, aus der in der Nachverarbeitung Ausreißer und andere Meßartefakte entfernt werden müssen. Zu diesen beiden Schritten werden bekannte und einzelne neue Verfahren diskutiert. Danach werden Bayes'sche Verfahren aus der 2-D Bildverarbeitung vorgestellt, die die Grundlage für das neue Schätzverfahren bilden. Hierbei wird einerseits die Lokalisierung des Interferenzmusters durch eine Likelihood-Funktion eingebracht, andererseits das Vorwissen über die Oberflächengestalt in Form eines lokalen Priors geliefert. Das Verfahren berücksichtigt zugleich den vollen 3-D Datensatz wie auch dieses Vorwissen und bestimmt so eine im Sinne des MPM (maximale lokale Randverteilung) -Schätzers optimale Oberflächenrekonstruktion. Desweiteren wird in der Arbeit die Entwicklung einer zum Vergleichen derartiger Höhenkarten geeigneten quantitativen Methode dargestellt und diese zur Bestimmung der Leistungsfähigkeit verschiedener Nachverarbeitungsverfahren herangezogen.

Contents

LONGUM ITER EST PER PRAECEPTA
BREVE ET EFFICAX PER EXEMPLA

(Seneca)

1. Introduction

Overview In this thesis, we will discuss a new approach for the reconstruction of height maps obtained from scanning white light interferometry, which unifies pre- and postprocessing by Bayesian inference. Compared to conventional approaches, especially for high scanning speeds more accurate results can be achieved.

Industrial image processing Industrial image processing is a field of sustained and expansive growth, now continuing for almost two decades. In the beginning, the possibilities were restricted to very simple tasks, like the detection of the presence of an object, without measurement or identification. But with both the increase in computing power and the development on side of better imaging systems, from video cameras to CCDs and on, the possible applications have become almost countless. Today image processing, still young and sometimes adventurous, has been established as a powerful measurement and testing technology in manufacturing industry.

Out of the many aspects of image processing, the analysis of object surfaces has been gaining of more and more importance, as a scientific interest as well as from side of industrial applications [Rose, 2003]. Surfaces come into focus not only as the primary interface of an object to its environment, i. e. by their form, color or haptics, but also as they can bear specific technical properties, which then can be measured and tested.

In the scope of this thesis, technical surfaces forming mechanical interfaces to other objects are of particular interest. The exact measurement of the surface height as a basis for inference to technical and even functional properties forms the background of our investigations.

As an example, let us look at metallic seals. These are surface structures turned out of a solid piece of metal and used in high-pressure fluid valves. The sealing functionality becomes manifest across a thin ring of e. g. 1 mm width and 20 mm diameter. Flanged to a counterpart, the junction is sealed only when the functional surfaces are planar, smooth and intact. Planar means that no waves, pits, humps or other larger irregularities may come up across it. The smoothness is a mixed requirement: on one hand, the surface must be smooth enough so that no significant leakage may occur, on the other hand it should be so rough to allow for a tight interlock. At last, the seal must be intact, so

no scratches, holes or tips may be present in the surface.

White light interferometry The requirement to automatically test all of these specifications leads to exact specifications of the height measurement device for extended surfaces. The necessary lateral resolution can typically be achieved by a standard CCD-camera with adapted optics. However, the height resolution of 0.1 to 1 μm can only be realized with a light-interferometric approach [Bohn, 2000].

Interferometry used as a measurement tool is surely one of the oldest applications of wave optics, dating back to Michelson's time. However, it has not found its way into industrial application until very recently, as setups better adapted to the rough environment of industrial manufacturing and manageable also for non-specialists are now slowly becoming available.

White light interferometry is here of particular interest, as it fills an important gap by allowing for the measurement of surfaces which are too rough for laser interferometers and too smooth for mechanical testing devices (ball-point testers). As we have experienced, it is a frequent coincidence that surfaces manufactured in this precision range often bear crucial functionality and so are categorically required to be tested after machining.

To fulfill this task, the data obtained from the white light interferometer have to be very reliable. The delivered height map contains the height values calculated for each pixel of the recording CCD-camera. To detect small defects in the surface, this map should be highly reliable, at best down to the level of single pixels.

This is particularly challenging because white light interferometry with rough surfaces is intrinsically error-prone. The reason lies with the physics of reflection, as we will further discuss in the course of this work.

When noisy pixels of the height map, be they single and scattered or in small groups, are detected faulty and assigned a wrong height, the test decision the device has found for a manufactured piece becomes unreliable and debatable. Therefore the height map should be either free of errors, or at least the reliability of each height estimate should be known.

To reduce errors of the height map, postprocessing is applied (in contrast to preprocessing, which is the primary estimation of the height values from the raw data). Here, as we will discuss, the traditional canon of image processing tools can be applied to height maps, augmented by specialized approaches making use of the additional information available with interferometry raw data.

New processing approach In the center of this thesis stands the discussion of a new approach to height estimation for white light interferometry, which is prepared in a Bayesian estimation framework and embodies pre- and postprocessing steps of conventional approaches into one procedure. In preparation of this, we look into Bayesian methods for image processing and reconstruction.

Conventionally, the postprocessing step has only a primary height map and no other information available. Therefore the correction of erroneous pixels can only be based on the neighboring pixels. In the novel approach, the height is

estimated jointly from both the raw data of each pixel and of its neighbors. Therefore an "information bottleneck" is removed and a better estimation is possible. We will discuss both the approach and the performance in greater detail to show what improvement is possible and when which data processing is advantageous.

Finally, we will reason about possibilities for loose correspondences or even links between the last developments, which were up to now focused on white light interferometry, to some (functional) image restauration techniques with an estimation background.

Guide to the thesis In Chap. 2, we start with the discussion of white light interferometry and associated methods of pre- and postprocessing. Next, in Chap. 3 Bayesian methods for image processing are presented. These two major ingredients are brought together in Chap. 4 for the development of the new height estimation approach for white light interferometry. Lastly, in Chap. 5 we look out for possible connections of that approach and methods from image restauration.

2. White light interferometry

Overview In this chapter we briefly describe the physical and technical foundations of white light interferometry. We also point out alternative approaches and recent developments around data processing for white light interferometry.

The first section 2.1 covers the foundations of this measurement principle in optical physics, with emphasis on the particularities when dealing with interferometric inspection of rough surfaces. A more general introduction to optical interferometry can be found in many optics textbooks, cf. [Hecht, 2001] or [Hariharan, 2003].

The next section 2.2 centers around the known data processing approaches for the raw video signal from the interferometer. It is the signal processing aspect that actually distinguishes white light interferometry as a measurement principle from the mere optical phenomenon. The development of white light interferometry went off from different starting points as one can track through the early papers, like [Davidson et al., 1987] (lateral metrology of semiconductors), [Kino and Chim, 1990] (Mirau interference microscope), [Lee and Strand, 1990] (profilometry), [Koch and Ulrich, 1991] (displacement sensor with fiber-optics), and [Dresel et al., 1992] (surface sensing). Until today, the application of white light interferometry in surface metrology and profilometry has come even more into focus of research—at least in part due to the measurement requirements surfacing from the semiconductor and manufacturing industries.

The physical properties of rough and smooth surface reflection are different, therefore different data processing procedures are used. So far research with a background in metrology application has often been focused around fast algorithms from 1-D signal processing, where the data could already be evaluated during the acquisition stage, i. e., “online”.

In the third section 2.3 we discuss current approaches to postprocessing of height maps obtained with white light interferometry. This is an essential step to obtain reliable height maps from interferometric scans of rough surfaces. Due to the lack of phase information and the strong variability of the reflected intensity, there is a certain probability that data points are erroneous or unreliable.

In the fourth section 2.4, after briefly touching mechanical height measurement approaches, we will give a brief outlook on other optical systems as alternatives to white light interferometry of rough surfaces. We discuss their possibilities and drawbacks in comparison and point out some suggestions for further reading.

2.1. Physics of white light interferometry

2.1.1. Measurement principle

Setup A white light interferometry is an optical interferometer which is specialized for spatially resolved height measurement of at least diffusely reflecting surfaces. It is usually built upon a standard Michelson interferometer setup. A broad band light source takes the place of the laser known in the classic setup. To enable the system for surface inspection, a video camera or digital imaging system (CCD array) makes up the detector arm. The surface that is to be inspected takes the place of one of the two mirrors. By use of a mechanical translation stage, the inspected surface is moved perpendicular with respect to the interferometer (cf. Fig. 2.2). During this *scan*, the data for the calculation of the height map is acquired.

The optical focus of the imaging system in the detector arm should be get to the inspected surface for a recording of high spatial resolution of the interference fringes. The scanning procedure however brings in a slight out-of-focus movements of the surface. Excessive blurring can be avoided by considering the depth of focus: The aperture of the objective lens should be limited so that the size of the blur discs stays well below the camera's pixel size for the whole travel length. A further restriction exists for the diameter of the illumination aperture, which limits the achievable lateral resolution due to the granularity of speckle that it is directly correlated to (cf. Fig. 2.1 and Sec. 2.1.2).

Coherence In white light interferometry, the coherence properties of the light source have a significant and limiting influence on the achievable height resolution. *Coherence* is the phenomenon which covers the spatial or temporal correlation of waves. It is widely present and can be observed with all kinds of waves as well as of course with light waves.

With the quantum mechanical dualism of particles and waves in mind, let us consider the light emitted from an arbitrary source as made up from many wavelets or wave trains. A single wave train is characterized by its determinate phase progress and its finite length. The length of emitted wave trains is only determined up to a probability law, which is related to the physical process causing the emission. Therefore a wave trains ends or undergoes a random phase shift after a certain time. The spatial correlation between wave trains is another characteristic of the light source. It can be observed that successive wave trains from neighboring sites of an emitting surface can have a correlated phase. It is therefore customary to differentiate coherence phenomena into *longitudinal (timely)* and *spatial* coherence.

Let us first discuss longitudinal coherence. The average time during which a wave train "exists" is defined as the *coherence time* τ_c. The *coherence length* is the corresponding length, depending on the speed of light c_n in its propagation medium with refractive index n: $l_c = c_n \tau_c$. Phenomenologically, the coherence length gives answer to the question how far two points can be apart along the direction of propagation, while the phase of the signal at one point can still

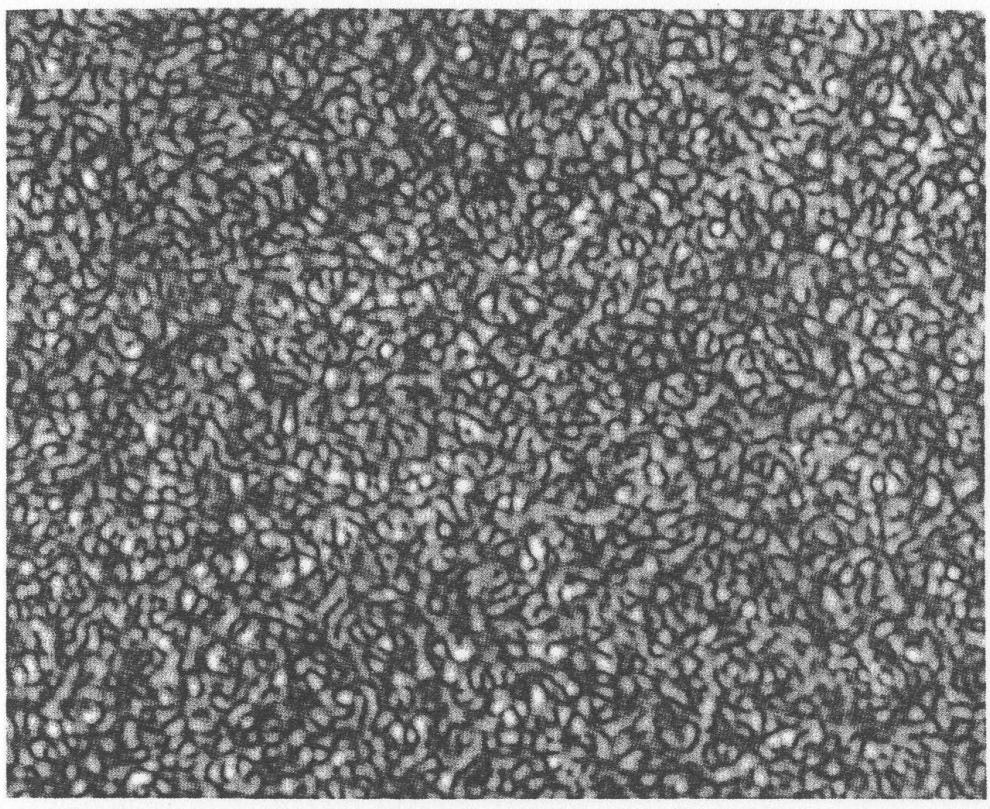


Figure 2.1: Speckle—granular coherence phenomenon observed with laser illumination on a rough surface (from [Dainty, 1984], with kind permission of the author and the publisher).

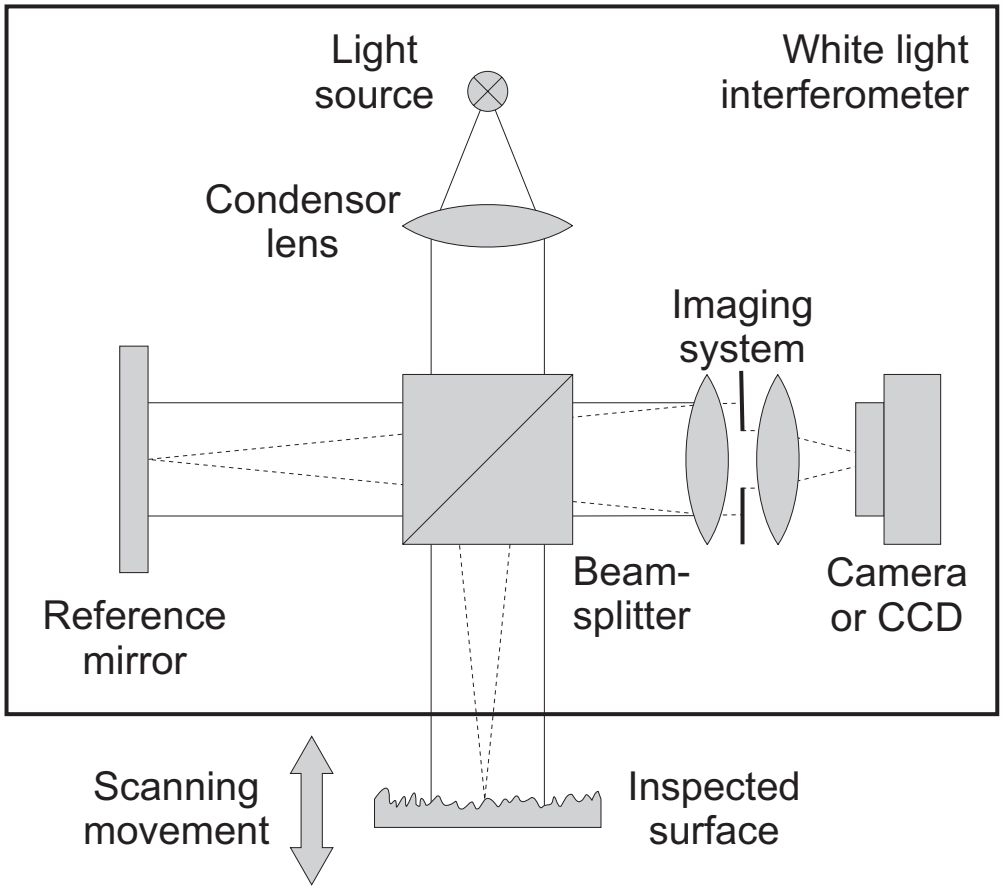


Figure 2.2: Optical setup for white light interferometry.

be determined from a measurement at the other point. A wave train finally ceases to "exist" when its phase progress undergoes a stochastic shift. The frequency bandwidth of such a signal is given by $\Delta\nu = 1/\tau_c$. This inverse proportionality can be understood from the properties of the Fourier transform or the Heisenberg uncertainty principle. It becomes clear that "incoherence" cannot exist in a strict sense, for it would require a source of infinite bandwidth. Instead, the notion of "incoherent light sources", which covers thermal source (e. g., a light bulb) or the sun, refers to their very short coherence length. With these sources, subtle coherence phenomena still exist and can be made visible [McKechnie, 1976], and interferometry is possible as well.

The concept of *spatial coherence* covers the characteristics of the emission resulting from the spatial extent of an area of emission, which can be a direct light source or, quite common, a virtual source formed by imaging a real light source. The emission of a thermal light source is a handy example of low spatial coherence emission: Due to spontaneous relaxation, each site of the emitter (that could be a surface atom in a solid emitter) shows photonic emission independent of other sites. The length of each wave train corresponds to the relaxation time. It is directly clear that any two centers of emission that are farther apart than about the length of the wave train cannot influence each other. Any two wave trains emitted at the same time will be of arbitrary, random phase difference, thus such a light source is spatially incoherent, beyond the diffraction limit. With the same argument as in the case of longitudinal coherence, nothing can be said from measuring the phase of one wave train about the phase of the other wave train.

A common approach to create a source of defined spatial coherence is to project a diffraction-limited light source onto a larger area: The resulting image is spatially coherent to the extent given by the original source's longitudinal coherence length—the imaging system "transforms" longitudinal into spatial coherence.

With the discussions in this paragraph we have seen that longitudinal and spatial coherence can both be brought back to the question whether the light waves measured in two points in space are in correlation of each other. Equally, both effects can be demonstrated and measured with Young's classical double-slit experiment and should therefore only be seen as aspects of the wave properties [Hecht, 2001].

Optical interference We start the theoretical discussion with the most simple case, namely the interference of perfectly monochromatic, planar waves. This means we assume infinite longitudinal coherence and additionally consider the problem translationally indifferent, that means we ignore any spatial effects.

In practice, a reasonable approximation to this idealization can be found with the interferometry within a single speckle, i. e. the electromagnetic field in a spatial volume limited by its longitudinal and spatial coherence length (see Sec. 2.1.2).

For our discussion, let us assume a linear homogeneous isotropic optical medium,

$$\varepsilon = \varepsilon_r \varepsilon_0 \qquad \mu = \mu_r \mu_0 \qquad c = \frac{1}{\sqrt{\varepsilon \mu}} \tag{2.1}$$

Under these settings, the basis solutions of Maxwell's equations are simple planar waves [Born and Wolf, 1999]. It is convenient to use the complex-valued analytic signal notation and keep in mind that only its real part is physically relevant:

$$\boldsymbol{E}(\boldsymbol{r},t) = \boldsymbol{E_0}\, e^{-i(\boldsymbol{kr}-\omega t)} \qquad \boldsymbol{H}(\boldsymbol{r},t) = \boldsymbol{H_0}\, e^{-i(\boldsymbol{kr}-\omega t)} \text{ with } \boldsymbol{H_0} = \frac{\boldsymbol{k}\times\boldsymbol{E_0}}{\omega\mu} \tag{2.2}$$

We define the intensity by the absolute value of the Poynting vector $\boldsymbol{S}$:

$$I = |\boldsymbol{S}| = |\boldsymbol{E}\times\boldsymbol{H}| \tag{2.3}$$

In the linear case it becomes

$$I = \sqrt{\frac{\varepsilon}{\mu}}\,|\boldsymbol{E}|^2 \tag{2.4}$$

In an interferometer setup, the optical waves from the two arms are recombined in the beamsplitter, which is expressed by summation of the respective fields:

$$\boldsymbol{E} = \boldsymbol{E_1} + \boldsymbol{E_2} \tag{2.5}$$

The optical path length difference $\Delta\boldsymbol{r}$ between the two arms leads to a residual phase shift between $\boldsymbol{E_1}$ and $\boldsymbol{E_2}$:

$$\begin{aligned} \boldsymbol{E_2}(\boldsymbol{r},t) &= \boldsymbol{E_1}(\boldsymbol{r}+\Delta\boldsymbol{r},t) && (2.6)\\ &= \boldsymbol{E_{1,0}} e^{-i(\boldsymbol{k}(\boldsymbol{r}+\Delta\boldsymbol{r})-\omega t)} = \boldsymbol{E_{1,0}} e^{-i(\boldsymbol{kr}-\omega t)} e^{-i(\boldsymbol{k}\Delta\boldsymbol{r})} && (2.7)\\ &= \boldsymbol{E_1}(\boldsymbol{r},t) e^{-i(\boldsymbol{k}\Delta\boldsymbol{r})} && (2.8) \end{aligned}$$

We assign the net effect of this as a phase shift $\varphi = -\boldsymbol{k}\Delta\boldsymbol{r}$ to the $\boldsymbol{E_2}$ arm of the interferometer.

At this point, it already becomes clear that longitudinal coherence is a mandatory criterion for stable interference: If the phase difference is larger than about the coherence length, most probably wave trains of different random phase shifts come into interference. Thus the phase shift φ of the resulting signal will also be random.

For simplicity, we consider our setup free of dispersion, thus we assume φ is linear in frequency. Experience with our laboratory white light interferometer setups however has shown that already the dispersion errors caused by propagation through thin ($\approx 1\,\mathrm{mm}$) neutral density filters in one arm are easily visible and significantly deteriorate the signal. The dispersion must therefore be compensated, e. g. by a glass of equivalent optical properties in the other arm. A further effect is the dampening of the electromagnetic fields originates from optical asymmetry of the beamsplitter and differences in reflective properties

of the reference and object surfaces in the two arms. We can try to compensate this by introducing dampening factors $\alpha \leq 1$, $\beta \leq 1$, which however is still only a first-order approximation. Also cf. [Pförtner and Schwider, 2001] for influences of optical element imperfections.

With these approximations, we write the resulting field after the beamsplitter as follows:

$$\boldsymbol{E'} = \alpha \boldsymbol{E_1} + \beta \boldsymbol{E_1} e^{i\varphi} \tag{2.9}$$

The detector array in the CCD camera cannot follow the time-domain oscillation of the interference field $\boldsymbol{E'}$, but outputs an electric signal proportional (within the limits of the converting electronics) to the intensity of the interference field $\boldsymbol{E'}$. To calculate the intensity according to Eq. (2.4), we make use of $|\boldsymbol{E}|^2 = (\boldsymbol{E}\, e^{i\varphi})(\boldsymbol{E}^*\, e^{-i\varphi})$ and get:

$$\begin{aligned} I_{\mathrm{CCD}} &= \langle |\boldsymbol{E'}|^2 \rangle & (2.10)\\ &= \langle |\alpha \boldsymbol{E_1}|^2 \rangle + \langle |\beta \boldsymbol{E_1}|^2 \rangle + \langle |\alpha\beta \boldsymbol{E_1}\boldsymbol{E_1}(e^{-i\varphi} + e^{i\varphi})| \rangle & (2.11)\\ &= \langle |\alpha \boldsymbol{E_1}|^2 \rangle + \langle |\beta \boldsymbol{E_1}|^2 \rangle + \langle |2\alpha\beta \boldsymbol{E_1}\boldsymbol{E_1} \cos\varphi| \rangle. & (2.12) \end{aligned}$$

We gather up the sum of the first two terms to the *fundamental intensity* I_0, while the last one becomes the *interference amplitude* I_1:

$$I_{\mathrm{CCD}} = I_0 + I_1 \cos\varphi \tag{2.13}$$

Thus the first summand of Eq. (2.13) describes the intensity as measured without interference effects. This is modulated by the second term with an amplitude I_1. The two contributions I_1 and I_2 do not need to be equally large. Quite the contrary, with white light interferometers I_2 is often much smaller than I_1, i. e., the interferometer is not symmetric in the light paths of its arms, cf. Eq. (2.12). Therefore the output signal of the interferometer does not drop to zero for fully destructive interference, nor does it double for constructive interference.

The *visibility* or *interference contrast* is defined by the following ratio of maximum and minimum intensity:

$$\mathcal{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{2.14}$$

For the interference modulation in Eq. (2.13), this becomes simply $\mathcal{V} = I_1/I_0$. For a symmetric interferometer ($\alpha = \beta$ in Eq. (2.12)), one can reach visibility $\mathcal{V} = 1$.

If the light arriving at the detector is not fully coherent, which is usually due to the properties of the light source or due to diffusely scattered (stray) light from rough surfaces, the incoherent part can be seen as another contribution to I_0. Therefore sometimes the—slightly misleading—designations "coherent" and "incoherent" for the two contributions of Eq. (2.13) can be found in literature with a more empirical approach.

Scanning procedure The first stage of the white light interferometric measurement process is the systematic variation of the phase of the interference signal. This is achieved by a controlled motion of the measurement setup perpendicular to the surface under inspection, during which the camera continuously acquires an image sequence, cf. the movement arrow in Fig. 2.2. The extension of one interferometer arm by Δz leads to a phase shift $\varphi = -2k\Delta z$.

A more or less precise correspondence between the positions of the moving translation stage and the frame numbers exists—for some setups the frame acquisition is even triggered by the stage movement. Based on this, each frame number is assigned a discrete height value, ranging from 1 to $h_{\max}$. The scale is found by a calibration measurement or is known from the transmission ratio of the translation stage used. The height of a pixel is set to that frame's height value, at which an algorithm (see Sec. 2.2) determines the center of interference.

Errors in the stage movement like speed variations or slip-stick-effects deteriorate the frame-to-height correspondence. Investigations on this have been described in [Schraud, 2000] and [Körber, 2004]. However, for the measurement of rough surfaces, which is the main focus of this thesis, these effects can be ignored as they are usually much smaller than the statistical uncertainty of the height map (cf. Sec. 2.1.4).

In a white light interferometer setup, a translation stage with either a continuous or step-wise motion mode can be integrated. With continuous motion and cameras with an asynchronous shutter, a systematic skew will be introduced by the movement during frame acquisition. Although it could be another source for impairment of the data quality, for our setup and field of application we have found that this issue can safely be ignored.

Influence of the light source Interferometry is generally possible with a variety of different light sources, like lasers, arc lamps and thermal light sources. We have seen that for a Michelson-type interferometer setup, the longitudinal coherence length is a central parameter.

The emission of a light source can be characterized by its spectrum, i. e. the distribution of the emission over different wavelengths, $dI/d\lambda$. The coherence length is generally inverse proportional to the spectral "width", i. e. the broader the spectrum, the shorter the coherence length.

In this paragraph we briefly discuss the use of some common light sources for interferometry.

Laser source In the beginning of this chapter, we have theoretically discussed interferometry for a single wavelength, which leads to a cosine-like intensity modulation, Eq. (2.13), in the detector. This case is however not achievable in practice, but is merely an idealization—a single wavelength would correspond to an infinite coherence length. Such a light source would have an arbitrarily low bandwidth and its emission reduced to a delta-peak spectrum. A quite fair approximation can however be found with laser sources. A stabilized He-Ne gas laser running on a single longitudinal mode has a coherence length of roughly about 300 m [Saleh and Teich, 1991], [Bergmann and Schaefer, 2004]. Such a

quasi-monochromatic laser source is a good approximation to the idealized case of a purely monochromatic source as described in last paragraphs. The interference modulation of the laser's signal is almost exactly as given in Eq. (2.13), with very good periodicity. The large coherence length reduces the interference amplitude when the path differences reach the order of magnitude of the coherence length. For practical purposes, the interference signal can be considered strictly periodic. This restricts the range of uniqueness Δz for the detection of phase differences to the period length; that means, only phase shifts smaller than 2π, or $\Delta z < \frac{\lambda}{2}$ measured. This restriction can be a significant drawback when using a laser of high coherence length in an height measuring interferometer (cf. Sec. 2.4 for additional discussion of *laser interferometer*).

Heterodyne source A longer range of uniqueness for the interference signal is generally desirable for height measurement. One approach is to introduce a signal modulation, known as *heterodyne principle*: Mixing of two or more periodic signals leads to a signal which has a larger period length than both of the original signals. For the application in an interferometer, one can combine two laser beams of slightly different wavelength with a beamsplitter. The summed signal (known as *beat* in physical acoustics) of two sinusoidal with wavenumbers $k = \frac{2\pi}{\lambda}$ and $k' = k + \Delta k$ has a longer periodicity, as one finds by simple mathematics:

$$\sin(kz) + \sin((k+\Delta k)z) = 2\sin\frac{(kz + (k+\Delta k)z}{2}\ \cos\frac{kz - (k+\Delta k)z}{2} \tag{2.15}$$

$$= \sin\left(\bar{k}z\right)\cos\left(\frac{\Delta k}{2}z\right) \quad \text{with } \bar{k} = k + \frac{\Delta k}{2} \tag{2.16}$$

The result in Eq. (2.16) has a high-frequency part, oscillating at the mean wavenumber $\bar{k}$, which is modulated with a slow oscillation at $K = \Delta k/2$, the *synthetic* or *heterodyne* wavenumber of this interferometer setup, cf. Fig. 2.3. The periodicity of such an interferometer setup is therefore $2\pi/2\Delta k$, which gives a synthetic wavelength that can be calculated like $\Lambda^{-1} = \left|\lambda^{-1} - \lambda'^{-1}\right|$. Half of this wavelength is the range of uniqueness of the *heterodyne interferometer*.

Interferometry with a broadband source A broadband light source is any source which features a spectrum not consisting only of line emission, i. e. a "broad" spectrum. In particular, this includes thermal light sources and light-emitting diodes (LEDs). Phenomenologically, the broad emission can be seen as the superposition of very many signals of only slightly different wavelengths. Consequently, interferometry with broadband light sources can be seen as a step from two-wavelength (heterodyne) to multi-wavelength interferometry, or *white-light interferometry*. The latter name became established even though in many cases the light source is anything but white, for example an infrared LED. For a mathematical description, correlation properties form a convenient approach:

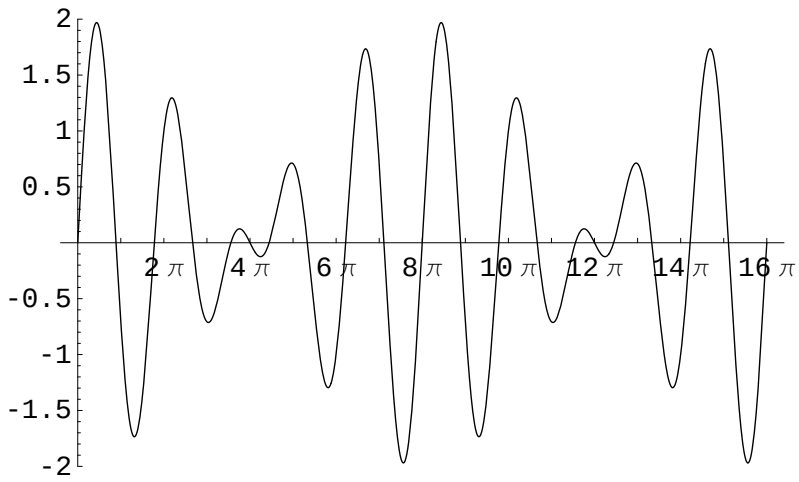


Figure 2.3: Illustration of a heterodyne oscillation, wavenumbers $k_1 = 1$ ($\lambda_1 = 2\pi$) and $k_2 = 5/4$ ($\lambda_2 = 8/5\pi$). Beat wavenumber is $K = 1/8$ and heterodyne wavelength $\Lambda = 8\pi$.

We look again at the two-beam interference. The optical path length difference corresponds to to a time lag for $\boldsymbol{E_2}$:

$$\boldsymbol{E_2}(\boldsymbol{r}, t) = \boldsymbol{E_1}(\boldsymbol{r}, t + \tau) \tag{2.17}$$

The interference term for real valued fields is then (cf. Eq. (2.12)):

$$I_1 \propto \langle \boldsymbol{E_1}(t)\boldsymbol{E_1}(t + \tau)\rangle \tag{2.18}$$

The mean value $\langle \dots \rangle$ is found by integration (the integration time should last over at least one period of the light wave—in practice the inertia of any current detector causes much longer integration times):

$$I_1 \propto \int_0^T \boldsymbol{E_1}(t)\boldsymbol{E_1}(t + \tau)\, d\tau \tag{2.19}$$

which we can identify as the auto-correlate c_{EE} of $\boldsymbol{E_1}$, if we accept the approximation

$$\int_0^T \dots d\tau \approx \lim_{T\to\infty} \int_0^T \dots d\tau, \tag{2.20}$$

i. e., if the integration covers the time during which the interference occurs, which is the case if T is sufficiently large against the correlation length l_c.

Using the Wiener-Khintchine theorem [Hecht, 2001], we can now form a mathematical link between the properties of a broadband light source and its interferogram. This theorem is an application of the Parseval theorem for signals $f(t)$ and $f(t + \tau)$ [Moon and Stirling, 2000]. In our case:

$$\mathcal{F}\{c_{EE}(\tau)\} = |\mathcal{F}\{E(t)\}|^2 \quad \left(\text{or } = |F(\omega)|^2\right) \tag{2.21}$$

$|F(\omega)|^2$ is the power spectrum (spectral energy distribution) of the electrical field of the light source. Therefore, by simple Fourier transformation of the spectrum one can calculate the outer form of the interferogram of an arbitrary light source.

In case of a highly coherent laser, the interferogram is a perfect sinusoid, as we have seen in an earlier paragraph. In the white light interferometer setup we use for our experiments, a near-infrared LED is used, which approximately has a Gaussian spectrum. Therefore, the interferogram here also has a Gaussian envelope, but with inverse width. Other common setups for white light interferometers use ordinary light bulbs: A thermal source can be approximated reasonably with a wide Gaussian around its peak, and the corresponding interferogram will have a rather tall Gaussian envelope. This form of the interferogram allows for a more precise analysis, as the envelope is steeper, but such a signal is also more difficult to detect within harsh noise. We will discuss more of the analysis of white light interferograms in detail in Sec. 2.2.

Modifications to the setup A high speed of data acquisition and an increased precision are two frequent, yet often contradictory aims for the development of interferometer setups. This thesis covers some possibilities from the data processing side. With changes in the technical setup however, ways to significant improvements especially in terms of acquisition speed and robustness against mechanical flaws of the setup, can be paved. We refer to the dissertation [Seiffert, 2005] for a discussion of the most recent developments, like the use of color cameras or augmentation of the Michelson-type white light interferometer setup by an additional laser interferometer.

Besides the classic Michelson-type setup (cf. Fig. 2.2), further variations have been devised which are applicable for more specialized applications. We take note of the following to show the breadth of development:

- White light interferometry with specialized microscope optics for conic surfaces [de Groot and Colonna de Lega, 2003], or a setup for inner surfaces hollow tubular structures [Aziz, 1998].

- A divergent beam setup, which allows for the measurement of non-perpendicularly polished, highly reflecting surfaces [Ammon et al., 1997].

- A non-perpendicular intrinsic movement procedure to enable the measurement of very elongated objects by longitudinal scanning [Restle, 2003].

- Variants of the interferometer's optical design to improve the zoom range [Windecker et al., 1999] and a Linnek-type setup for the microscope range [Windecker and Tiziani, 1999].

- Approaches how to control shifts of the optical path and to provide an intrinsic calibration [Olszak and Schmit, 2003].

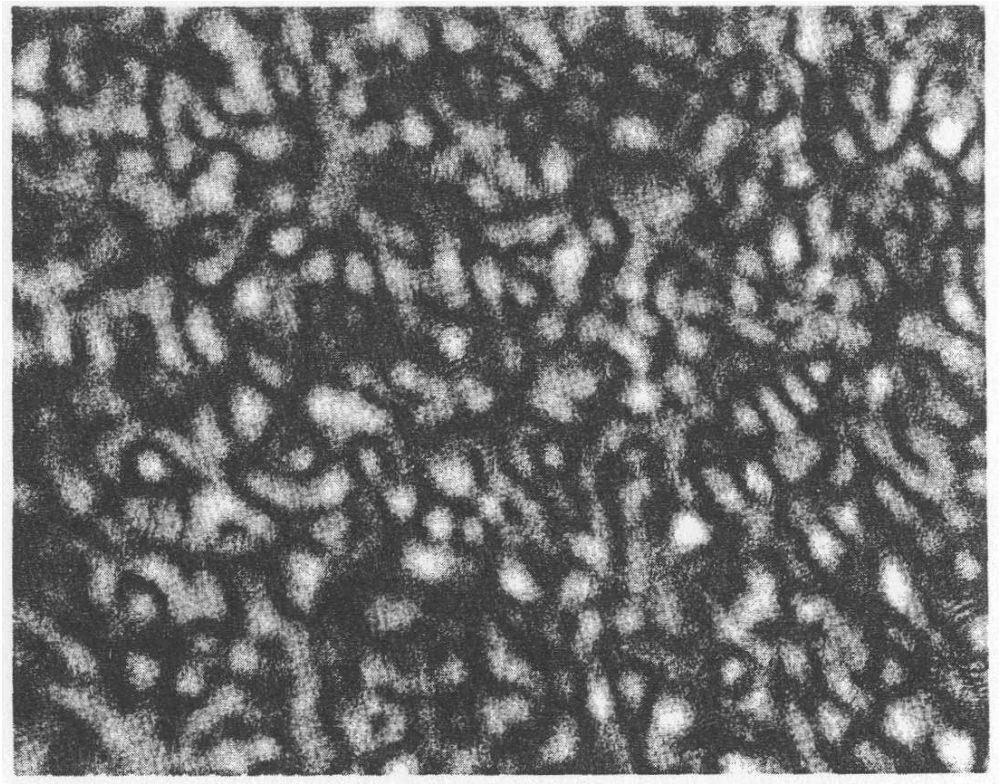
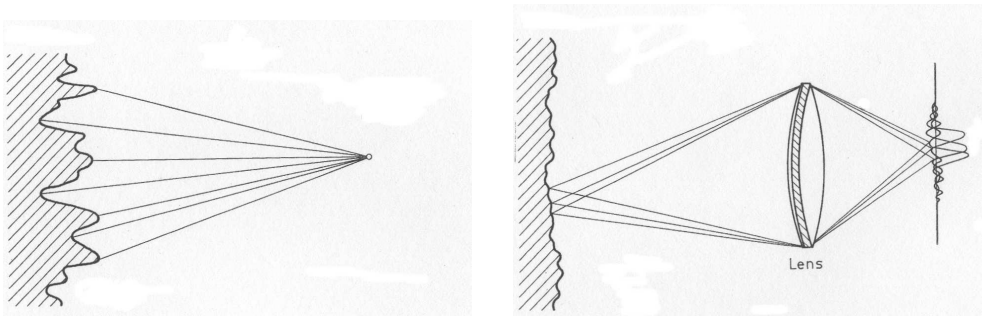


Figure 2.4: Speckle as observed under laser illumination of a rough surface (from [Dainty, 1984], with kind permission of the author and the publisher).

2.1.2. Speckle

Speckle are a common phenomenon, which can be found with almost any system dealing with waves. It is even present in daily life, as the apertures of the optical system *sun* → *rough surface reflection* → *human eye* fulfills the necessary prerequisites, but often goes unnoticed there. If at least partial coherence between wave packets in a spatio-temporal neighborhood exists, speckle can be observed. Examples can be found with optical, radio and even acoustical waves, they are of practical use in speckle interferometry (astronomy), speckle holography and others.

Still, speckle phenomena are best known and easiest to experience with laser sources: Most surfaces, when illuminated with a laser source, expose a granular appearance (cf. Fig. 2.4). More precisely, if the height of a reflective surface varies by more than the scale of the illumination wavelength, and the coherence length of the light source is smaller than this variation, speckle can be observed. In this case, the interference condition is fulfilled, therefore brightness variations due to constructive and destructive interference occur. The height variations of a rough surface lead to interference of light scattered from many micro-facets of this surface, already on a microscopic lateral scale. Due



(a) (b)

Figure 2.5: Situations in with (a) objective and (b) subjective speckle can be observed under coherent illumination (from [Dainty, 1984], with kind permission of the author and the publisher).

to diffraction a granular pattern of larger spatial extent, the speckles, can be observed.

Subjective and objective speckle Speckle phenomena in optical systems are often differentiated into *objective* and *subjective* speckle, depending on the relationship between scatterers, apertures and observer. Fig. 2.5 shows exemplary situations in which both types can be observed.

Objective speckle arise after free propagation of coherent waves scattered at a rough surface. The speckle pattern does only depend on the surface and thus is independent of the individual observer. This explains its name "static" or "objective speckle".

Subjective speckle arise when light is scattered from a rough surface and observed after propagation through an imaging system. The aperture of this system limits the resolution that can be achieved when observing the surface. At the same time, coherent interference is only possible for light reflected from an area within this aperture. Therefore the size d of speckles is determined by the imaging aperture [Bohn, 2000], [Goodman, 1984]:

$$d \sin \theta_{\text{obs}} \approx \frac{\lambda}{2} \tag{2.22}$$

2.1.3. Reflective properties of rough surfaces

Roughness Although the meaning of the term "roughness" seems to be clear from everyday life, nevertheless quite a number of definitions can be found. All the more when investigating surfaces of complex shape, it is not clear how to discriminate roughness from other surface features such as waviness or the overall surface profile. Even definitions that are written down in normative

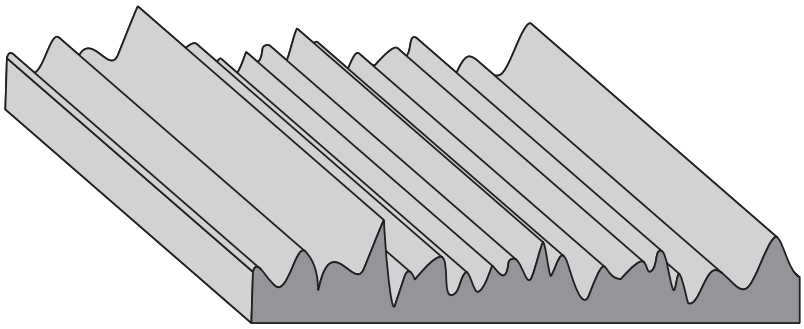


Figure 2.6: Sketch of a typical roughness reference standard used for calibration of mechanical roughness testers. Although the standard is an extensive piece of metal, the test structures are engraved only in one direction in its reference surface.

documents can appear somewhat ad hoc (cf. [DIN, 1990] for German standards concerning surface description and measurement procedures). The measurement procedures defined in this document are well adapted to the requirements of tactile mechanical testers that measure only along a line of the surface. The procedures come into difficulties with the requirements and possibilities of optical measurement devices with spatial resolution. A good example are roughness standards, which are used to calibrate mechanical roughness testers. As can be seen in Fig. 2.6, the surface of such a standard shows statistical roughness only in one dimension, while it does not vary at all along the other. This type of standard cannot live up to the power of a spatially resolved surface measurement device. We here observe a certain disequilibrium between what is defined and what can be measured. Some discussions on how this gap in the tool chain could be bridged can be found in the dissertation [Eberle, 2005] and the references cited therein.

From the optics point of view, roughness cannot merely be seen as a surface property, but has to be considered in conjunction with the surface's influence on the light reflection process. A widely accepted definition considers a surface *optically rough*, if its height variation within a coherently illuminated area is larger than the illumination's mean wavelength [Fercher et al., 1985]. While this property is easily accessible in experiments, it cannot hold for a definition of surface roughness: An optical measurement can qualify a surface both rough *and* smooth, only depending on the wavelength and the spatial coherence of the illumination used in the measurement:

$$\text{“rough”:} \quad \Delta h_{\text{speckle}} > \lambda_{\text{mean}} \tag{2.23}$$

We cannot offer a solution for this complex problem. But as we do not need a tight definition in this work, we consider roughness under the optics point of view, taking an illumination of small bandwidth at $\approx 800\,\text{nm}$ wavelength as a reference.

2.1.4. Statistics of rough-surface reflection

The discussion of rough-surface reflection is quite complicated if one tries to calculate the interaction of real surfaces and illumination from real light sources. One then has to revert to simulation methods, which however still rely upon some simplifications (cf. end of this section).

For the case of perfect coherence and roughness, as well as a statistically "large" number of scattering centers, also ignoring any polarization issues, a rather easy analytic calculation of the statistical law of reflected intensity exists. We recite the central points laid out in [Goodman, 1984] in the next paragraph.

Random walk model In the random walk model, the reflecting rough surface is considered to be made up from numerous planar facets ("micro-facets") of different size and orientation. Each facet is a unique scatterer: it is flat, and reflects a random fraction of the incident light into a random direction. The light field after reflection is made up from contributions of these scatterers, of different intensity and phase across the whole surface. Each facet and its contribution is named an *elementary phasor*. For the spatial area and path differences which allow coherent superposition, i. e. the speckle extent in time and space, the electrical fields interfere and contribute to I_1. On the other side, light that is scattered under higher angles, from other regions of the surface and with path lengths differing by more than the coherence length, is added up to the intensity I_0.

This approach is both valid for objective and subjective speckle (cf. Sec. 2.1.2 and Fig. 2.5). For any kind of speckle, the amplitude of the electric field after reflection is made up from contributions of a large number of phasors. They are located in different regions of the scattering surface, therefore we assume that the phase of their contribution is arbitrary:

$$E = \sum_{i=1}^{N} \frac{1}{\sqrt{N}} E_i = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} |E_i| e^{i\varphi_i} \tag{2.24}$$

The name "random walk model" originally describes a statistical model, in which (in its simplest form) at each time step a particle makes a random movement by one step in either possible direction. If one looks at the distance the particle moved after some time, one finds it obeys a Gaussian probability distribution and so the model actually describes a diffusion process. This is very similar to the summation of the electrical field vectors in Eq. (2.24) which its contributions of different length and direction, hence the name.

We now go on with first looking at the situation arising from a single speckle, with the reflection measured in a single point above the surface (the first-order properties). In addition, the calculation (as detailed in [Goodman, 1984]) requires validity of the theorem of large numbers, which means we have to assume contributions of very many (not necessarily independent) phasors within a speckle. This last postulate has been proven sufficient for most practical cases [Goodman, 1984] and simulations based on it coincide with measurements [Ettl, 2001].

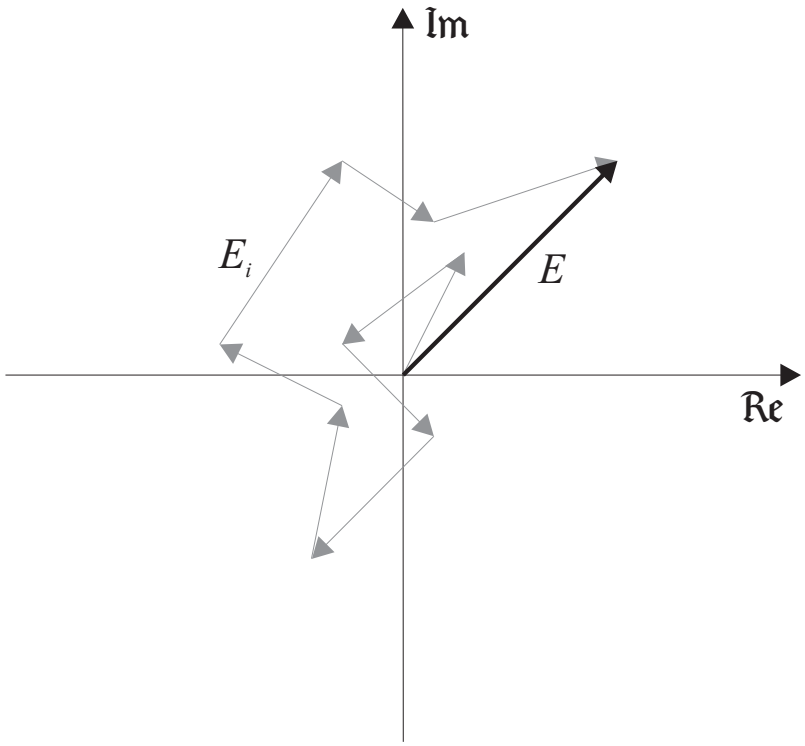


Figure 2.7: Random walk model: Random contributions for the scattered amplitude (after [Goodman, 1984]).

In this model, the following assumptions about the reflective properties of the elementary phasors have to be accepted:

- The amplitude $|E_i|$ and phase φ_i of each unique scatterer should be statistically independent of each other:

  $$p(|E_i|, \varphi_i) = p(|E_i|)\, p(\varphi_i) \tag{2.25}$$

  Furthermore, they should be independent of any other scatterer's amplitude or phase. We will discuss the consequences of this restriction in the course of this paragraph.

- The phases φ_i should be uniformly distributed on the interval $[0, 2\pi]$:

  $$p(\varphi_i) = \mathcal{U}(0, 2\pi) \tag{2.26}$$

  This requirement can be fulfilled best if the surface's height differences (or, more precisely, the phase shifts that arise thereupon) are so large that they exceed the 2π-interval for the interferometer's wavelength significantly. That is, the surface height should vary by at least $\pm\lambda/4$.

Note that these requirements restrict the admissable size of a unique scatterer: On one hand, it must be so small as to ensure independence from other scatterers, but on the other hand, it must be so large as to cover an area of height differences sufficient for a uniform phase distribution.

We now can calculate the statistical law for the real and imaginary part of the reflected field's amplitude. We immediately find:

$$\mathrm{Re}\{E_i\} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} |E_i| \cos\varphi_i \tag{2.27}$$

$$\mathrm{Im}\{E_i\} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} |E_i| \sin\varphi_i \tag{2.28}$$

One sees that the averages (first order moments) of both parts are zero:

$$\langle \mathrm{Re}\{E_i\} \rangle = \frac{1}{\sqrt{N}} \sum_{i=1} N \langle |E_i| \cos\varphi_i \rangle \tag{2.29}$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1} N \langle |E_i| \rangle \langle \cos\varphi_i \rangle \quad \text{by assumption Eq. (2.25)} \tag{2.30}$$

$$= 0 \quad \text{by assumption Eq. (2.26)} \tag{2.31}$$

and equally we find:

$$\langle \mathrm{Im}\{E_i\} \rangle = 0 \tag{2.32}$$

The same way the second moments can be calculated. For the variances one gets:

$$\langle \mathrm{Re}\{E\}^2 \rangle = \langle \mathrm{Im}\{E\}^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} \frac{\langle |E_i|^2 \rangle}{2} \tag{2.33}$$

And similarly, it can be calculated that the covariance is zero:

$$\langle \mathrm{Re}\{E\}\, \mathrm{Im}\{E\} \rangle = 0 \tag{2.34}$$

The reflected electrical field is added up from contributions of many small surface facets within one speckle (in the formulae, this corresponds to the transition $N \to \infty$). We have already discussed that the law of large numbers holds true. So the joint probability density function of real and imaginary part must be Gaussian:

$$p_E(\mathrm{Re}\{E\}, \mathrm{Im}\{E\}) = \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{\mathrm{Re}\{E\}^2 + \mathrm{Im}\{E\}^2}{2\sigma^2} \right\} \tag{2.35}$$

The variance σ^2 is, with the results from Eq. (2.33) and (2.34), found to be

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{\langle |E_i| \rangle}{2} \tag{2.36}$$

From Eq. (2.35), one can derive the probability density functions for the directly accessible variables, namely amplitude and phase. We make use of the

transformation theorem for probability densities, which states for two density functions p_X to p_Y for the transition (cf. [Moon and Stirling, 2000]):

$$p_Y(y) = p_X(g^{-1}(y))\,||\boldsymbol{J}(g)|| \quad \text{with the Jacobian } J_{i,j}(g) = \left(\frac{\partial g_i}{\partial y_j}\right)_{i,j} \tag{2.37}$$

In our case, the relationship

$$I = |\mathrm{Re}\{E_i\}|^2 + |\mathrm{Im}\{E_i\}|^2 \tag{2.38}$$

$$\tan\varphi = \frac{\mathrm{Im}\{E_i\}}{\mathrm{Re}\{E_i\}} \tag{2.39}$$

is inverted, and for the Jacobian's determinant one finds $||\boldsymbol{J}|| = \frac{1}{2}$. Then we get:

$$p_{I,\varphi}(I,\varphi) = \frac{1}{4\pi\sigma^2}\exp\left\{-\frac{I}{2\sigma^2}\right\} \tag{2.40}$$

By marginalization, one finds:

$$p_I(I) = \int_{-\pi}^{\pi} p_{I,\varphi}(I,\varphi)\,d\varphi = \frac{1}{2\sigma^2}\exp\left\{-\frac{I}{2\sigma^2}\right\} \tag{2.41}$$

$$p_\varphi(\varphi) = \int_0^{\infty} p_{I,\varphi}(I,\varphi)\,dI = \frac{1}{2\pi}, \tag{2.42}$$

each with the restriction $I \geq 0$ and $-\pi \leq \varphi < \pi$.

This result shows that the probability density for the intensity follows an exponential law, so the probability to observe a certain intensity becomes exponentially smaller the larger the intensity is (cf. plot for $c_{12} = 1$ in Fig. 2.8). For the phase we obtained a constant probability density functions, so any phase is equally probable.

In experimental setups, the ideal conditions founding this theory are not always met:

- The scattering surface may depolarize the light, which is therefore "lost" for the above calculations where we take only one polarization into account. For a correct result, the two polarizations involved must be calculated separately. However, if the detector is blind to different polarizations, the above results should still remain valid, as long as depolarization effects do not remove the correlation.

- While the reflected intensity follows an exponential distribution, the signals obtained from bright speckles tend to be more stable than those from darker speckles, which however dominate the image. In addition, for those dark speckles external, typically mechanical disturbances to the interferometer setup lead to relatively larger intensity fluctuates [Ettl, 2001]. The reason is that for these speckles, the intensity integrated over many microfacets is mostly compensated, thus singular additional contributions have a large impact. With bright speckle, such extra intensity is small compared to the overall signal level and has little influence.

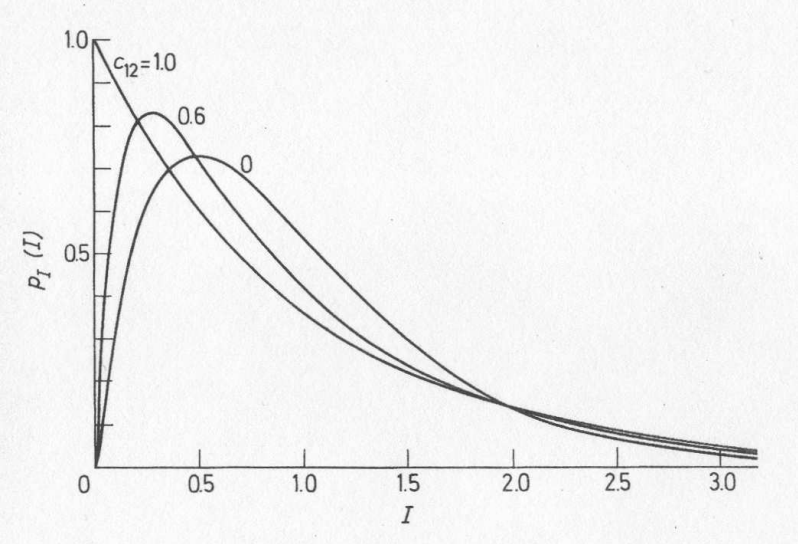


Figure 2.8: Incoherent sum of two speckles of equal mean intensities: probability distribution of the reflected intensity (from [Goodman, 1984], with kind permission of the author and the publisher).
The parameter c_{12} describes the intensity correlation coefficient: technically, $c_{12} = 1$ corresponds to equal σ's as we have discussed (Eq. (2.35)), $c_{12} < 1$ to lower correlation, as reflected in a larger degeneracy of the coherence matrix $\boldsymbol{J}$ (Eq. (2.37)).

- If additional background illumination is present, the results can change. Incoherent illumination simply adds an offset to the intensity distribution, but the phase distribution remains unchanged. If the background however is coherent, the resulting intensity distribution reflects this contribution by broadening up from the exponential form. Accordingly, the contribution shows up in the phase statistics as an additional peak around the phase of the background signal (cf. [Goodman, 1984] for further details).

  This scenario is also an approximate description for surfaces which are neither ideally rough, nor smooth (as an example, cf. Fig. 4.2). These surfaces come up frequently in industrial high-precision manufacturing, and often exhibit a rough surface with leveled regions. The smoothed areas account for the coherent part of the reflection, which broadens up the exponential intensity distribution from the rough parts of the surface (cf. the discussions in [Ettl, 2001]).

- As a secondary technical aspect, one cannot expected that speckle take the rectangular form of camera pixels. Therefore, speckle are mostly not imaged exactly onto the camera pixels, but a pixel sees the reflection combined from two or more speckles. The light adds incoherently, so the resulting intensity is the *sum* of the *intensities* of individual speckles. If the speckles' electromagnetic fields were added (coherent case), the scenario would not differ from the case of a single speckle detailed above.

White light interferometry with rough surfaces In the previous paragraph, we have derived that for ideally rough surfaces, the backscattered intensity follows an exponential law, with k a normalization constant (cf. Fig. 2.8):

$$p(I)dI = \frac{1}{2\sigma^2} \exp\left\{-\frac{I}{2\sigma^2}\right\} dI \tag{2.43}$$

The negative exponential probability distribution already explains a large part of the challenge that interferometry is faced with rough surfaces: The intensity directly backscattered from the object is predominantly low; most of the light is scattered diffusely. The probability that the reflection has a high intensity is only small. On the other hand, the light returned from the reference mirror has an insignificant loss and is so of constantly bright intensity.

The light diffusely scattered from other areas gives another significant contribution, again to the incoherent background signal and so the non-interfering intensity (I_0 in Eq. (2.12)) dominates even more against the interference amplitude (I_1). Therefore interferometry with rough surfaces suffers from a volatile and often low interference contrast (visibility, cf. Eq. (2.14)).

The visibility gives a good estimate of the reliability of current approaches for height calculation from the interference fringes: If it falls below a critical number, it has been found that the probability of misdetecting the surface height strongly increases [Restle et al., 2004], and outliers or missing values in the height map occur.

Empirically, we consider those pixels as *outliers* (cf. Sec. 2.3.2), for which the processing algorithm unintentionally yields a wrong height value. This will often be a value outside the range defined by the roughness and surrounding pixels. In contrast, *missing values* are those pixels, for which the processing algorithm sees itself incapable of assigning a correct value and leaves a defined "hole" in the height map.

Bearing the statistical nature of the scattering process in mind (cf. also Eq. (2.43)), one sees that these errors are intrinsic to rough surface interferometry and cannot be avoided. We will discuss procedures for the determination and elimination of these errors in Sec. 2.3 and of course Chap. 4.

White light interferometry with smooth surfaces The reflection from optically smooth surfaces is much more stable than from rough surfaces, which generally simplifies interferometry. The phase shift of the reflected field can be evaluated for the height estimation. Therefore it is simple to achieve a much higher height resolution if only the mechanical stability is sufficient.

By definition, optically smooth surfaces impose a phase shift on the reflected field that varies by significantly less than the wavelength λ. This is particularly helpful when processing the interference signal, as the 2π-ambiguity can be ignored, and the evaluation of the inner phase, based on the results from neighboring pixels is possible. White light interferometry here comes into competition with laser interferometry as it cannot benefit from the unique height reconstruction when smoothness is an a priori assumption.

The phase of the reflected signal can be calculated as before with the random walk model. The smoothness prerequisite allows a series expansion of Eq. (2.24) without assuming independence (cf. Eq. (2.25)):

$$E_i \approx |E_i| (1 + i\varphi_i) \tag{2.44}$$

That way, one can find that the phase of the reflected signal is the average of the (marginally different) phase shifts induced by each micro-facet within a speckle.

2.2. Signal processing for white light interferometry

Evaluation of white light interferograms We have seen that the interferogram of a white light source is defined by the auto-correlate, and so the spectrum of the source, cf. Eq. (2.19). In the following, it is handy to use the following equation for the interference fringes, which is a generalization from Eq. (2.13) for arbitrary light sources:

$$I(z) = I_0 + I_1 G(z - z_0) \cos(kz + \varphi_0) \tag{2.45}$$

Here $I(z)$ is the intensity for a given scan position z, z_0 a reference height, $G(z - z_0)$ the envelope of the interference and φ_0 the inner phase shift.

The envelope G is the Fourier transform of the power spectrum of the light source. For the ideal single-frequency source with a delta-peak spectrum, the transform is constant, thus one regains Eq. (2.13). For a broadband source, the envelope peaks at $z = z_0$ (cf. Eq. (2.19)), i. e. when the interferometer is balanced. During the scanning procedure, the arm lengths and so the parameter z is systematically changed. The aim is now to detect the maximum of the envelope G, while this is modulated by the cosine term, the *inner modulation*.

This modulation carries an inner phase shift φ_0 which is different from the signal phase φ used earlier in the chapter: The former describes the shift of the sinusoidal interference modulation against the signal's envelope. It is the residual of the phase shifts occurring at each micro-facet of the surface. As these are of random height and inclination (cf. Sec. 2.1.4), their effect sums up to a random phase shift of uniform distribution, at least for an ideally rough surface.

Limitation of height resolution in speckles Although amplitude and phase of the reflected signal are determined by a random process, the detectable phase shift is not independent from the surface. The residual phase shift after reflection leads to a corresponding shift of the envelope in Eq. (2.45), thus the detected height is subject to the same shift. As the height distribution of the surface micro-facets within a speckle is limited, so are the phase shifts each can contribute to the random sum. Therefore the highest and lowest height that can be detected for a speckle always lie within the limits of the micro-faces' height distribution [Ettl, 2001]. The resulting height is therefore random and uniformly distributed, but within the minimum and maximum of the surface patch that makes up the speckle.

Pre- and postprocessing We use the term *signal processing* for the whole chain of data processing happening around scanning interferometry, from the raw intensity data recorded to the final height map ready for a reliable classification. The final height map should be characterized by its reliability: The height values written down match the true height by a high reliability or even within fixed bounds. The limits which when rough surfaces are measured due to the speckle statistics (see Sec. 2.1.2) should have been taken into concern.

Within the conventional approach, used in most white light interferometer setups, this procedure can be subdivided into two parts: The first step is the preparation of a primary height map, which we name the signal *preprocessing* stage. The second step is then the denoising and correction of this height map, which we name the *postprocessing* or *denoising* stage.

The primary height map is obtained by an algorithm determining the center of fringe contrast. To that end, each pixel is processed separately from any other pixel. While the height in each pixel is estimated, it is also possible to derive a *confidence* measure from the raw data sequence of that pixel. We discuss the possible use of this measure later in the course of this section.

In the postprocessing stage, the aim is to free the primary height map from errors, so that any misdetected height values are corrected. Here, the height map can be considered as a conventional, possibly real-valued, gray value image: The height of each pixel is encoded into a gray value, so that differences in height translate into differences in gray values. Therefore the whole tool box of digital image processing [Jähne, 2002] is available to help removing erroneous and outlying values. The confidence measure acquired in the first stage can support this: It can also be translated into a gray value image, which then can be used for weighting operations of smoothing and restoration. We discuss the current techniques for postprocessing in Sec. 2.3.

While during the preprocessing stage the information acquired for each pixel is used separate from the others, during postprocessing the correlations of neighboring pixels are explicitly used. The approach based on Bayesian estimation, which we will present in Chap. 3, is significantly different, as it joins the pre- and postprocessing stages and so keeps the dimensionality of the original data until the very last step.

Performance of processing algorithms As we focus on postprocessing in this thesis, the following sections discuss only the ideas and not the details and performance of preprocessing algorithms. The diploma thesis [Schraud, 2000] discusses several real-time enabled preprocessing algorithms, with emphasis on the signal processing point of view. It also gives results for their performance. The thesis [Eberle, 2005] deals with the performance of processing algorithms in the high-precision regime and gives some newer results. Additional algorithms and comparisons can be found in [Caber, 1993], [Deck and de Groot, 1994] and [Fleischer et al., 2000].

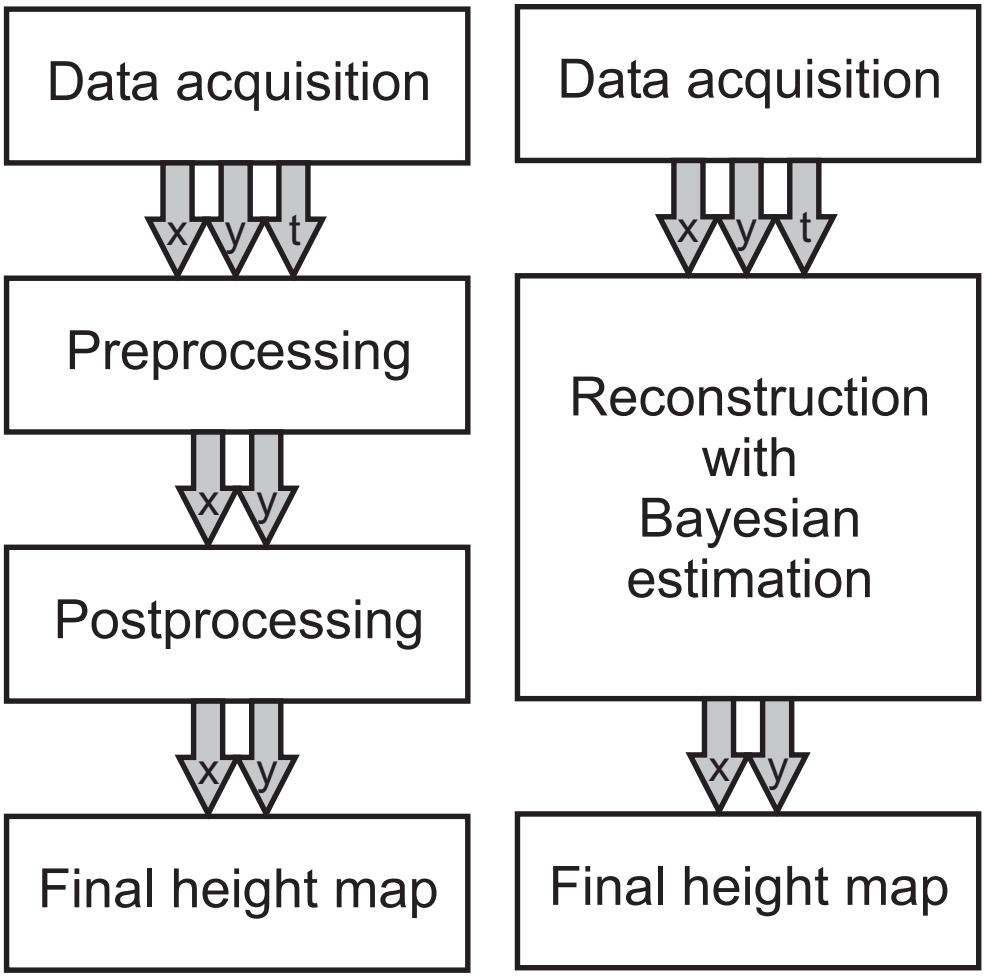


Figure 2.9: Draft sketch of the data flow for (a) conventional strategy versus (b) Bayesian estimation approach for white light interferometry signal processing and height map estimation.

2.2.1. Processing for rough surfaces

For rough surfaces, only the envelope of the interferogram gives information on the height. Therefore, preprocessing methods are aimed at a fast and precise determination of the envelope's center, where it takes its maximum absolute value. However, the desired maximum of the envelope does not coincide with the actual maximum of the interferogram due to the inner phase shift, but is slightly displaced—unless φ_0 is equal to 0 or $\pi/2$.

We will first go over various fast algorithms which do not take this detail into account but accept this error. A possibility to *deconvolve* the envelope from the inner oscillation to get a better estimate of its maximum is given with the Hilbert transformation, which we discuss later in this section.

Further discussions of algorithms for the processing of rough surfaces can be found in the references of the last section, and in [Schraud, 2000] and [Bohn, 2000].

Maximum / minimum method The maximum / minimum method makes use of the basic observation that the global maximum (or minimum) of the interference signal can be found near the desired maximum of the envelope, cf. Fig. 2.10. This approximation becomes the more robust, the higher the wavenumber of the inner oscillation is compared to the breadth of the envelope, so that it has several oscillations around the maximum. The figure shows the approximate situation for the setup used in our experiments. A rough, but fast estimate of the envelope's center is thus given by the absolute maximum or minimum of the interferogram:

$$\hat{z}_0 = \max_z I(z) \qquad \text{or} \qquad \hat{z}_0 = \min_z I(z) \tag{2.46}$$

This very straightforward approach is easy to implement and can be set up in real-time enabled hardware, requiring only a pixel-wise comparison for each frame, with a temporary storage for one value per each pixel.

Even for low noise, this method bears an error due to the indeterminate relation between the envelope's absolute maximum and the phase shift of the inner oscillation. This leads to an average error of this estimation of

$$<\Delta\hat{z}> = \frac{\bar{\lambda}}{4} \tag{2.47}$$

The relation between noise level and coherence length is critical for the correct detection of the maximum: For a high coherence length, the envelope is broad, it has a low curvature near its maximum. Therefore misclassifications of noisy peaks occur more often and lead to jumps of multiples of $2\pi = \bar{\lambda}/2$.

A slight modified approach could use the maximum of the signal's absolute value as an estimator. In that case, due to the doubled wavenumber, the errors and uncertainties would be half of those calculated above, however the computational effort of a hardware implementation would be larger.

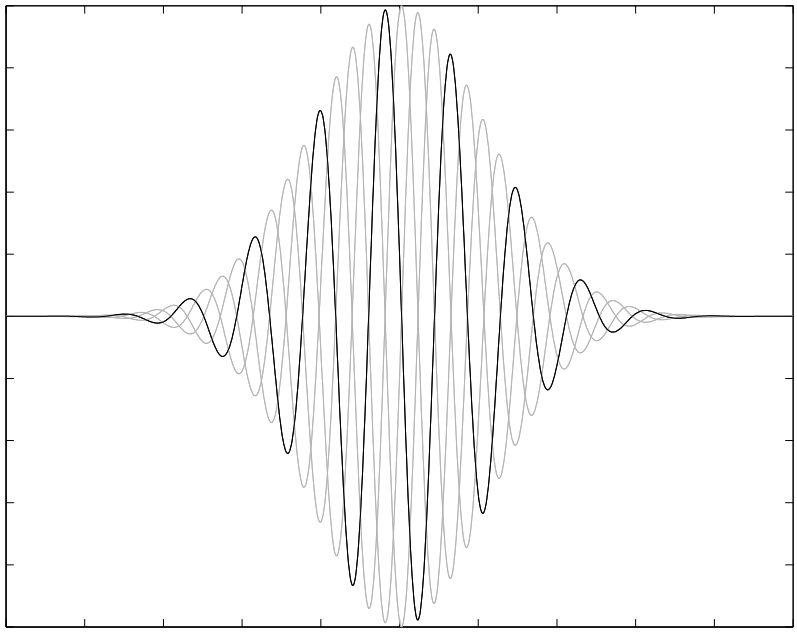


Figure 2.10: Synthetic interference signals with different phase shifts (black: 0, gray: 0.5π, π, 1.5π), but the same envelope, thus representing a common height value.

Contrast method The contrast method is similar to the minimum / maximum approach mentioned in the last paragraph: Here the maximum absolute difference of successive intensity values is used as an estimator for the envelope's absolute maximum:

$$\hat{z}_0 = \max_z |I(z) - I(z-1)| \tag{2.48}$$

Again, this method can be implemented easily in specialized hardware. The above formulation effects a small forward shift (in scan direction) of z_0 from the true maximum. In almost any application, such a constant overall shift of the resulting height map is irrelevant and can be ignored.

The above filter becomes maximum where the interferogram oscillations have a maximum gradient, which is coarsely around the maximum of the envelope. This approach should show a slightly increased robustness compared to the maximum / minimum method, which stems from the high-pass characteristics of the difference operations: low frequencies of the noise are suppressed. However, the characteristics of this filter are not particularly well suited to the characteristics of the noise, which is timely uncorrelated and therefore also very present at high frequencies.

As before, an increased number of misdetections due to a low-curvature envelope can occur in noisy measurements and lead to $\bar{\lambda}/2$ jumps.

Sliding average algorithm As the raw signal is band-limited, a significant increase of robustness against noise can be achieved by adding a low-pass filter

characteristics. A quite simple measure which however still can be implemented in basic real-time enabled hardware, is to average the contrast signal over a number of successive frames after the high-pass processing:

$$\hat{z}_0 = \max_z \{|I(z) - I(z-1)| + |(I(z-1) - I(z-2)| \tag{2.49}$$

$$+ \cdots + |I(z-k) - I(z-k-1)|\} \tag{2.50}$$

As with the contrast method, the asymmetric processing leads to a constant overall shift. A generalized notation for this operation can be found by making use of a window function B and the convolution operation $*$:

$$\hat{z}_0 = \max_z \left(B * |I(j) - I(j-1)|_{j=1\ldots k}\right)(z) \tag{2.51}$$

In this case, the window function is a rectangular window, which corresponds to an unweighted sliding average.

Hilbert transformation The Hilbert transformation (Hilbert filter) is an operation which extracts the analytic phase [Oppenheim and Schafer, 1999] of a signal.

The operator $\mathcal{H}$ of the Hilbert transformation imposes a $\pi/2$ shift on a signal $f(x)$. The phase φ_f of the signal is then easy to extract:

$$f'(x) = \mathcal{H}f(x) \qquad \varphi_{f(x)} = \arctan\left(-\frac{f'(x)}{f(x)}\right) \tag{2.52}$$

The Hilbert transformation does not change the amplitude of the signal, and its transfer function is purely imaginary. It can therefore be described best by the Fourier transform of the transfer function:

$$\hat{h}(k) = \begin{cases} -i & \text{for } k < 0 \\ 0 & \text{for } k = 0 \\ i & \text{for } k > 0 \end{cases} \tag{2.53}$$

The Hilbert transform allows for a decomposition of the envelope and the inner oscillation of the interference signal. So it delivers a signal for which the detection of the maximum which here coincides with the desired maximum of the envelope, is very easy. This can be seen easily as we write down a generalized interferogram (cf. Eq. (2.45) and Sec. 2.2):

$$I(z) = I_0\, G(z - z_0)\, \cos(kz + \varphi_0) \tag{2.54}$$

The Hilbert transform shifts the phase of the signal by $\pi/2$, thus we get

$$\mathcal{H}I(z) = -I_0\, G(z - z_0)\, \sin(kz + \varphi_0) \tag{2.55}$$

The analytic signal can be constructed by adding the real-valued signal I and its $\pi/2$-phase shifted complement. Its name stems from the fact that it is an

analytic function inside the complex unit circle—that is, its Fourier transform is only non-zero for positive frequencies. It can thus be calculated like

$$\mathcal{A}I(z) = I(z) - i\mathcal{H}I(z) \tag{2.56}$$
$$= I_0\, G(z - z_0)\,(\cos(kz - \varphi_0) + i\sin(kz - \varphi_0)) \tag{2.57}$$
$$= I_0\, G(z - z_0)\, e^{i(kz-\varphi_0)} \tag{2.58}$$

With $|e^{ix}| = 1$ the envelope can now be directly obtained as the absolute value of the analytic signal. The height z_0 is then estimated as:

$$\hat{z}_0 = \max_z |\mathcal{A}I(z)| = \max_z I_0 G(z - z_0) \tag{2.59}$$

The Hilbert transformation is analytically elegant, but for its application as a preprocessing filter we have a higher computational effort than with the filters mentioned earlier. Two options are possible in practice: Either, the signal is transformed to the Fourier domain, the filter Eq. (2.53) is applied and the result is transformed back. Or, the filter is transformed into the height (variable z) domain and is applied directly to the signal. Its impulse response would then have infinite terms in z and must be truncated, which leads to a limited precision of the calculation. As well, a slight shift of the filtered signal due to causality can be observed. These effects can be reduced by combining the Hilbert transform with a low-pass filter or windowed/sliding average operation.

Wavelet analysis Wavelets have been proposed as a band-limited filter in white light interferometry. They are used as filters that combine flexibility and tunability within a sound theoretical framework and, in case of dyadic wavelets, fast computability.

Therefore wavelets can be used as a preprocessing filter, followed with a maximum detection to single out the center of the white light interferogram. On the basis of synthesized interference signals, a detailed discussion and evaluation with encouraging results has been reported in [Recknagel and Notni, 1998], cf. also [Sandoz, 1997] and [Sandoz and Jacquot, 1997]. Within a diploma thesis, we have performed a re-evaluation based on real interferometric data, a full account of the results can be found in [Natter, 2003]. Here, we only give a very brief overview of wavelets here and recite the results of this diploma thesis. A detailed introduction to wavelet theory can be found in most signal processing textbooks, especially [Mallat, 1999].

Overview wavelets Wavelets $\psi(z)$ are the kernels of the wavelet transformation, a time-frequency transformation. Wavelets are defined by two properties[1]:

- Wavelets have zero average:

$$\int_{-\infty}^{\infty} \psi(z)dz = 0, \tag{2.60}$$

[1] In this paragraph, the notation of the continuous wavelet transform is used, cf. [Grossmann and Morlet, 1984], [Mallat, 1999].

- they can be scaled and translated (parameters s and u):

$$\psi_{s,u}(z) = \frac{1}{\sqrt{s}}\,\psi\left(\frac{z-u}{s}\right). \tag{2.61}$$

The wavelet transform for the intensity function $I(z)$ is calculated with this *wavelet atom*:

$$WI(u,s) = \int_{-\infty}^{\infty} I(z)\,\frac{1}{\sqrt{s}}\,\psi^*\left(\frac{z-u}{s}\right)dz \tag{2.62}$$

Benefits of wavelet analysis Other than the kernel of the Fourier transform, e^{-ikz}, the wavelet atom and its Fourier transform cease to zero for large values of z or k. The Fourier transform is global and delocalizes the original data in the transformed domain—the whole height series is transformed into one frequency spectrum. In contrast, the properties of the wavelet kernel ensure that the wavelet transform depends on the magnitude of the signal and it's Fourier transform and so allow for a combined time-frequency analysis of the signal [Schwarzer et al., 1996]. In the time-frequency plane, the wavelet coefficients $WI(u,s)$ are large where a chunk of time varying harmonic oscillation matches the scaled and translated wavelet atom.

Procedure As we already know the shape of the interference signal to some degree, but mainly lack its position along the height axis z, a full time-frequency analysis is not required. Instead, only a transformation with the (discretized) best performing wavelet is sufficient. followed by a simple maximum detection, i. e.,

$$\hat{z}_0 = \max_z WI(z, s_{\text{opt}}). \tag{2.63}$$

The wavelet transform is therefore used to set up a particular matched filter.

The *M*orlet wavelet is the recommended wavelet basis for white light interferogram processing [Recknagel and Notni, 1998], as it already has an appearance very close to a interference signal. It is obtained by localizing a complex sine wave with a Gaussian envelope [Daubechies, 1992]:

$$\psi(z) = C\,e^{-z^2/2}\cos(5z), \tag{2.64}$$

with C a normalization constant. The wavelet and its scaled versions are discretized just before use in the computations.

Next, the best scaling level of the wavelet atom needs to be determined. It shows that the interference signal is always manifest in the first three scaling levels. For most scanning speeds, the first or second level shows the highest energy, decreasing strongly towards higher levels. The height estimate is the localization of maximum energy in the transformed domain.

Results In comparison with the real-time enabled methods such as sliding average, preprocessing with the wavelet transformation leads to slightly higher average absolute deviations for $\hat{z}_0$. This approach can therefore not be recommended in general, as other preprocessing approaches deliver better data with less or equal computational effort.

2.2.2. Processing for smooth surfaces

We have seen in Sec. 2.2.1 that the height of a rough surface can only be determined from the envelope of the interferogram (cf. Eq. (2.45)), as the phase shift φ_0 of the inner oscillation is random. With smooth surfaces, we can now use this information, as the height changes only very little from microscopic facet to facet of the surface, significantly less than the interferometer's mean wavelength.

The fast algorithms proposed for rough surface estimation (like contrast method or sliding average algorithm), the envelope cannot be divided from the inner oscillation, which gives rise to a height error of approximately $\lambda/4$.

When the information of the inner phase shift is valid, i. e. when the smoothness assumption for the surface is reliable prior measurement, we can obtain a much higher precision from processing φ_0: The phase of a sinusoidal oscillation can be estimated up to about 1 per cent of the wavelength, $\lambda/100$.

In order to reach this precision, inaccuracies in the physical setup have to be treated much more serious than with rough surfaces: The most important issues are deformations of the optical elements, namely of the reference mirror, due to manufacturing errors and thermal stress, leading to extended artifacts. Another spurious effect has been found originating from mismatching angles of beam-splitter prisms, which lead to a continuous, spatially varying dispersion error that manifests itself as "wiggles" all over the reconstructed height profile, cf. here the investigations in [Pförtner and Schwider, 2001].

2.2.3. Processing for semi-rough surfaces

In everyday practice of measuring high-precision tooled objects one often comes across surfaces, which can neither be considered optically smooth nor rough. That is, the roughness is so low that the spatial speckle statistics does not follow an uniform phase distribution, as one would expect from an ideally rough surface. On the other side, phase jumps of neighboring speckle are still frequent, something which one would not expect from smooth surfaces. Therefore, algorithms devised for rough surfaces deliver a worse height resolution than necessary, and algorithms for smooth surfaces fail as they oversmooth the phase jumps from height discontinuities.

In [de Groot and Deck, 1995] a possible solution to this problem was first published while it was even earlier implemented into these authors' white light interferometer. This approach (named *frequency domain analysis*, FDA or *white light phase-shifting interferometry*, WLPSI) was independently published and reworked in [Larkin, 1996], [Sandoz et al., 1997] and others.

The idea behind these algorithms is to form a hybrid from a peak-detection algorithm, working on the full-signal or the envelope, as presented for rough surfaces (cf. Sec. 2.2.1), and a follow-up phase estimation for the inner oscillation (cf. Sec. 2.2.2). The phase estimation bears a 2π or $\lambda/2$ uncertainty, which can be cancelled with the coarser knowledge of the peak position obtained in the first step, therefore no elaborate phase-unwrapping procedure is necessary.

Alongside edges on these rather smooth surfaces, under large numerical apertures, artifacts described as "bat-wings" due to diffraction can become the dominant source of reconstruction errors. [Harasaki and Wyant, 2000] identified this problem and both [Harasaki et al., 2000] and [de Groot et al., 2002] provide a solution by further refining this hybrid algorithm.

2.2.4. Confidence measure

Fig. 2.11 shows two interference measurements for two points of a metallic surface (cf. photograph in Fig. 4.2). The points are not far apart from each other, however are the measurement conditions very challenging due to the fast scanning speed of $84\,\mu$m/s. While the form of the upper interference signal is significantly deteriorated due to the fast recording, the envelope and it's center or maximum is simple to detect. In case of the lower signal however, the interference is completely submerged in the noise, and intensity variations cannot be safely assigned as contributions from noise or interference signal.

As one expects intuitively, the uncertainty in determination of the envelope's center becomes larger the lower the signal-to-noise ratio is[1]. Based on this, a characterizing parameter has been devised. It is often known as the *confidence* measure, named like this by one of the pioneering groups in white light interferometry, but it exists in similar form in other interferometer setups. For the system used in our investigations the maximum difference between adjacent minima and maxima is used to define this measure [Ettl, 2002], other definitions are equally possible. With this particular definition, a calculation of the confidence number is possible both after each data series is taken and on line, by continuously updating the number from the most recent data. Unfortunately, this confidence number is only poorly connected to other measures of reliability, in particular the ensemble variance of the detected height, which has a stringent mathematical foundation, cf. [Restle et al., 2004].

With the confidence number and the detected height, for each pixel two complementary values are available. This opens up more possibilities for denoising (postprocessing) of the height map, starting from thresholding pixels of insufficient confidence to more advances options, as we will discuss in Sects. 2.3.3 and 2.3.4.

[1] This statement is a main result of the *detection theory*, cf. [Moon and Stirling, 2000] for a starting point of reading.

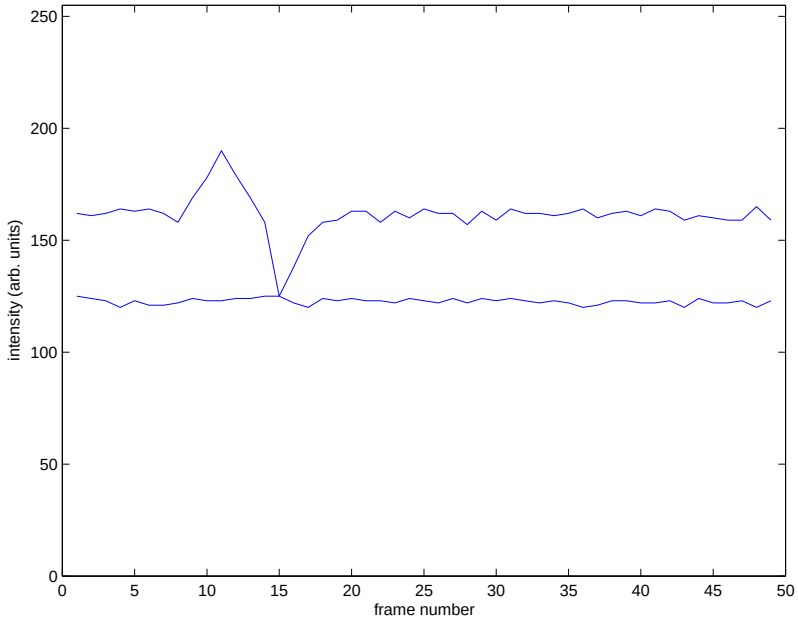


Figure 2.11: Extreme examples of superior (upper) and almost negligible (lower) signal-to-noise ratio in case of 8-fold subsampling of the inner oscillation. The vertical shift of the two curves originates from different background intensity of the reflections. Scanning speed is 84 μm/s, the scale is 1.68 μm per frame.

2.3. Denoising of height maps from interferometry

As it has been stated in Sec. 2.1.4, height maps of rough or quasi-rough surfaces necessarily contain a number of erroneous pixels, i. e. missing values or *outliers*. Depending on the further use of the acquired height map, correction of these pixels is at least desired, if not strictly required when the height map is handed to non-robust operations.

2.3.1. Linear filtering

A linear filters in image processing returns a value that is a linear combination of a pixel gray value and those of it's neighborhood. The coefficients of this polynomial characterize the filter.

Simple linear filters The calculation of the arithmetic mean is the simplest linear filtering operation. The height value of a pixel is replaced by the mean height of the pixel and its neighbors. With the arithmetic mean, all pixel sites have the same weight, therefore all coefficients of this filter are equal. This operation implicitly assumes that the height map should be smooth, at least for the pixel and its neighbors–this assumption of course fails near edges and irregular structures.

Mathematically, linear filtering can be expressed by operations in which filter

masks are convolved with the height map or image to be processed. Therefore, the matrix notation is preferable rather than writing down nested sums.

Let I be an image. Linear filters G are those image processing operators, for which the following linearity constraint holds:

$$G(\alpha I_1 + \beta I_2) = \alpha G I_1 + \beta G I_2 \tag{2.65}$$

In particular, one can see that the filter response is proportional to the size of the input signals, which are scaled with factors α, β.

As an example let us consider the simple *rectangle filter*, which calculates the arithmetic mean over all pixels within the neighborhood of any pixel. The size of the neighborhood is determined by the filter's size. It is common to name the filter according to its size, e. g. the mask of a 3×3-rectangle filter is:

$$G_{\text{sq}} = \frac{1}{9} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{2.66}$$

The filter response can be calculated by a discrete convolution:

$$I'_{x,y} = \sum_{x'=-r}^{r} \sum_{y'=-r}^{r} G_{x',y'} I_{x-x',y-y'} \tag{2.67}$$

Another filter applied in many image denoising problems is the *Gauss-filter*. Its filter mask is a discretization of the two-dimensional Gaussian curve. This filter has two parameters: as usual the size of the filter mask, and the width σ of the Gaussian curve that is discretized onto the filter mask. This filter puts highest weight on the central pixel of its mask, it then becomes smaller the farther a pixel is away from the center. For useful filter masks, these parameters should be chosen jointly, so that the mask is large enough to cover the distance over which the Gaussian is influential over the noise background. For $\sigma = 0.5$ of the lattice spacing, the 3×3 mask is one reasonable choice for many application:

$$G_{\text{Gauss},\sigma=0.5} = \frac{1}{1.64} \begin{bmatrix} 0.02 & 0.14 & 0.02 \\ 0.14 & 1.0 & 0.14 \\ 0.02 & 0.14 & 0.02 \end{bmatrix} \tag{2.68}$$

However, linear filters are generally not robust (cf. Sec. 2.3.2): After processing the raw data, the errors found in the primary height map are often outliers—pixels, whose height range is far outside the reasonable range. With a linear filter, these (like any) pixels have a linear influence on the output of the filter. With one pixel of the input data having an arbitrarily wrong value, the output becomes as well arbitrarily wrong. This should be illustrated with a 3×3-rectangle filter G_{sq} working on a 5×5-patch of an image, where an outlier

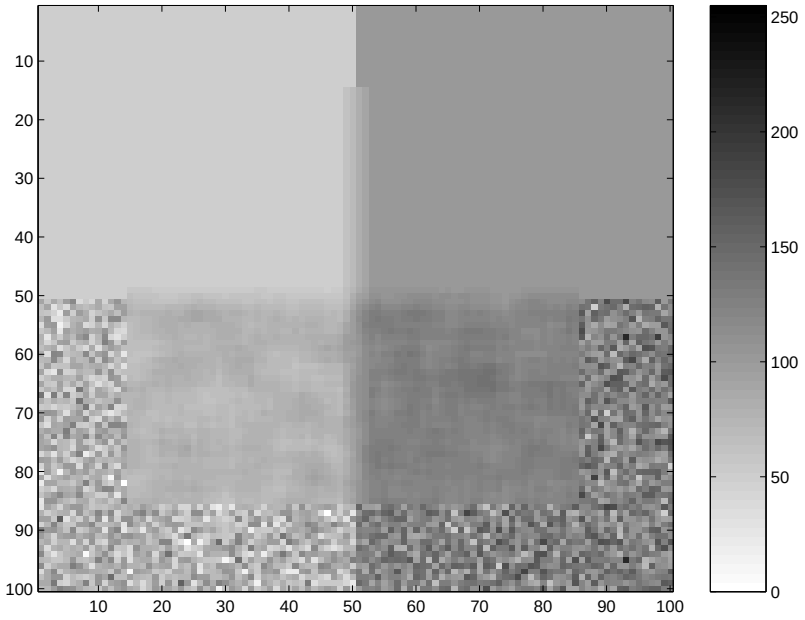


Figure 2.12: Illustration of the effects of a 5×5 rectangle filter (applied in inner area) for an image edge between two smooth and two noisy quadrants (Gaussian white noise).

is put in the center pixel:

$$G_{\mathrm{sq}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \dots & & & & \\ & 1.1 & 1.1 & 1.1 & \\ & 1.1 & 1.1 & 1.1 & \\ & 1.1 & 1.1 & 1.1 & \\ & & & & \dots \end{bmatrix} \tag{2.69}$$

$$G_{\mathrm{sq}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1000 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \dots & & & & \\ & 112 & 112 & 112 & \\ & 112 & 112 & 112 & \\ & 112 & 112 & 112 & \\ & & & & \dots \end{bmatrix} \tag{2.70}$$

A further, often undesired effect of linear filters is the blurring of image edges or object boundaries: if a part of the pixels under the filter mask belongs to one gray value or height domain and another part to the other, a linear filter returns values lying in-between the two domains, cf. Fig. 2.12.

One can of course understand this situation similar to the case calculated above, now with a series of adjoining outliers. Each is spread out over the neighbors leading to the blurry edge. Another view is to conceive that here the quietly underlying assumption of linear filtering, that the original image be smooth, is simply failed. Robust filtering can be seen as a way of weakening this assumption and still gaining better results by differentiating between outliers and edges.

Weighted mean filters *Weighted* filters are adjustable, which can be used to take the quality of the data into account. That way, it is possible to achieve practical robustness with this modification of linear filters.

The *weighted* arithmetic mean is defined as follows: For a set of data points x_i with variance σ_i^2, the weighted arithmetic mean is given by

$$< x_i >= \frac{1}{k} \sum_i \frac{x_i}{\sigma_i^2} \quad \text{with the normalization } k = \sum_i \frac{1}{\sigma_i^2} \tag{2.71}$$

For each data point weights are given, in this case we have $w_i = \sigma_i^2$. This idea easily carries over to image processing where weighted filtering is also known as *normalized convolution* [Granlund and Knutsson, 1995]. For an image I, an image of weights W and a filter mask G, the operation is defined as:

$$I' = \frac{G * (WI)}{G * W}, \tag{2.72}$$

thus the name becomes obvious.

Some possibilities of choosing the the weights other than as the inverse variances are interesting. In particular, the variance of a single pixel is not directly available in image processing, and a corresponding measure for white light interferometry is problematic [Restle et al., 2004]. For smoothing interferometric height maps, we consider the following two approaches:

1. The weights can be derived from the spatial variability on a local scale of the image. It is assumed that the image is overall smooth. Areas of high variability in the gray values (or height) then hint towards a low measurement confidence, which could also be seen in the pixel-wise variance if a sequence of image were acquired. The local variability can be measured with numerous different filters, like gradient-based filters and other edge detectors.

   However, it should be noted that it is not obvious at all how to deduce the variances (or at least some measure of uncertainty) of a single pixel from the spatial variability and assumptions have to be made: If the image were generated from an ergodic stochastic process, the spatial variance and gray value variance a pixel would be exchangeable—this however can almost never be assessed.

   In white light interferometry, as we have discussed in Sec. 2.1.4, the surface height profile, as seen by the interferometer, itself is generated by a stochastic process. The above-mentioned assumption of general smoothness is thus not valid for an image representing a height map from white light interferometry.

2. With scanning interferometry we have the special case that the height map is not the primary information acquired, but already product of a prior processing step. Additional insight into the variability of each pixel could thus be gained from looking at the raw data. One way to do this is to derive empirical weights from the confidence values that are available with some preprocessing algorithms. We discuss this in the next Sec. 2.3.3.

2.3.2. Robust filtering

Within image processing, we consider all those operations as *robust filters*, which do not lead to arbitrarily large results when an arbitrarily large input (like with outliers) occurs. Borders of objects that come up as edges in an image can be considered as outliers [Black and Sapiro, 1999]. We can thus expect that a robust filtering algorithm can avoid to blur or wash out these features. Robust filters can return, depending on the quality of the input signal, a response of substantially different quality, they are therefore generally not linear.

Median filter The median filter is the best-known *rank-order filter*. This class of filters have in common to perform sort and re-ordering operations on the input set. For example, the maximum or minimum operation are primitive rank-order filters.

The median filter, for a set $K = \{k_1, k_2, \ldots, k_n\}$ of input values, chooses the value of the middle index from the ordered list $\{k_a \leq k_b \leq \cdots \leq k_n\}$[1]. If the number of input values is even, by convention the mean value from the two values next to the middle of the list is returned. The returned value is therefore always within the range of the input set, although intermediate values not on the input carrier may come up.

Due to its non-linearity, the median filter in image processing cannot be expressed by a convolution with a filter mask, but a sort operation has to be performed on each pixel. To that end, a neighborhood system is defined to set the spatial range of the filter, often a square $n \times n$ field of pixels is used. The gray values within the neighborhood around the pixel currently processed are taken into a list (here, of size n^2), reordered and the value at position $|n^2/2|+1$ of the list is written into the filtered image.

The median filter shows maximum robustness against outliers in the Huber sense [Donoho and Huber, 1983]. The filter result does not change significantly if the list to be filtered is augmented with outliers, i. e. entries of grossly different values, as long as the list does not contain more outliers than actual values. Therefore the median filter is well adapted to the removal of outliers from a height map. It preserves image edges well, from the filter's view, these can be seen as a group of outliers, which is ignored until it becomes the majority of the input list.

In contrast to the weighted filters discussed below, the median filter cannot be made adaptive to the quality of the original data it is processed on. Both in areas with many outliers and in those with few, it performs the same filtering. This is why the median filter has a predisposition to oversmooth in image areas with little or no outliers. It can possibly remove image features that only appear like outliers, but which are positively backed up by their good data quality and hence could be saved if one could recognize this. The tendency to oversmooth becomes greater the larger the filter mask is chosen. It therefore depends on the properties of the image if and with which mask size a median filter should be applied.

[1] Although discussing image filtering, we start adopting a simple notation of ordinal numbers.

Adaptive median filtering Similar to the ideas that lead to weighted filters (cf. Sec. 2.3.1), one could imagine applying a robust filter selectively only to those pixels which are suspected to be erroneous, and to leave the others intact.

We wish to apply a smoothing filter to only those pixels in the image or height map, which we can identify as outliers. In the realm of robust statistics, the spatial median filter seems appropriate, and the same filter can provide us a way of detecting outliers.

We apply the median filter to a data set that consists of a central pixel of gray value x_0 and its neighbors, $x_1, \ldots, x_k$, where it yields

$$\tilde{x}_0 = \mathrm{med}\{x_0, \ldots, x_k\} \tag{2.73}$$

Let us now look at the distance $|x_0 - \tilde{x}_0|$: If the central pixel is an outlier, this difference is outside the range of variation that we expect given the roughness of the surface under inspection. For an optimum distinction, this threshold should be adaptable to the fraction of outliers and the local variability in the image or height map.

In a study on breakdown points[1], Hampel [Hampel, 1985] proposed to identify outliers by their statistical variability, which is (by the definition of outliers) significantly larger than the variability of "normal" samples.[2] The variability of the dataset is measured with the MAD (median of absolute deviations):

$$\mathrm{MAD}\{x_0, \ldots, x_k\} = \mathrm{med}\{|x_0 - \tilde{x}_0|, \ldots, |x_k - \tilde{x}_0|\} \tag{2.74}$$

The MAD-value is a robust measure for the variability; it does not change significantly if one adds gross outliers to the dataset. The *Hampel detector* for outliers can then be written:

$$|x_0 - \tilde{x}_0| \geq c\,\mathrm{MAD}\{x_0, \ldots, x_k\} \tag{2.75}$$

The pixels fulfilling this equation are identified as outliers and can be replaced by better estimates, like the median value $\tilde{x}_0$ itself. The parameter c in Eq. (2.75) gives freedom to tune the outlier detector: the larger it is chosen, the more tolerant we are to values lying far-off, possibly missing out some outliers. If the probability distribution of the dataset is known, the optimum value for c can be found by simulation, as demonstrated by [Davies and Gather, 1993] in case of a normal distribution for the underlying dataset. In our experiments, we measured the quality of a reconstruction with the average absolute error of the estimate against a reference height map (cf. Sec. 4.3.2). c was chosen to minimize this error, and we found optimum values slightly off the ones simulated for a normal distribution. A computer implementation of this filter is easy, and for our experiments we used a fast MATLAB implementation requiring about 0.8 s per height map.

It is crucial for the Hampel detector that the variability of the dataset is measured with a robust approach like the MAD-value. Non-robust measures would get large if outliers exist in the dataset and so automatically become insensitive to them.

[1] For the theory of breakdown points, cf. also [Huber, 1981] or [Donoho and Huber, 1983].

[2] The author would like to thank Chr. Hennig (Seminar for Statistics, ETH Zürich) for pointing out the possibility of using Hampel's approach in this filtering scheme.

2.3.3. Specialized filtering approaches

In this section we describe filtering approaches which make use of the specialties of white light interferometry. Due to the preprocessing required, additional information comes up during the formation process of the height map, which we try to use for filtering.

Confidence thresholding The confidence measure, which is provided with some white light interferometer setups, gives a coarse reflection of the reliability of an obtained height value. A straightforward approach to make use of this is to establish a threshold filtering based on the confidence value. Pixels below the threshold are assumed to be unreliable and can be marked, either as "invalid" to leave them out of further processing, or they can be subjected to additional filtering as to replace those pixels. However, the first option must often be disfavored, when afterwards additional image processing will be used, especially neighborhood operations: Often it is not clear how to handle invalid pixels within a filter mask, and the correct choice to mark the whole affected neighborhood as invalid makes these regions grow with every processing step. The data set usable for classification or feature extractions stays reliable, but becomes smaller with each iteration. For the second option, a number of possibilities exist how to estimate a missing pixel from its surroundings. E. g., one could proceed with the median value of the neighborhood, which is similar to adaptive median filtering (cf. Sec. 2.3.2).

Variance-weighted (nonparametric) smoothing The confidence measure can also be used for a weighted filtering approach: One assigns low weights to pixels of low confidence, and higher weights to those of a higher confidence value. Intuitively, this should reduce the influence of outliers and invalid pixels on the smoothing process if both have a rather low confidence measure.

More precisely, we weight each height value with its uncertainty which we derive from the confidence measure. To that end, we try to relate the variance of the calculated height values to the confidence measure obtained from the interferometer setup. The smoothing is then done by weighted averaging over a local neighborhood, the size of which is again chosen according to the uncertainty. That way, this approach is an extension of smoothing by normalized convolution incorporating information from the confidence measure. A detailed description can be found in [Restle et al., 2004] and the thesis [Restle, 2003], we here give a sketch of the basic ideas.

Relation of confidence values to variances From practice we know that a low confidence corresponds to a high uncertainty in the obtained height value (cf. Sec. 2.2.4). The confidence values are therefore related to the empirical variance of the height data. To put this finding on solid ground, in the thesis [Restle, 2003] investigations with multiple measurements of challenging surfaces have been carried out. Based on a series of 25 height and confidence maps, for

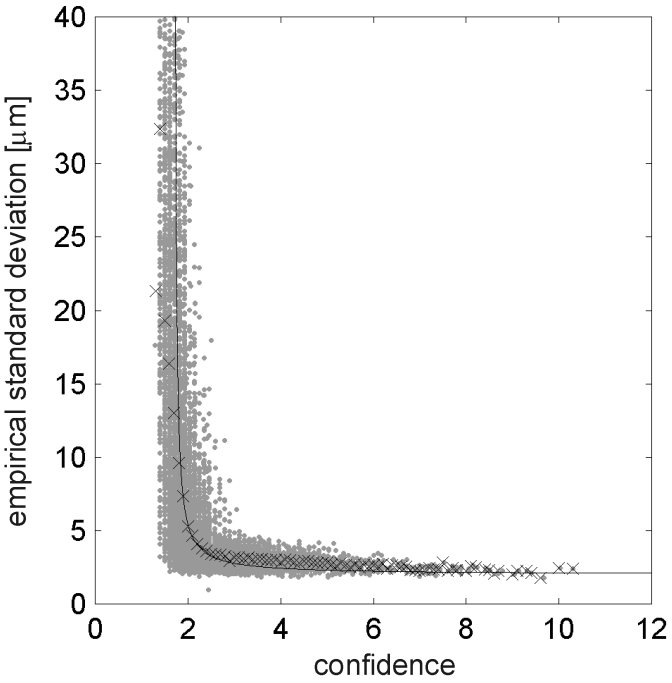


Figure 2.13: Robust empirical standard deviation against confidence values for a rough, uncooperative surface (taken from [Restle et al., 2004]).

each pixel the standard deviation of the height values was calculated and plotted against the arithmetic mean of the corresponding confidence value.

Fig. 2.13 shows the results. From the hyperbolic form of the scatter plot, one sees saturation in two directions: First, the empirical standard deviation cannot fall below a certain residual error. Second, the recorded confidence values have a lower limit, and the corresponding standard deviation has an upper bound. This however can be related to the limited scanning range and height output range and thus limited error in the data.

The data is parameterized with a shifted hyperbola (cf. line plot in Fig. 2.13); the functional form of the parameterization is chosen empirically. The algorithm is not tunable by single parameters, but has full flexibility with the adaptable interpolation function. The surface is reconstructed by local smoothing instead of a functional fitting, which explains the naming "nonparametric smoothing".

This function assigns each confidence value an estimated standard deviation. From the figure it is however clear that only for large confidence values the deviation can be deduced, for small confidence values it bears a large error. In that sense the confidence value is a poor estimator for the ensemble standard deviation.

Smoothing with variable-width kernels The smoothing of the height map is performed by normalized convolution (cf. Eq. (2.72)), with the weights chosen as the inverse empirical variances. While it is possible to fix the size of the

smoothing kernels support to an arbitrary value, a different approach has been investigated:

The size of the smoothing kernels is fine-tuned to the local uncertainty, so that in areas where the expected noise (calculated from the confidence values) is larger, a bigger smoothing mask is used, while in areas of expectedly low noise the mask is kept narrow and blurring is avoided the best possible. By iteratively and selectively applying a filter of increasing size, one can control the smoothing process and reach a pre-defined target uncertainty for the whole image.

Performance With this extension of normalized convolution smoothing one can smooth the image so that the height values of all pixels are under a given target variance. By the iterative approach with variable sized filter kernels, the blurring and loss of spatial resolution, unavoidable with linear filters, is reduced to a minimum. This approach however requires the confidence-to-variance correspondence to be known in before, at least for the rough class of surfaces under investigation.

On the test images in [Restle et al., 2004], the filter performs slightly better than a 3×3 mask median filter, which on the other hand is much faster to perform. Despite adapting the size of the smoothing mask, this filter increases the error of pixels which have a very low uncertainty, probably due to estimation errors for the variance around low confidence values (cf. Fig. 2.13).

For the sample piece we use to evaluate postprocessing algorithms in this thesis (cf photograph in Fig. 4.2), the actual correlation of confidence values to the ensemble variance could not have been established. With the shifted hyperbola standard correlation however, the algorithm generally yields a slightly inferior average absolute error compared with the median filter and other algorithms (cf. Sects. 4.3.4 and 4.3.5).

2.3.4. Further possibilities

Using a local parameter from the image formation process like the confidence measure in postprocessing is not necessarily limited to white light interferometry. Also other imaging systems like SAR[1], MRI[2] or CT[3] require complex operations on the raw data before an image can be built up. Only for MRI, an extension which creates weights for postprocessing of noisy data is known to us ([Prüssmann et al., 1999] use the local coil sensitivity).

On the other side, there are a number of image denoising approaches that make use of a reliability measure to weight or tune their algorithm. These algorithms can possibly easily be adapted to utilizing a confidence measure instead.

A variance measure, derived from features and properties of the spatial neighborhood is used in [Recknagel et al., 2000] to constrain smoothing of height maps obtained by confocal microscopy.

[1] Synthetic Aperture Radar
[2] Magnetic Resonance Imaging
[3] Computer Tomography

2.4. Alternative approaches to interferometric height measurement

White light interferometry is a "young" technology, its application did not yet spread wide outside laboratories, and other approaches for measuring surface height, namely tactile mechanical testing, are still very well established.

Mechanical methods for height measurement Especially in industrial manufacturing, mechanical devices which acquire line-like surface profiles with a tactile sensor have a long tradition and are continuously being used. Discussion of and comparison with these systems is however not in the scope of this thesis. The reason is that with these systems, much time and a great effort is needed in order to record spatial height information instead of line-like profiles. Mechanical tactile systems always add directional artifacts and the spatial resolution differs significantly alongside and perpendicular to the sensor movement direction. Therefore they should play no role in spatial surface metrology.

In the following paragraphs, we will briefly discuss the alternatives to white light interferometry that are based on optical measurement principles and give some further reading.

Alternative optical approaches For white light interferometry, a height resolution δz down to about 0.01 μm has been reported [Harasaki et al., 2000], here with a hybrid algorithm, explicitly also using the phase information[1]. For the resolution that is achievable with the sliding average algorithm (cf. Sec. 2.2.1) theoretical considerations exist [Fleischer et al., 2001], but the simulations are still too optimistic. With significantly reduced instrumental investment and lower requirements in computational power and algorithmic complexity, triangulation approaches cover the field of lower height resolution, that is, above $1\,\mu m$ [Bohn, 2000]. The confocal microscopy approach has a height resolution roughly comparable to white light interferometry. For laser interferometry, resolutions down to approximately 0.1 nm [Wang, 2003] have been achieved. Such a precision can only be achieved with surfaces that have a smoothness in the same order of magnitude, as then the measurement range is correspondingly low or an elaborate phase unwrapping procedure has to be applied subsequently.

Triangulation approaches The basic principle of triangulation is the estimation of the height of a surface patch from its angle of view. Two principal approaches can be distinguished:

Passive triangulation: The surface is illuminated diffusely and observed with two cameras which are tilted against each other. The height is calculated from the lateral displacement (stereo disparity). Here it is necessary to

[1] The two terms *resolution* (meant as smallest height difference that can be differentiated) and *repeatability* (stability of a height measurement over a measurement sequence) are easy to confuse. One should be aware of this when one comes across comparisons of the capabilities of different measurement techniques.

match corresponding surface patches from the two cameras. This is, at least for a machine vision system, a non-trivial task requiring prior knowledge [Schlesinger, 2003].

Active triangulation: A line or a geometric pattern is projected onto the surface. The measurement triangle is formed between the projector, the camera and the surface, so that the lateral displacement of a projected feature is proportional to the height of that patch. When the projection consists of parallel lines, the height can only be calculated for a corresponding orthogonal line-like region. When a two-dimensional pattern is used, a spatial measurement is possible. In that case it is necessary to vary the pattern over several image acquisitions to obtain uniqueness. Examples for realizations of this approach are *Moiré* triangulation and *phase-measuring triangulation* (PMT), cf. [Bohn, 2000].

Laser interferometry We have already discussed that interferometry with a laser source can technically be seen as a predecessor to white light interferometry. The optical setup for both is similar, the central difference lies with the interference signal, which in the ideal case is purely sinusoidal (for Eq. (2.45), so the "envelope" would be constant: $G \equiv 1$). As the interference signal is strictly periodic, the range of uniqueness is limited to the laser wavelength used. Without further assumptions on the spatial structure of the surface, the height of a pixel can only be determined within

$$\Delta z = \frac{\lambda}{2}. \tag{2.76}$$

This restriction limits the field of application for laser interferometry to the inspection of de facto optically smooth surfaces, like optical mirrors, or semiconductor dies. Discontinuities in form of height steps larger $\frac{\lambda}{2}$ are folded back to within the range of uniqueness, i. e., the signal's phase is *wrapped*.

There are two main approaches to extend the range of uniqueness:

Larger wavelength One option to enlarge the unique measurement range is the use of a larger wavelength. Precision-worked metal surfaces typically become optically smooth with respect to a infrared source of $\lambda \gtrsim 1\,\mu\text{m}$. If the height is estimated from the local phase of the interference pattern, one has to expect a reduction of the height resolution when using a larger wavelength source: The slope of the signal, dI/dz is reduced as the phase progress $d\varphi/dz$ is smaller. Therefore the phase becomes more difficult to detect with a detector having a discretized output—that is the case for all digital imaging systems.

Heterodyne principle We have already discussed heterodyne interferometry in the context of multi-wavelength interferometry (cf. Sec. 2.1.1, see also [Sodnik et al., 1991]). The synthetization of the longer heterodyne wavelength can be done directly in the interferometer, or "virtually" by overlaying of successive measurements (λ-shift interferometry). The detection

of the contributions with different wavelength is then done separately. In any case, for the synthesized wavelength we have:

$$\frac{1}{\lambda_{\text{synth}}} = \sum_i \frac{1}{\lambda_i} \tag{2.77}$$

One can see that the longest heterodyne wavelength can be achieved by combining light that differs only little in wavelength.

The differences between synthesizing on-line (directly in the interferometer) and off-line become clear when we bring back to mind (cf. Fig. 2.3) that for little differences in wavelength, the outer modulation becomes very slow. Thus the challenge of the direct approach becomes the *fringe order identification*, which requires thoughtful sampling of the signal with the fewest number of frames. The main difficulty is to differentiate between neighboring sinusoidal "fringes" that differ only slightly in height for a small Δk, thanks to the then-slow cosine modulation. With noisy measurements, misclassifications become more frequent and can lead to $\pi/\bar{k}$ jumps in the height estimate [Dändliker et al., 1995].

This last issue can be circumvented if the heterodyne wavelength is only synthesized virtually, i. e. off-line in the processing computer. This can be done either having both wavelengths in the interferometer and making the detector wavelength sensitive, so as to register the two wavelengths separately. Another approach, based on speckle-interferometry, continuously switches between the two injected wavelength and has the detector synchronized accordingly [Meixner et al., 2003]. With any these approaches, it is possible to gain a larger range of uniqueness without compromising the depth resolution, albeit at the price of higher system complexity.

3. Bayesian estimation in image reconstruction

3.1. Foundations

Bayesian image restoration A central paradigm of Bayesian approaches is the idea of a hidden truth behind what is observable. If we move our focus on images and image processing, this means that the visible image is then only considered as a (often degraded) instance of the underlying true image. The degradation process usually accounts for noise, artifacts and other unwanted features within the data. As this process is usually both unknown and dominated by random effects, it causes that the result is badly correlated to the original, the true data.

Some groups of researchers (cf. [Chu et al., 1998]) consider Bayesian image restoration as one of two principal paths to image denoising, contrasted with filtering (linear, non-linear, robust, morphological, and others) on the other side.

As originally defined by Hadamard in 1902 in a very general context (accessible via [Tikhonov and Arsenin, 1977]), a mathematical problem's solution is *well posed* if

(a) it exists,

(b) it is unique, and

(c) it has a continuous dependency on the data.

So far, due to its non-uniqueness and discontinuous dependency on individual measurements, we have to consider the image reconstruction problem in fact severely *ill-posed:* By the high interrelation of the variables representing image pixels, the optimization problem in reconstruction is usually non-convex, which then usually requires time-consuming sampling strategies, for example via Markov chain Monte Carlo (MCMC). The restauration problem is otherwise simply not accessible to any attempt of brute-force solution as it represents an overwhelmingly large computational load.

The transformation of an ill-posed problem towards a problem with a stable inverse is known as *regularization*, this term has also been adopted to the field of image restoration [Katsaggelos, 1989]. In Bayesian image restoration, this is achieved by adding prior knowledge about the unknown truth within a Bayesian approach.

The central and maybe most serious problem—which immediately arises when we consider an image as a random system—is the number of possible

configurations that could be taken into account: For a x-by-y pixel image with z gray values and no further restrictions, $z^{x \cdot y}$ configurations are possible. For a 100×100 binary image, this is 2^{10000} or $\approx 10^{3000}$ configurations, by far exceeding the number of particles in our universe. For any estimation procedure this sets up a sampling problem which is often regarded as the main drawback of Bayesian approaches: it requires either sophisticated strategies to quickly reach the high-probability regions of the configuration space, or specialized approaches like the one we work out in this thesis that fit into niches and cover only selected classes problems. On the other hand, Bayesian approaches allow for a more abstract formulation of the optimization target and so are expected to require less experience with filter selection and their tuning [Chu et al., 1998].

The image labeling problem Under a more general point of view the image processing problem we are discussing all through this thesis can be considered as a *labeling problem*. Our task is to attach labels to sites, in our practical application that is to attach labels of discrete height values to pixel positions. Another image processing problem that fits into this framework is edge detection, which is to assign the dichotomous labels "edge" and "no edge" to the dual lattice of pixels. Some further labeling problems are stereo reconstruction, with labels identifying the depth of recognized objects in each pixel [Schlesinger, 2003] and the classification of ground and soil features on the basis of possibly multispectral satellite images (SAR data). For earlier works in this field, yet hampered by the computer power of their time, cf. [Kittler and Föglein, 1984], [Hjort and Mohn, 1984], [Mohn et al., 1987], and also [Zhang et al., 1990], as well as the later publication [Shekhar et al., 2002]. Obviously still before the arrival of adequately powerful computers, the approach mostly came out focus for geoscience.

Our height estimation task can therefore be seen as a further example of the general labeling problem, with labels from a set of discrete height values assigned to image sites.

Consistency of a problem For the computational treatment of the corresponding optimization problem the question of *consistency* is important. A problem is consistent, if its definition in local properties can be consistently transformed into a description in global properties. Consistency is required to make use of the Markov random field framework [Chalmond, 2003], [Li, 2001a]. However, the height estimation task can only be solved directly if it is not formulated consistently. In Sec. 5.1, we will discuss modifications that could build a bridge to Markov models.

3.1.1. Setting of the problem

Throughout this and the following chapter, we adopt a notation merged from the textbooks [Chalmond, 2003] and [Winkler, 2003], hoping it will give a legible and not-so-cluttered symbolic description of the ideas. Furthermore, we will

restrict ourselves in most cases to the discrete notation, which can be transfered straightforwardly into practice.

We start off at a single pixel, or *site*, for which a finite number of settings is available. These are usually identified with gray levels. The sites are usually located in a fixed pattern, which is a rectangular matrix for most image sensors. To identify single sites, we simply require some enumeration of pixels, be it row-wise, column-wise, or else. A site can then be identified by its site index s, its setting be $x_s \in X_s$, with X_s the finite state space of this particular site. The setting of site s is then a point in X_s.

The set of all sites (pixels) will be named S. Settings of several pixels taken together make up a single point in the associated product space. For the whole of S, $\boldsymbol{X} = \prod_{s \in S} X_s$ let be this state space (configuration space). Fixing the values for all sites gives a realization of an image that is a single point in this $||S||$-dimensional space: $\boldsymbol{x} \in \boldsymbol{X}$. Vectors and matrices will usually be written in bold letters.

In most cases, we will not discuss settings of the whole set of sites, but restrict ourselves to small groups of sites, which are usually *neighbors*. Let a set of neighboring sites be A, with $A \subseteq S$. In this case, the sites take a setting in the space $\boldsymbol{X}_A = \prod_{s \in A} \boldsymbol{X}_s$, which can be projected onto $\boldsymbol{X}$ by means of some mapping.

When discussing a probability distribution P of a random variable $\mathcal{X}$ and its realizations x, we leave away the random variable from the notation:

$$P(\mathcal{X} = x) \equiv P(x) \tag{3.1}$$

This will usually go without any ambiguity. Adopting that convention, we gain some simplicity and clarity.

3.1.2. Bayesian estimation

The Bayes theorem [Bayes, 1763] is the central idea and foundation of this work. It is an interesting side note that it cannot be proven from very first principles, but has to remain axiomatic. To make it more plausible in a graphical way, Venn diagrams are often used [Stahel, 2002].

The Bayes theorem is often deduced from properties of a joint probability distribution $P(\mathcal{X} = x, \mathcal{Y} = y) \equiv P(x, y)$ of two random variables $\mathcal{X}$ and $\mathcal{Y}$ and their respective laws. For the conditional probabilities we have:

$$P(a, b) = P(a|b)\ P(b) = P(b|a)\ P(a) \tag{3.2}$$

The last terms from Eq. (3.2) can directly be transformed into:

$$P(a|b) = \frac{P(b|a)\ P(a)}{P(b)} \tag{3.3}$$

The denominator has to be assumed non-zero, which is usually sensible in the context of Bayesian inference (see below).

y	measured data
x	hypothesis of the true image
$\Pi(x)$	*a priori* probability / probability distribution ("prior")
$f(y\|x)$	likelihood
$P(y)$	evidence
$\mathbb{P}(x\|y)$	*a posteriori* probability / probability distribution

Table 3.1: Naming conventions for Bayesian inference: symbols and their designation as used in Eq. 3.4.

The basis of Bayesian inference is this last equation (3.3). Traditionally, its constituent parts are given a special naming as well as a specific notation, which is usually similar to the following:

$$\boxed{\mathbb{P}(x|y) = \frac{f(y|x)\ \Pi(x)}{P(y)}} \qquad \text{Bayesian paradigm} \tag{3.4}$$

The names of the symbols used here are given in Table 3.1.

These particularities hint towards the origin and the special importance of this formula, Eq. (3.4), as well as its designation as Bayesian *paradigm*.

The formula provides a powerful tool for statistical inference from an imperfect measurement and prior knowledge about the underlying truth. Both ingredients are brought together to form the a posteriori probability. In addition, an estimation function Φ, with $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is introduced which chooses the "best" image hypothesis $\boldsymbol{x}$ from the a posteriori probability distribution according to some measure. The choice for Φ can also be motivated by a minimization problem for the more abstract *cost function* (cf. Sec. 3.1.4), which assigns costs to the estimates depending on how "far" the estimate lies off the truth. A frequent choice for Φ is the maximum operation: here, the hypothesis with the highest a posteriori probability is selected as an estimate of the unknown truth.

In most applications, only one measurement $\boldsymbol{y}$ is considered, and the *evidence* is therefore only a constant scaling parameter. From there, we can leave it away and are thus left with a proportional expression of only the numerator and with $\boldsymbol{y}$ fixed:

$$\mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) \sim f(\boldsymbol{y}|\boldsymbol{x})\Pi(\boldsymbol{x}) \tag{3.5}$$

The Bayesian inference problem is therefore the following:

$$\hat{\boldsymbol{x}} = \Phi(\mathbb{P}(\boldsymbol{x}|\boldsymbol{y})) \quad \text{with } \mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{x})\ \Pi(\boldsymbol{x}) \tag{3.6}$$

In the following, we discuss the a priori probability or short, *prior* $\Pi(\boldsymbol{x})$ and the likelihood $f(\boldsymbol{y}|\boldsymbol{x})$.

3.1.3. Prior and likelihood

Prior The prior $\Pi(\boldsymbol{x})$ is the quantity that introduces our abstract prior knowledge about the true image properties into the estimation process. For each image hypothesis $\boldsymbol{x}$ from the hypothesis space $\boldsymbol{X}$ it assigns a value which describes the probability of that image. The truth itself is unknown, only expected values for some of its properties exist and are implicitly contained in the prior functional. Therefore the distance between hypothesis and truth can only be assessed by these properties. The deviations can be weighted differently by use of a cost function, which itself can be derived from a more abstract cost functional (cf. Sec. 3.1.4).

The prior is a probability distribution, and so we assume semi-positivity and a normalization,

$$\boldsymbol{x} \in \boldsymbol{X}: \quad \Pi(\boldsymbol{x}) \geq 0 \quad \text{and} \quad \sum_{\boldsymbol{x}} \Pi(\boldsymbol{x}) = 1, \tag{3.7}$$

that is, Π is a probability distribution.

The prior is a measure for the probability of a hypothesis. For two hypotheses $\boldsymbol{x}^a$ and $\boldsymbol{x}^b$, we say $\boldsymbol{x}^a$ has a higher probability, or is more favorable than $\boldsymbol{x}^b$, if the following holds:

$$\Pi(\boldsymbol{x}^a) > \Pi(\boldsymbol{x}^b) \tag{3.8}$$

The magnitude of distance between two images is determined by the actual cost function implemented.

Likelihood The second ingredient to Bayesian estimation is the likelihood. It is a probability function that forms the bridge between the data that are actually measured, $\boldsymbol{y}$, and the hypotheses on the truth, $\boldsymbol{x}$. The likelihood then states the probability of a certain measurement given a hypothesis about the true value. For image data, the likelihood of $\boldsymbol{x}$ and $\boldsymbol{y}$ representing two full images can rarely be quantified sensibly, as the joint configuration space is vast. Instead, the problem is brought down to individual pixels and the pixels in their neighborhood. Beyond the neighborhood area, pixels are then considered independent and the likelihood can be factored into contributions for each spatially independent pixel configuration.

With the likelihood describing the probabilities for a given hypothesis of the true image, i. e., for a hypothetical physical situation, an access to the problem with additional benefit is opened.

The likelihood is used to describe how a (hypothesized) true image could be linked to its deteriorated measurements. This approach is “inverse” to the original setting of the estimation problem, and constraints are often much easier to quantify if they can be based on a hypothesis.

The degradation of electronically recorded images is an example: with a likelihood function, the effect of noise sources on this measurement process can be accounted for in its full combination, when only the physical background of the processes is known.

When the likelihood and the prior are formulated, the major part of the image reconstruction problem is defined and the a posteriori probability can be calculated.

As a next step, a cost function and a corresponding estimator for the a posteriori probability are discussed, cf. Sec. 3.1.4.

The configuration space for images is extremely large, which imposes a serious problem on the effort to draw samples from a probability distribution. When subsampling with a uniform probability, most configurations drawn will have a very small a posteriori probability. Estimators will easily become biased or very volatile, strongly influenced by the few (if any) samples of high weight. We discuss ways how to efficiently generate samples with high a posteriori probability in Sec. 3.2.

The approaches available for efficient sampling restrict the possible choices for prior probabilities.

A central issue for the design of image reconstruction priors is to adjust the balance between the smoothing of noisy pixels and the preservation of edges and solitary image features. In addition, efficient sampling from the posterior probability imposes its own restrictions on the range of possible priors, which will also be part of our focus in the detailed discussion of priors in Sec. 3.3.

3.1.4. Cost functions and a posteriori estimators

A mathematical description of the ideas in Sec. 3.1.2 can start with a *cost function*. It gives us a measure and a simple tool at hands, with which to decide which of the possible realizations is a more "reasonable" choice. This is done by fixing a cost value for each decision and, as a general paradigm, the decision should be made so that minimum costs are achieved [Moon and Stirling, 2000].

From the mathematical point of view, a cost function is a measure, which convert the complex differences between images $\boldsymbol{x}^a$ and $\boldsymbol{x}^b$ into a plain number. Mathematically, we require that an elementary cost function $\Gamma : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}^+$ fulfills the following criteria [Chalmond, 2003].

1. The cost should be commutative and positive semi-definite:

$$\Gamma(\boldsymbol{x}^a, \boldsymbol{x}^b) = \Gamma(\boldsymbol{x}^b, \boldsymbol{x}^a) \geq 0 \tag{3.9}$$

2. The cost should be zero if and only if the images are equal:

$$\Gamma(\boldsymbol{x}^a, \boldsymbol{x}^b) = 0 \iff \boldsymbol{x}^a = \boldsymbol{x}^b \tag{3.10}$$

Costs are usually calculated in reference to the truth, which is unfortunately unknown for the Bayesian estimation problem. This issue can be circumvented by the introduction of *Bayesian costs*, which are the average costs of all posterior settings in the light of a measurement $\boldsymbol{y}$, weighted with the associated probability:

$$\tilde{\Gamma}(\boldsymbol{x}) = \sum_{\boldsymbol{x}' \in \boldsymbol{X}} \Gamma(\boldsymbol{x}, \boldsymbol{x}') \mathbb{P}(\boldsymbol{x}' | \boldsymbol{y}) \tag{3.11}$$

As mentioned in the introduction, the best estimate is axiomatically postulated as that of minimum cost:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\boldsymbol{X}} \tilde{\Gamma}(\boldsymbol{x}) \tag{3.12}$$

In the following paragraphs, we discuss the estimators which are connected to three cost settings frequently used.

Maximum a posteriori estimate (MAP) A simple setting of the cost function imposes constant costs on any estimate $\hat{\boldsymbol{x}}$ that is different from the best choice $\boldsymbol{x}^{\text{best}}$, independent from the number of erroneous pixels or how large the differences are:

$$\Gamma(\hat{\boldsymbol{x}}, \boldsymbol{x}^{\text{best}}) = \begin{cases} 0 & \text{if } \hat{\boldsymbol{x}} = \boldsymbol{x}^{\text{best}} \\ 1 & \text{otherwise} \end{cases} \tag{3.13}$$

The Bayesian costs then are

$$\begin{aligned} \tilde{\Gamma}(\boldsymbol{x}) &= \sum_{\boldsymbol{x}'\in\boldsymbol{X}} \Gamma(\boldsymbol{x}, \boldsymbol{x}')\, \mathbb{P}(\boldsymbol{x}'|\boldsymbol{y}) && (3.14)\\ &= \sum_{\boldsymbol{x}'\in\boldsymbol{X}} \mathbb{P}(\boldsymbol{x}'|\boldsymbol{y}) - \mathbb{P}(\boldsymbol{x}'|\boldsymbol{y})|_{\boldsymbol{x}'=\boldsymbol{x}} && (3.15)\\ &= 1 - \mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) && (3.16) \end{aligned}$$

The cost minimization postulate corresponds to a maximization of the a posteriori probability:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\boldsymbol{X}} \{1 - \mathbb{P}(\boldsymbol{x}|\boldsymbol{y})\} = \arg\max_{\boldsymbol{x}\in\boldsymbol{X}} \mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) \tag{3.17}$$

Under this cost definition, the best estimate is found by choosing the image $\boldsymbol{x}$, which has a maximum a posteriori probability. This estimator is therefore called *maximum a posteriori* estimator or short, MAP.

This estimator has one obvious shortcoming, as it can be extremely hard to find the best estimate: By using only the cost definition in Eq. (3.13), we have no way of knowing which one of two wrong estimates is nearer to the best choice. Algorithms which evolve one estimate out of an other would search the state space without knowing how far off the current estimate is. In fact, this problem has a setting equivalent to the *maximum clique* problem [Flach, 2002], which has been proven to be NP-complete.

Nevertheless and albeit the reasoning behind its cost definition is rather plain, the MAP estimator is well-accepted and popular in application [Hunt, 1977], [Geman and Geman, 1984], [Bouman and Sauer, 1993], [Medeiros et al., 1998].

Marginal posterior mode estimate (MPM) In some cases, finding the minimum-cost estimate for the complete scene is not required. Instead, it may be sufficient to use a minimum-cost estimate which covers only a particular site and the sites within its proximity:

$$\Gamma(\hat{\boldsymbol{x}}, \boldsymbol{x}^{\text{best}}) = \sum_{s\in S} \Gamma_s(\hat{x}_s, {x_s}^{\text{best}}) \tag{3.18}$$

When defining the costs on a per-site basis, we again set it to a constant contribution for each deviation from the best value, regardless how large the error is. The influence of the surroundings becomes only effective with the a posteriori probability, which is influenced by the prior:

$$\Gamma_s(\hat{x}_s, {x_s}^{\text{best}}) = \begin{cases} 0 & \text{if } \hat{x}_s = {x_s}^{\text{best}} \\ 1 & \text{otherwise} \end{cases} \tag{3.19}$$

The Bayesian costs can be accumulated as the sum over all sites:

$$\tilde{\Gamma}(\boldsymbol{x}) = \sum_{s \in S} \tilde{\Gamma}_s(x_s), \tag{3.20}$$

with the following contribution from each site:

$$\begin{aligned} \tilde{\Gamma}_s(x_s) &= \sum_{x'_s \in X_s} \Gamma_s(x_s, x'_s)\, \mathbb{P}(x'_s, \boldsymbol{y}) && (3.21) \\ &= 1 - \mathbb{P}(x_s|\boldsymbol{y}) && (3.22) \end{aligned}$$

By the semi-positivity of $\tilde{\Gamma}_s$, minimization of Eq. (3.12) can again be achieved site-wise. Analogous to Eq. (3.17), this yields the site-wise estimate:

$$\hat{\boldsymbol{x}} = (\hat{x})_s \quad \text{with } \hat{x}_s = \arg\max_{x_s \in X_s} \mathbb{P}(x_s|\boldsymbol{y}). \tag{3.23}$$

This estimator is called *marginal posterior mode*, or short, MPM estimator. The Bayesian estimation problem can be worked through independently for each site with its neighborhood, which however in practice does not imply a significant improvement in computability. In Chap. 4 we will discuss the direct computability for this estimator with a certain class of priors.

Minimum mean squares estimate (MMS) For this estimator, the costs are not constant as with the others so far, but chosen as the quadratic distance per site. This leads to costs that grow larger with distance from the best image. Similar to the MPM estimate, we can split up the costs for a complete image into single-site costs:

$$\Gamma_s(\hat{x}_s, {x_s}^{\text{best}}) = \left(x_s - {x_s}^{\text{best}}\right)^2 \tag{3.24}$$

The Bayesian costs for each site s then are:

$$\tilde{\Gamma}_s(x_s) = \sum_{x'_s \in X_s} (x_s - x'_s)^2\, \mathbb{P}(x'_s|y_s) \tag{3.25}$$

The costs in Eq. (3.25) should be minimum. We choose to locate that minimum analytically:

$$\begin{aligned} \hat{x}_s &= \arg\min \tilde{\Gamma}_s(x_s) && (3.26) \\ \frac{\partial}{\partial x_s} \tilde{\Gamma}_s(x_s) &\stackrel{!}{=} 0 && (3.27) \end{aligned}$$

The differentiation leads to the following necessary, and also sufficient condition for the optimum, cf. [Winkler, 2003]:

$$x_s \, \mathbb{P}(x_s|y_s) = \hat{x}_s \, \mathbb{P}(\hat{x}_s|y_s) \tag{3.28}$$

With the last identity we get for the minimum mean squares, or MMS a posteriori estimate:

$$\hat{\boldsymbol{x}} = \sum_x \hat{x}_s \, \mathbb{P}(\hat{x}_s|y_s), \tag{3.29}$$

This formula is actually just like Eq. (3.11), the definition of Bayesian costs. That is why MMS estimators are the Bayesian (minimum-risk) estimators for the case of additive, quadratic costs.

3.1.5. Deterministic approaches

In Sec. 3.1 we have discussed that the image reconstruction problem in general is ill-posed and any brute-force approach is severely hampered by the sheer enormity of the configuration space that needs to be searched.

Still however for some settings of the problem, specialized approaches have been developed. In contrast to the stochastic methods discussed in Sec. 3.2.2, these are deterministic in the sense that they need no sampling by a random process for operation.

The algorithms discussed in this section are capable of finding the optimum solution, or the minimum costs setting of the image restauration problem within finite time. Likewise, the approach we propose for height map estimation (Chap. 4) is deterministic and bound to find the optimum configuration within the limits of its specialized settings.

Analytical integration With certain settings for the probability distributions of prior and likelihood, a Bayesian estimation problem can take a mathematical form that allows for a analytic integration of the posterior. Thus, the a posteriori probability density can be obtained directly.

In [Kendall et al., 1987], a number of examples for such "conjugate" settings can be found:

- Gaussian likelihood + Gaussian prior $\longrightarrow$ Gaussian posterior
- Binomial likelihood + Beta function prior $\longrightarrow$ Beta function posterior
- Poisson likelihood + Gamma prior $\longrightarrow$ Gamma posterior

The central drawback of this solution is of course the restriction of the probability distributions to certain settings. For some applications, the choice of the prior probability is to a certain degree discretionary, mostly justified by lack of prior knowledge on the problem.

But as the functional form of a prior directly influences the a posteriori probability of an estimation problem, a prior probability function should not

be chosen simply for the sake of a less complex computability—at least not unless its consequences on the estimation problem are understood. For the height estimation prior that is introduced in Chap. 4, we will discuss this issue in Sec. 4.2.3).

Ford-Fulkerson The Ford-Fulkerson algorithm is a well known algorithm in network theory [Mehlhorn, 1984]. It is used to solve the *maximum flow problem*, which addresses the question of how to maximize the flow between a given source and sink through a network of edges with given capacity.

To carry this over to image reconstruction, an image model must be converted into a flow network (a directed graph). For this purpose, for each pixel a node is set up; two more nodes are added that serve as source and sink. The source is connected to each node, the same holds true for links from the pixel nodes to the sink. The capacities of these links are chosen according to log-likelihoods of the observations. The a priori knowledge sets the interdependence of neighboring pixels which is represented by two links connecting back and forth between two neighbors. Their capacities are set as the coefficient of the model describing the pair interaction of the neighborhood, cf. Eq. (3.55). The Ford-Fulkerson algorithm calculates the maximum flow through this network, which is the MAP estimate of the original problem.

The Ford-Fulkerson algorithm has however some limitations concerning its application to image reconstruction:

- The prior must be of a generalized Ising type (cf. Sec. 3.2.1), i. e. it can be expressed as a sum of pair potentials.
- Edge models (line processes) for robust reconstruction (cf. Sec. 3.3) can only be be incorporated into the Ford-Fulkerson framework if they can be represented as a pair-potential.
- The computation requires many operations on the computer, which makes handling of larger images intricate.

Within image processing, the Ford-Fulkerson algorithm is therefore used rarely in actual applications and mostly serves for theoretical considerations.

GNC algorithm The GNC algorithm (Graduated Non Convexity) was invented by Blake and Zisserman and is a minimization algorithm for the *weak membrane model*. It incorporates an explicit edge model for edge preservation, cf. Sec. 3.3.2. The energy function (Hamiltonian) H of this model takes the following form [Winkler, 2003]:

$$H = \sum_{s \sim t} \phi(g_s - g_t) + \sum_s (y_s - g_s)^2 \tag{3.30}$$

In this formula, s and t are pixel indices, g_s and g_t denote gray values and y_s is the data term (the observed image), $s \sim t$ denotes a neighborhood relationship between s and t.

During the optimization, the roughness penalty ϕ in the Hamiltonian is approximated by convex functions. These gradually approach the original functional values by variation of an overall parameter p during the course of optimization, hence the name GNC. With this approximation one tries to avoid dropping into local minima of the energy function. That way, the GNC algorithm shares its principal ideas with the simulated annealing approach on Markov random fields (cf. Sec. 3.2.2), but it has the drawback that its energy functions are restricted to the form of Eq. (3.30). An in-depth discussion can be found in [Blake and Zisserman, 1987].

3.2. Bayesian estimation with Markov random fields

In the last section, we have discussed how to set up the prior and likelihood for Bayesian estimation, and how the cost function for an appropriate a posteriori estimator can be selected.

Estimates $\mathcal{E}$ for functions f derived from the a posteriori probability can be calculated according to the general formula

$$\mathcal{E}\{f(\boldsymbol{x})\} = \sum f(\boldsymbol{x})\mathbb{P}(\boldsymbol{x}|\boldsymbol{y}) \tag{3.31}$$

In the strict sense, the sum should be evaluated for the full configuration space $\boldsymbol{X}$, which has a cardinality of $z^{x_{\max} y_{\max}}$, z the number of gray levels. Even for simple binary images, the associated state space is overwhelmingly large (cf. introduction to Sec. 3.1). But not only the space contains far too many elements to be handled, but also most configurations are of very small a posteriori probability and contribute little to the estimate, as they do not resemble the true image at all.

It is therefore necessary to reduce the number of samples drawn. If we can restrict ourselves for a given number of samples to the ones with a high probability, the error for the estimator can be reduced, as only samples of low probability are left away. This "intelligent" sampling is usually referred to as *Monte Carlo* sampling [Hastings, 1970]. Still, the sampling method itself can be optimized, by the introduction of *dynamic* Monte Carlo or *simulation* methods, which in the course of the simulation gradually move the sampling towards regions of higher probability and there let it "settle in".

Markov models and in particular Markov random fields (MRFs) give a solution to the—actually very general—problem of sampling from high-dimensional random distributions. Markov methods are used in many scientific fields, such as speech recognition, tracking and control problems or archeometrics, cf. also [Besag et al., 1995].

3.2.1. Markov random fields

In this section, we will discuss how to draw samples from a Markov random field. To that end, we first consider the properties of random fields, then the sampling from a Markov chain—the one dimensional precursor to Markov fields—in the

next paragraphs. After the Markov chain properties, the questions of setting the burn-in period and reliable sampling, we discuss Markov random fields with the Ising/Potts example.

For many applications built upon Markov random fields, the efficiency of sampling becomes the next central optimization issue. We discuss solutions to this issue, also referred to as *simulation methods*, in Sec. 3.2.2. For some methods and settings of parameters, as is the case with simulated annealing, convergence can be strictly proven, in other cases this is not possible.

Random fields

A random field is a set of sites that each can be in a certain configuration, with a probability distribution defined for any configuration of the set.

In a more mathematical formulation, we take the set of sites S and a configuration space $\boldsymbol{X}$ for this set as defined in Sec. 3.1.1. A probability distribution Π on $\boldsymbol{X}$ is then defined as $\Pi = (\Pi(\boldsymbol{x}))_{\boldsymbol{x}\in\boldsymbol{X}}$ with the restriction $\Pi(\boldsymbol{x}) \geq 0$ and the normalization $\sum_{\boldsymbol{x}\in\boldsymbol{X}} \Pi(\boldsymbol{x}) = 1$ that make it a probability measure.

If Π is strictly positive, that is, for all configuration $\boldsymbol{x} \in \boldsymbol{X}$ we have $\Pi(\boldsymbol{x}) > 0$, then $(\boldsymbol{X}, \Pi)$ is called a *random field*.

Neighborhoods and cliques Prior knowledge on a random field is usually available only within small neighborhoods of a site. An example: For an image we have a common idea about the property "smoothness". However, we cannot deduce the gray value setting in one corner of the image from that in another, far away corner. Instead, the notion of smoothness can only be expressed on a local scale, this could be by assigning similar gray values of pixels in the spatial proximity of a certain pixel a higher a probability than gray values farther apart.

Intuitively, neighbors are sites "next" to each other, and a neighborhood is the set of those sites, and cliques can be seen as groups formed within a neighborhood.

Moving over to a mathematical notation, a *neighborhood system* for sites s is defined as a collection ∂ of sets $\partial(s) \equiv \{\partial\{s\} : s \in S\}$ with the following properties:

(a) $s \notin \partial\{s\}$ and

(b) $s \in \partial(t)$ if and only if $t \in \partial(s)$.

In that case all $t \in \partial(s)$ are *neighbors* of s. We also use the notation $t \sim s$ if t is a neighbor of s, then of course s is a neighbor of t. To express the set of sites that contains the neighborhood for a site x_s, we write the shorthand $\partial(x_s)$.

A subset C of the set of sites S is called a *clique*, if all elements of C are neighbors: $C \subseteq S : s, t \in C : s \sim t$. Given an appropriate neighborhood system, it is also possible that the empty set $\emptyset$ and S itself are part of the clique.

All common imaging systems have a rectangular pixel pattern. Due to the symmetry of this pattern, the *4-neighborhood* and the *8-neighborhood* are particularly important, cf. Fig. 3.1.

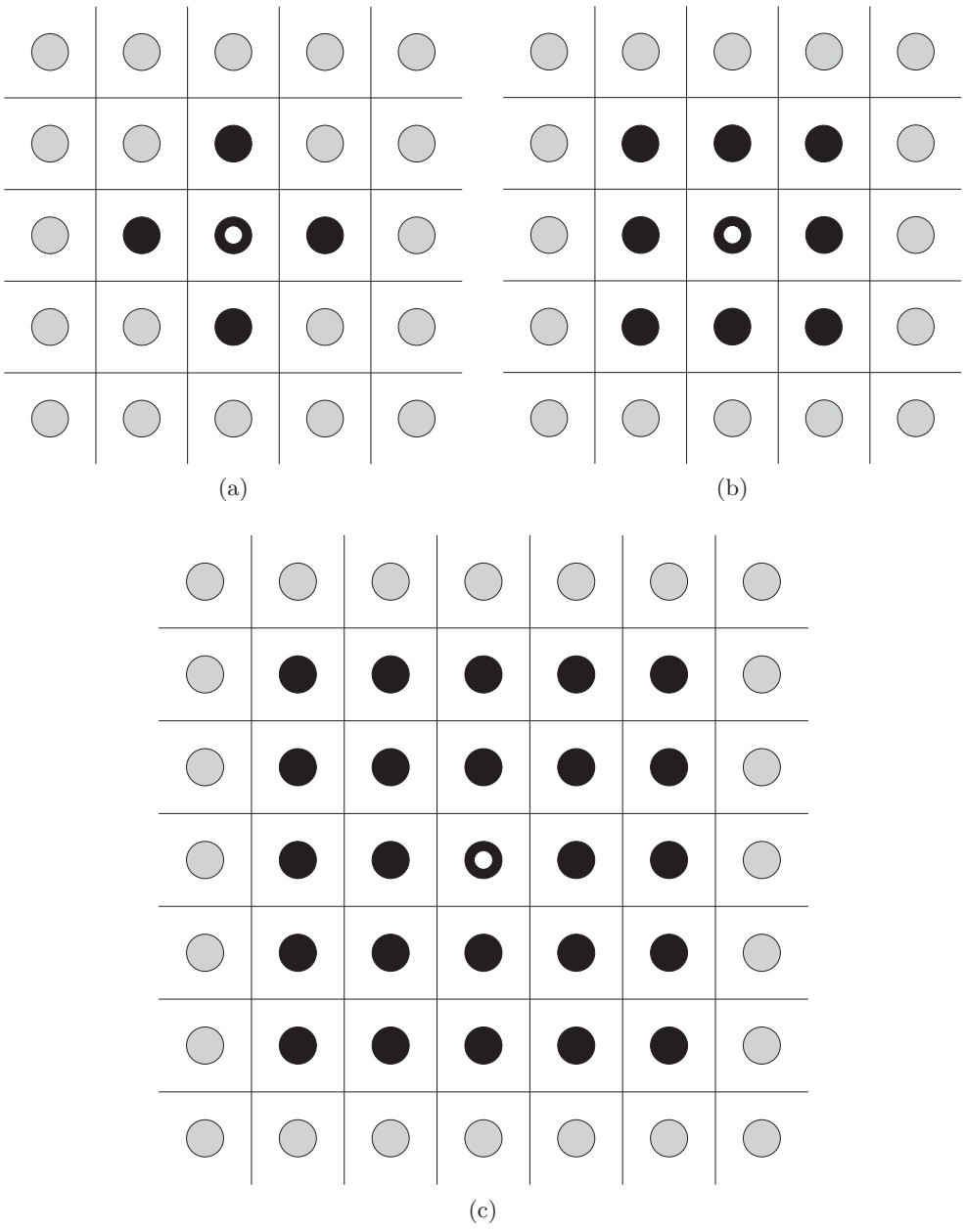


Figure 3.1: Examples of symmetric neighborhoods for the rectangular pixel pattern: (a) 4-neighborhood, (b) 8-neighborhood and (c) 24-neighborhood. The reference (center) site s is marked by a white dot.

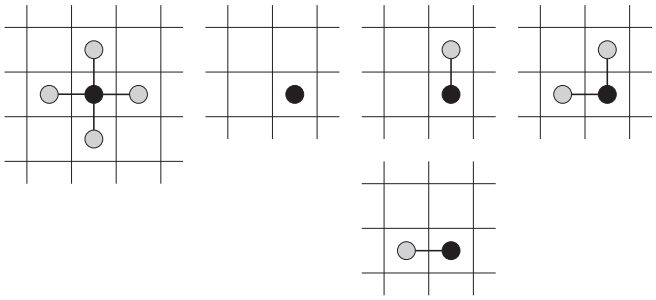


Figure 3.2: Neighborhood relationships (lines) and cliques for the 4-neighborhood case (cf. Fig. 3.1a). The rotated variants are omitted.

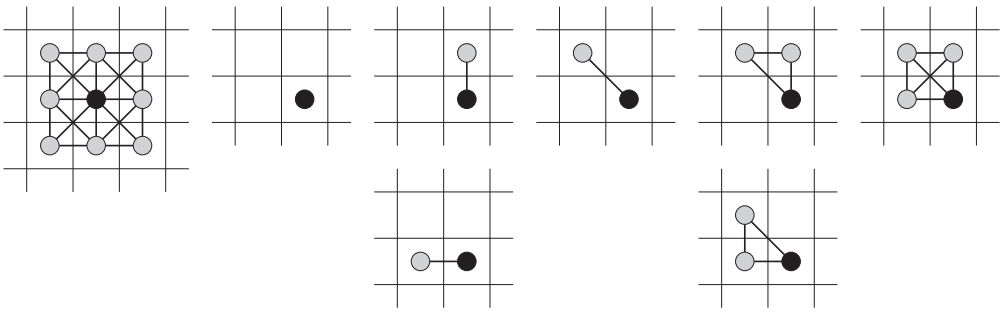


Figure 3.3: Neighborhood relationships (lines) and cliques for the 8-neighborhood case (cf. Fig. 3.1b). The rotated variants are omitted.

The cliques that correspond to their neighborhood system for a particular site s are shown: in Fig. 3.2 for the 4-neighborhood, and in Fig. 3.3 for the 8-neighborhood. The empty set $\emptyset$ clique and the symmetrically rotated cliques are left away from the illustration. Still it is visible that the number of clique grows fast with increasing size of the neighborhood.

Handling of image boundaries In the proximity of the outer image boundaries, neighborhood relationships fail and exceptions have to be introduced. A common option is to first extend the image domain by half of the filter size for the boundary pixels. Then the gray values of the additional pixels could be set to zero, or all copied from the former outermost pixel, which is usually a smoother transition. Another possibility is to mirror the gray values at the border. The Fourier transform of an image can only be strictly invertible if the image is continued at the border of the other side, which, applied for both axes, effectively puts the image around a torus [Jähne, 2002]. As the original scene is usually not periodical, this approach often leads to stronger high-frequency artifacts in the Fourier domain.

Finally, a rather simple option we adopt for the filtering procedures discussed in Chaps. 4 and 5, is to process only those pixels for which the filter mask does not cross the border. Therefore the output image is reduced by half the filter mask size along each border. This however is not a viable option if large filters

have to be applied repeatedly, as the usable area shrinks in each iteration.

Markov processes

Markov random fields are derived from Markov processes, also known as the *Markov chain model*. A Markov process is defined on sequential data, like time series, for which transition probabilities from one state to the other are given. Markov processes are a tools for modeling a sequential development ruled by a statistical law.

Markov chains Let us first review the one-dimensional *Markov chain* model: A Markov chain of first order describes the time development of a random variable X_i and its realization x_i, with $i \in \mathbb{N}$ being the time index.

By definition, the random variable fulfills the (first-order) Markov property, if the following holds true:

$$\mathbb{P}(x_i|x_{i-1}, x_{i-2}, \ldots, x_1) = \mathbb{P}(x_i|x_{i-1}) \tag{3.32}$$

The probability distribution at time i therefore may only depend on the distribution at the latest time step, $i-1$. Higher order Markov processes have dependencies also back to distributions earlier in time. In practice, these can only be found in special applications, as higher order interrelations are often difficult to model.

The evolution of a Markov process can be described by the associated *transition probability matrix* or *Markov kernel* $\mathbb{P}$, which specifies the transition probabilities from any state to another:

$$x \in X,\ \mathbb{P} : X \times X \to [0,1] :\ \mathbb{P}(x, \cdot) \tag{3.33}$$

A Markov transition can then be described by the probability to move from any state x in the present distribution ν to state y; the right hand side gives a common shorthand:

$$\sum_x \nu(x)\, \mathbb{P}(x, y) \equiv \nu\mathbb{P}(y) \tag{3.34}$$

A *homogeneous Markov process* is defined as one with a constant kernel. A Markov kernel $\mathbb{P}$ is *primitive*, if any y can be reached from any x within a finite number of transitions:

$$\mathbb{P}^k(x, y) > 0 \quad \text{with } k \in \mathbb{N}, \text{ and for any } x, y \tag{3.35}$$

In this notation, multifold transitions are marked with the exponent to the kernel $\mathbb{P}$. A Markov process is *aperiodic*, if there is only one way to come from a distribution back to the same, and finally, a Markov process is *positive recurrent*, if in the limit of infinite cycles, any state is revisited with unit probability.

The practical impact of Markov processes arises from the *limit theorem*: For a primitive Markov kernel $\mathbb{P}$ and its positive recurrent, aperiodic process x on

a finite state space X, there exists a limit distribution and the Markov process converges to this invariant distribution:

$$\text{with } \mu\mathbb{P} = \mu : \quad \lim_{n\to\infty} \nu\mathbb{P}^n = \mu \tag{3.36}$$

With this theorem, we know that a Markov process eventually can give samples from a distribution very close and approaching to its invariant distribution. So if we are able to define a Markov process for the probability distribution we want to sample from, evolving the Markov process gives us the desired samples for the calculation of estimates.

By the law of large numbers, reliable estimates of arbitrary functions can be calculated from all samples produced by the Markov process: To make use of this in practice however, a Markov sampler which starts from some initial distribution should be let run for a certain time (*burn in*), then samples can be drawn. This avoids unnecessary bias on the estimates while the distribution is yet converging. Still it is often unclear, when the burn in period ends and when sampling can begin.

The abstract procedure is therefore as follows:

1. Find a Markov kernel $\mathbb{P}$ which describes the system,

2. input an initial distribution into the Markov process and let it "burn in" towards the invariant distribution,

3. approximate the MMS estimator of functions by its empirical average, obtained from sampling a long, partial realization of the process.

Markov random fields

Transition to random fields The idea of extending the notion of Markov processes to probability distributions of dimensions higher than one, i. e. the Markov chain, was already discussed in early years [Dobruschin, 1968], its powerful application to image restoration became widely recognized with the seminal paper by Geman & Geman, [Geman and Geman, 1984].

The central difference between Markov processes and Markov random fields lies the domain on which they "live" and the way how to evolve new states: Markov random fields are defined on random fields, which can be two- or higher dimensional. What makes Markov random fields of particular interest for us is that a planar image or height map can be modeled as a Markov random field.

In a Markov process, the dependence of a new state is based on earlier states and the development itself is also in time. On a Markov random field however, there is no natural notion of direction as in a sequence. In this case, we also develop the field in time, but consider the dependence of a new state only on each site's neighboring sites at the same instant of time. Plainly speaking, "time indices are considered as spatial indices" [Li, 2001b].

Local characteristics For the discussion of Markov properties for random fields, it is helpful to define *local characteristics* (also known as "local specifications" [Chalmond, 2003] or "full conditionals" [Winkler, 2003]), which are the local conditional probabilities for a neighborhood system $\partial(\boldsymbol{x})$:

$$\Pi(x_s|x_t, \text{ with } t \in \partial(s)) \qquad \text{local characteristics,} \tag{3.37}$$

which is actually a shorthand form for the probability of configuration $\mathcal{X}_s = x_s$ at site s under the conditions that $\mathcal{X}_t = x_t$ for all t in $\partial(s)$.

Handling random fields by their local characteristics is useful for Gibbsian form estimators (see below), which contain a partition function.

Markov property for random fields A *Markov random field* is defined as a random field Π with a neighborhood system $\partial(\boldsymbol{x})$ if for any two sites $s, t \in S$ and all configurations $\boldsymbol{x} \in \boldsymbol{X}$ the following holds true for the local characteristics:

$$\Pi(x_s|x_t, \text{ with } s \neq t) = \Pi(x_s|x_t, \text{ with } t \in \partial(s)) \tag{3.38}$$

That is, any site t' outside of $\partial(s)$ may not influence the conditional probability distribution for s on a Markov random field. Its properties do only depend on the configuration in its neighborhood.

The Markov property can also be formulated as follows: the probability laws $\mathcal{X}_{\{s\}}$ (i. e. for the set of site s) and $\mathcal{X}_{\{t:t\notin\partial(s)\}}$ are *conditionally independent* under a given $\mathcal{X}_{\partial(s)}$, cf. [Lauritzen, 1996].

While for Markov processes each state only depends on its immediate precursor, for a Markov random field each site only depends on the states of a site and its neighbors (cf. Eq. (3.38)).

Gibbs fields and Gibbs-Markov-equivalence The *Gibbs formula* gives a physically motivated relation between an energy or *potential* U[1] and a probability law or density function (see box below for this motivation):

$$\Pi(\boldsymbol{x}) = \frac{1}{Z}\, e^{-\beta U(\boldsymbol{x})} \tag{3.39}$$

Here $U(x)$ is the potential (energy) for a configuration $\boldsymbol{x}$, β a tunable parameter and Z a normalization parameter (partition function):

$$Z = \sum_{\boldsymbol{x}} e^{-\beta U(\boldsymbol{x})} \tag{3.40}$$

By specifying the potential $U(\boldsymbol{x})$ for a configuration $\boldsymbol{x}$, the corresponding probability $\Pi(\boldsymbol{x})$ can be found. If the potential U is a *neighborhood potential*, i. e. if it can be partitioned into cliques:

$$U(\boldsymbol{x}) = \sum_{c\in C} V_c(\boldsymbol{x}), \tag{3.41}$$

then the probability Π defines a *Gibbs random field*.

[1] We choose the symbol U instead of H as used in some places earlier in this chapter, as we want to reserve U for potentials that are neighborhood potentials, cf. Eq. (3.41).

Physical motivation for Gibbs potentials [Reif, 1965]

The Gibbs potential is the probability distribution of the canonical ensemble. This model is characterized by a constant average energy for its statistical ensemble.

The energy of the whole ensemble is E, summed from small energy contributions of individual elements (states). If we look at one particular contribution s with energy V_s, we therefore have:

$$V_s \ll E \tag{3.42}$$

The probability for this state is the number of states available, divided by the total number of states. The latter equals to the microcanonical sum for the ensemble energy, $\Omega(E)$. If the particular element had energy V_s, the remaining ensemble had $\Omega'(E-V_s)$ states available. Therefore, the element itself has $\Omega'(E-V_s)$ states available. We get for the probability:

$$\Pi(s) = \frac{\Omega'(E-V_s)}{\Omega'(E)} = \frac{1}{C}\,\Omega'(E-V_s) \tag{3.43}$$

with $C = \Omega'(E)$ a constant, as the overall energy E is fixed.

A series expansion for $\ln \Omega'(E-V_s)$ around E leads to

$$\ln \Omega'(E-V_s) = \ln \Omega'(E) - \frac{\partial}{\partial V_s} \ln \Omega'(E) V_s + \ldots \tag{3.44}$$

We ignore terms of higher order and use the definitions of entropy ($S = k_B \ln \Omega(E)$) and temperature ($T^{-1} = \partial S(E)/\partial E$) to get:

$$\ln \Omega'(E-V_s) = \ln \Omega'(E) - \frac{1}{k_B T} V_s \tag{3.45}$$

This we put in the expression for the probability to finally get:

$$\Pi(s) = \frac{1}{Z}\, e^{-\beta V_s} \tag{3.46}$$

with the partition function Z as the new normalization constant, to be calculated like $Z = \sum_s e^{-\beta V_s}$ and the inverse temperature $\beta = (k_B T)^{-1}$.

The formula Eq. 3.39 has a large practical importance for Markov random fields by several aspects:

(a) there exists an exact correspondence between the definition of a random field by its Gibbs or by its Markov properties (Hammersley-Clifford theorem, see below),

(b) a problem can often be better described by neighborhood potentials, as we show in the example given at the end of this chapter,

(c) the additional parameter β allows for efficient simulation strategies, like simulated annealing (cf. Sec. 3.2.2) and others.

The parameter Z is the normalization of the right hand side of Eq. (3.39) that makes it a probability. The parameter β is known as the *inverse temperature*, or $\beta = (kT)^{-1}$. In the limit of lowest inverse temperature, $\beta \to 0$, the right side of Eq. (3.39) vanishes, leaving a uniform probability distribution. In the high-β limit the exponential distribution becomes arbitrarily steep, leaving only the lowest energy configuration with non-zero probability. This situation has the physical correspondence of "freezing" when a system comes to zero temperature.

The Hammersley-Clifford theorem[1] proves the Gibbs-Markov equivalence: If the Gibbs field Eq. (3.39) has a potential H that is a neighborhood potential (cf. Eq. (3.41)), then it is a Markov random field for exactly this neighborhood.

Consistency of Markov random fields In the course of the proof for the Hammersley-Clifford theorem, one also obtains a formula relating the local characteristics to the Gibbs energy specified by neighborhood potentials:

$$\Pi(x_s|\partial(x_s)) = \frac{1}{z}\, e^{-U(\boldsymbol{x})+U(\boldsymbol{x}_{-s})} \tag{3.47}$$

with z the partition function for a single site,

$$z = \sum_{x_s} e^{-U(\boldsymbol{x})+U(\boldsymbol{x}_{-s})}, \tag{3.48}$$

with $\partial(x_s) = \{x_t\}$ with $t \in \partial(s)$ and $\boldsymbol{x}_{-s} = \{x_1, \ldots, x_{s-1}, x_{s+1}, \ldots, x_{|S|}\}$. With this formula, the two equivalent formulations of a Markov random field can be transformed into each other.

One can see that in order to establish the Markov property for a Gibbs field, it is necessary that the energy of a Gibbs field can be separated into independent contributions $U(\boldsymbol{x})$ (cf. Eq. (3.41)) and $U(\boldsymbol{x}_{-s})$, the energy for the neighborhood of s, excluding s itself. This is a prerequisite for a system of neighborhood potentials; if the separation is not possible, the field cannot be Markovian.

The model for height estimation we elaborate on in Chap. 4 is unfortunately not Markovian—at least for a locally restricted neighborhood definition—as we will discuss also in Chap. 5.

[1] It remained unpublished, see also [Grimmett and Welch, 1990]; in the former eastern world an early proof was given in [Averintsev, 1972].

As a side note, a general construction scheme how to set up a Markov field's local characteristics from neighborhood potentials is presented in a proof for the Hammersley-Clifford theorem in [Besag, 1974].

Example: The Ising/Potts models The *Ising model* is a simple, yet powerful system based on a Markov random field. While its original version assumes binary states for each site, an extension to a higher number of discrete states in known as the *Potts model*.

The Ising model was originally devised by E. Ising in 1925, independent of Markov theory, as an approach for the theoretical understanding of ferromagnetism (cf. [Winkler, 2003] for historical remarks). In this model, it was assumed that each atom in the ferromagnetic crystal can take either a "spin up" or "spin down" configuration, $S = \{-1, 1\}$, to entirely represent its magnetic properties. It is assumed that the forces between neighboring atoms are such that similar settings (co-linearity of the spins) are favorable over dissimilar settings. This is translated into a energy difference and incorporated in the following *pair potential*:

$$U(x_s, x_t) = -\beta\, x_s\, x_t \tag{3.49}$$

The parameter β is again the "inverse temperature" with $\beta = (kT)^{-1}$, k the Boltzmann constant.

Under exposiure to an external magnetic field B (a scalar here, we consider only one direction for all fields), the ferromagnetic material favors a spin configuration co-linear to that field. The model can then be further augmented by an external potential:

$$U_{\text{ext}}(x_s) = -\mu B\, x_s \tag{3.50}$$

Together with the potentials for single sites we obtain the energy functional, or Hamiltonian, for the complete system:

$$\begin{aligned} H(x) &= U_{\text{pair}} + U_{\text{ext}} && (3.51)\\ &= -\beta \sum_{s\sim t} x_s x_t - \mu B \sum_s x_s && (3.52) \end{aligned}$$

We want to calculate the corresponding local characteristics with Eq. (3.47). To that end, we split up the energy into clique-wise contributions, we use a 4-neighborhood. The difference $U(\boldsymbol{x}) - U(\boldsymbol{x}_{-s})$ has a contribution for the single-site clique C_1 and the four pair-wise cliques C_2

$$U(\boldsymbol{x}) - U(\boldsymbol{x}_{-s}) = \underbrace{-\mu B x_s}_{C_1} \underbrace{-\beta \sum x_s x_t}_{C_2(s,t)} \tag{3.53}$$

We then immediately obtain the local characteristics:

$$\Pi(x_s|\partial(x_s)) = \frac{1}{z}\, e^{-\mu B x_s - \beta \sum_{C_2} x_s x_t} \tag{3.54}$$

As the energy of a site within Ising model can be fully spilt into contributions from cliques only, the Ising model is a consistent Markov random field.

The Bernoulli energy model [Chalmond, 2003] is actually very similar to the Ising model, Eq. (3.51), but uses the "spin" configurations $S = \{0, 1\}$ and then obeys to Bernoulli's probability laws as local characteristics.

If we further want to introduce directionality for the pair interaction, like when it should behave different along rows and columns of a 2-D image, this can be done with a small modification to the Ising model. The additional parameter $\theta_{s,t}$ changes the coupling for the horizontal and vertical component of the pair interaction:

$$U_{\text{pair}} = -\beta\, \theta_{s,t}\, x_s\, x_t \tag{3.55}$$

If it is chosen homogeneously negative ($\theta_{s,t} < 0$), an anti-linear setting of neighboring spins is encouraged.

The Potts model is an extension of the Ising model allowing more configurations between "up" and "down". In the Potts model, only neighbors of same state are assigned a low energy, any other setting is given a higher, but then equal energy. Using the discretized delta-distance ($\delta(a, b) = 0$ for $a = b$, otherwise $\delta(a, b) = 1$), in the absence of an external field this can be written like:

$$H(x) = -\beta \sum_{s \sim t} \delta(x_s, x_t) \tag{3.56}$$

The Ising/Potts models have been studied extensively as they can represent most features and particularities of Markov random fields. They are part of a broader class of Markov models, known as Besag's auto-models [Besag, 1974]. For up to pair-wise interactions, these models take the following general form:

$$H(x) = \sum_{s} x_s G_s(x_s) + \sum_{x \sim t} \beta_{s,t} x_s x_t, \tag{3.57}$$

where G_s is an arbitrary weighting function, and the $\beta_{s,t}$ give a weighting for any pair of neighbors s and t.

3.2.2. Stochastic sampling approaches

In contrast to Sec. 3.1.5, we here describe stochastic sampling methods that do not necessarily find the optimum solution or end after a certain time, but rather return approximations, often with bound errors. Theses methods are also known as *Monte Carlo* algorithms or Metropolis *sampler* or sometimes just *simulation* methods.

We will first describe the classical Monte Carlo algorithms, which can generate samples from an arbitrary statistical distribution. Afterwards we discuss the Markov chain Monte Carlo approach, which is an efficient way to sample from Markov random fields. Finally, we briefly touch the simulated annealing method as a modifications to this sampler.

Metropolis algorithm The original Metropolis algorithm, introduced to calculate chemical equations of state [Metropolis et al., 1953], is part of a class known as *(von Neumann) rejection algorithms*. The name *Monte Carlo sampler* hints towards its most-general purpose, which is to provide samples according to a probability distribution $\Pi(x)$. This algorithm is sometimes also named *Monte Carlo integration*, as it can be used to numerically integrate arbitrary functions. The Metropolis algorithm usually comes into action when the probability distribution is theoretically known, but cannot be evaluated in its general form or is otherwise too complex to draw samples from. Instead, a better accessible distribution Γ is used to provide samples, which are accepted by a ratio determined by the target distribution. The Metropolis algorithm is sketched in Fig. 3.4.

It can be shown that using this approach, the empirical distribution of x when it is chosen this way is the target distribution Π. For computational efficiency, a small rejection probability is critical, thus the distribution providing samples Γ should be as similar to the target distribution Π as possible; we assume $\Pi(x) \leq c\,\Gamma(x)$. In practice, this requirement can become difficult to fulfill, especially with very high-dimensional distributions. Finally, the normalization parameter κ provides optimal efficiency if it is chosen like $\kappa = \sum_x \frac{\Pi(x)}{\Gamma(x)}$ [Chib and Greenberg, 1995].

1 draw a sample x from Γ
2 evaluate $\Pi(x)$ for this very sample
3 generate a sample κ from the uniform distribution $\mathcal{U}(0,1)$
4 *acceptance/rejection step:*
5 **if** $c\,\frac{\Pi(x)}{\Gamma(x)} > \kappa$ **then**
6 | accept x
7 **else**
8 | reject x
9 | try another sample from Γ
10 **end**

Figure 3.4: Metropolis algorithm for rejection sampling.

Markov chain Monte Carlo / Gibbs sampler The Gibbs sampler is an approach for efficient sampling from a Markov random field. Its name goes back to the paper [Geman and Geman, 1984] and suggests that the samples are drawn from the local characteristics of the Gibbs field representation, cf. Eq. (3.47).

To that end, a Markov chain is defined which its transition probability (Markov kernel) $\mathbb{P}$ made up as a composition of transitions with the local characteristics for each site of the image. We here make use of the fact that Gibbs fields are reversible and invariant for their local characteristics [Winkler, 2003]. The local characteristics can be evaluated independently from the settings outside their respective neighborhoods.

The *visiting scheme* T defines a sequence for the sites s of the image:

$$T = \{s_1, \ldots, s_n\} \quad \text{with } n = |S| \tag{3.58}$$

One transition of the Markov kernel for the full image is the composition of transitions for each site, in the sequence as set in the visiting scheme T:

$$\mathbb{P}(\boldsymbol{x}, \boldsymbol{x}') = \Pi_{T_1} \cdots \Pi_{T_n}(\boldsymbol{x}, \boldsymbol{x}') \tag{3.59}$$

with Π_{T_i} designating the local characteristics for the gray value of site $T(i)$ given its neighborhood: $\Pi(x_i|\partial(x_i))$.

The Markov chain Monte Carlo algorithm starts with drawing a sample from an initial distribution ν. Within one transition of the composed Markov kernel Eq. (3.59), every site is updated: After a given site is updated from x_i to x_i', this new configuration is used for the update at the subsequent site. This is a valid Markov transition thanks to the localization of the $\Pi(x_i|\partial(x_i))$.

One can finally show that the Markov kernel Eq. (3.59) fulfills all requirements for the limit theorem (Sec 3.2.1, and Eq. (3.36)):

$$\lim_{n\to\infty} \nu\mathbb{P}^n(\boldsymbol{x}) = \Pi(\boldsymbol{x}) \tag{3.60}$$

With the last formula, we know that the Markov chain Monte Carlo algorithm samples from the Gibbs field. After the burn in period, or with a good guess as initial configuration, we can easily obtain samples for the Markov random field close to its actual probability distribution. This approach has the additional benefit that it is only required to calculate the local characteristics. Therefore we must only handle the probability distribution for each pixel and its respective neighborhood, but do not need to work with the configuration space of the full image.

Simulated annealing / Relaxation The parameter β (the inverse temperature) of the Gibbs field representation for a Markov random field (cf. Eq. (3.39)) can be used to tune the convergence of the sampling.

Then β is not kept fixed, but used as an additional parameter for the Gibbs sampler:

$$\Pi^\beta(\boldsymbol{x}) = \frac{1}{Z^\beta}\, e^{-\beta U(\boldsymbol{x})} \tag{3.61}$$

One can show that with increasing β the probability distribution Π^β concentrates more and more around its maximum modes, i. e., the global minima of the energy function $U(\boldsymbol{x})$ [Winkler, 2003].

The key to this is the choice of the *cooling schedule* which is a sequence of temperature settings $\beta(n)$. With each sweep of the Gibbs sampler, a different β can be used, so actually the Markov random field is inhomogeneous now.

The designation as "simulated annealing" stems from the analogy to the annealing of crystal lattices. A perfect lattice is the minimum-energy configuration. This optimum configuration can be reached by first heating up the

crystal which increases the mobility of defects to cross energy barriers, and then cooling it slowly down while it "relaxes" to the ground state.

With the following cooling schedule, Π^{β} "condensates" to its global optimum:

$$\beta(n) \leq \frac{1}{\Delta} \ln n, \tag{3.62}$$

here Δ is a scale parameter. The inverse temperature has to be increased logarithmically to ensure convergence. The original simulated annealing algorithm therefore requires a very long time to settle to the optimum.

In practice, for many problems also faster cooling schedules (like "exponential cooling") work reasonably, although convergence to the optimum is not strictly ensured. The ICM (Iterated Conditional Modes) algorithm [Besag, 1986] can be seen as a variant to simulated annealing with instantaneous cooling.

3.3. Robust priors and retaining of edges

The a priori probability is the function that gathers the prior knowledge about the problem to be estimated. For image processing and height map estimation, this knowledge could particularly cover aspects like:

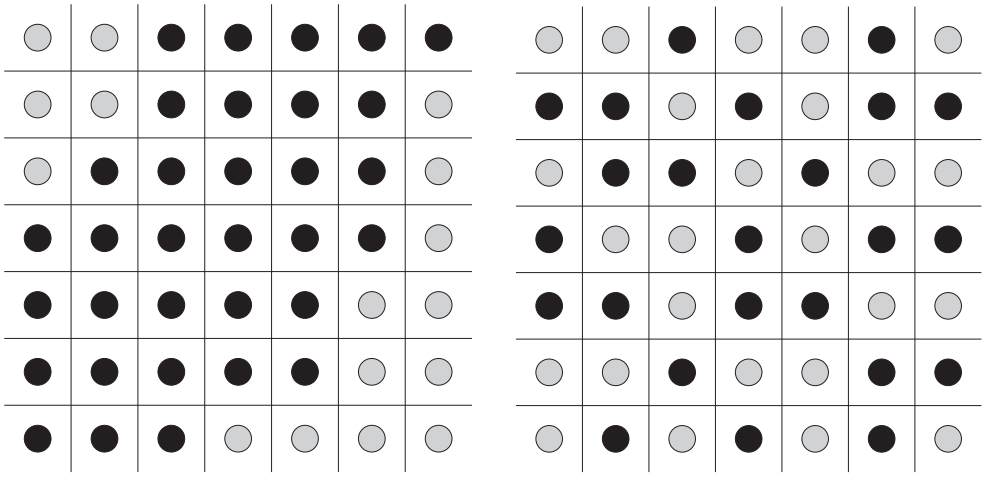
- the general smoothness (corresponding to the roughness for height maps),
- the correlation between neighboring pixels,
- the distribution of gray values on different scales—image waviness and overall skewness, or
- the density and characteristics of edges, corners and other image features.

A prior can be set up in an ad hoc manner, similar to the informal way prior knowledge is described, and without a clear idea of the influence on the estimation process. However, a thoughtful definition can have a large influence on the power of an image restoration approach.

A mathematically elegant approach is of course the definition of a prior via the specification of a cost or energy function $\Gamma_P : \boldsymbol{X} \to \mathbb{R}$ that assigns each image configuration $\boldsymbol{x} \in \boldsymbol{X}$ real-valued costs (cf. Sec. 3.1.4), which of course should be minimum for a configuration close to the design target of the prior. The probability distribution is then derived with the Gibbs formula (cf. Eq. 3.39).

Edges are very common and frequent image features, and they often carry important information. It is therefore particularly interesting to see how the choice of smoothing prior can still preserve edges as far as possible.

With a prior that shows a *robust* behavior towards outliers, edges can be easily handled in an implicit manner, rather than including these image features in the prior definition. As discussed in [Black and Sapiro, 1999], image edges can generally be regarded as an aggregation of outliers, which are not penalized excessively by robust priors. If tuned correctly, such a robust prior suppresses single erroneous pixels, but otherwise leaves edges intact (cf. Sec. 3.3.3 for a short introduction to the general notion of outliers and robustness).



(a) (b)

Figure 3.5: Binary images illustrating the notion of smoothness as a a priori assumption: (a) shows a smooth scene, (b) a less smooth (rough) scene.

3.3.1. Simple priors

Smoothness prior for binary images We take the following a priori assumption for smoothness of a binary image: *Same-colored neighboring pixels are "more favorable" than different-colored pixels.* Along this smoothness paradigm, Fig. 3.5a shows an image of higher a priori probability than that in Fig. 3.5b.

For a simple smoothing prior, we assign unit costs or energy for every pair of pixel that differ, and zero energy for every pair of equal color. We denote $s \sim t$ the pair made of the pixels s and t. The total energy for an image then is:

$$H(\boldsymbol{x}) = \sum_{s \sim t} (1 - \delta(x_s, x_t)) \tag{3.63}$$

With the Gibbs formula, we find for the a priori probability:

$$P(\boldsymbol{x}) = \frac{1}{Z} \exp\{-\beta H(\boldsymbol{x})\} \tag{3.64}$$

$$= \frac{1}{Z} \exp\left\{-\beta \sum_{s \sim t} (1 - \delta(x_s, x_t))\right\} \tag{3.65}$$

$$= \frac{1}{Z} \prod_{s \sim t} \exp\{-\beta(1 - \delta(x_s, x_t))\} \tag{3.66}$$

The constant β is a free parameter in this context, Z is a normalization constant. For most purposes, it can be left indefinite. By principle, it could be determined by summation over the state space, that is, all possible settings of all pixels:

$$\sum_{\boldsymbol{x}} P(\boldsymbol{x}) = 1 \tag{3.67}$$

Linear smoothness prior for gray value images For gray value images, ideas from the preceding paragraph can easily be transferred. For smoothness: *The lesser the gray value difference between two adjacent pixels, the smoother we consider that region.*

We consider a prior as linear, if its energy or cost function has a linear dependency on gray value changes. This leads to a logarithmic formulation for the prior itself, thanks to the Gibbs formula.

To quantify the difference between two pixels, a metric or distance measure is required. We can use the symmetric linear distance or $\mathbb{L}_1$-measure:

$$d(x_s, x_t) = |x_s - x_t| \tag{3.68}$$

Another common choice is the quadratic distance:

$$d(x_s, x_t) = (x_s - x_t)^2 \tag{3.69}$$

The associated energy function becomes:

$$H(\boldsymbol{x}) = \sum_{s \sim t} d(x_s, x_t) \tag{3.70}$$

As before, the prior can be set up like:

$$P(\boldsymbol{x}) = \frac{1}{Z} \prod_{s \sim t} \exp\left\{-\beta d(x_s, x_t)\right\} \tag{3.71}$$

Prior of the Ising / Potts model We have already discussed the Ising and Potts model in the context of Gibbs/Markov fields in Sec. 3.2.1. Comparison of Eq. (3.51) with the energy function for the linear smoother Eq. (3.70) immediately shows a formal correspondence:

$$d_{\text{Ising}}(x_s, x_t) = x_s x_t \tag{3.72}$$

Still, this expression is not a distance measure in the mathematical sense, as for $x_i = \pm 1$ we do not have $d_{\text{Ising}}(x_s, x_t) \geq 0$. However, for the the Potts model distance we get:

$$d_{\text{Potts}}(x_s, x_t) = \delta(x_s, x_t) \tag{3.73}$$

This expression takes only the values 0 (for an equal setting) and otherwise 1, regardless of the difference between the input values. Therefore it can be considered a very robust measure (cf. Sec. 2.3.2).

3.3.2. Line processes

Line processes have been introduced by [Geman and Geman, 1984] as a method to explicitly describe edges, or any other form of spatial discontinuities in images within a Bayesian framework. That way, piecewise smooth surfaces can be recovered.

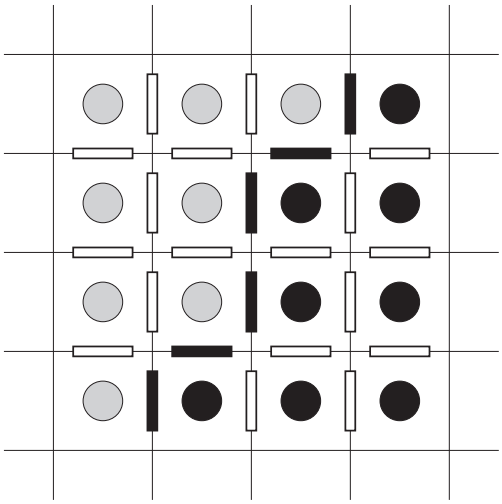


Figure 3.6: Pixel sites (circular) and activation of micro-edges on the dual lattice (4-neighborhood) alongside an edge in the image.

Description of line processes Line processes describe explicitly whether there exists an edge between two pixels or not. Between any two neighboring sites, "micro-edges" are put which indicate the presence of an edge, cf. Fig. 3.6.

For a rectangular pixel array, the line process edges form a dual lattice, made up from micro-edges between every two neighboring pixels (Fig. 3.6). A indicator variable $e_{s,t}$ describes whether there should be an edge between s and t or not:

$$e_{s,t} = \begin{cases} 0 & \text{no edge} \\ 1 & \text{edge exists between } s \text{ and } t \end{cases} \tag{3.74}$$

This state of the micro-edge directly triggers the smoothing between its adjacent pixels and can so inhibit oversmoothing and blurring of edges.

On the other side, to avoid complete disabling of any smoothing by creating an active line between every pair of pixels, an additional penalty is put upon each edge created.

The following energy function balances these antagonists:

$$H(\boldsymbol{x}) = \sum_{s \sim t} d(x_s, x_t)(1 - e_{s,t}) + \kappa e_{s,t} \tag{3.75}$$

The energy is summed over every pair of neighbors, $s \sim t$. $d(x_s, x_t)$ is a distance measure that enforces smoothness, but controlled by the state of $e_{s,t}$. The additional summand $\kappa e_{s,t}$ is the penalty, here chosen simply proportional to the number of micro-edges created.

With the Gibbs formula, the prior of this model is specified. We can now incorporate a model for the noise and image degradation by a likelihood. Together with a measurement (a recorded image), the a posteriori probability is completed. From this we can calculate estimators for the "true" scene. However, to obtain accurate estimates, samples of high a posteriori probability are

necessary. These we only can obtain by intelligently sampling of the configuration space. As the a posteriori probability describes a Markov random field for the original line processes, this can typically be done with the Markov chain Monte Carlo method.

3.3.3. Robust priors

The explicit distance measure in Eq. (3.70) gives an opportunity to incorporate robustness:

- "Small" deviations should have linear influence,
- while the influence of "large" deviations that presumably correspond to edges should be somewhat reduced.

The energy function with a robust error norm d_r then is as before:

$$H(\boldsymbol{x}) = \sum_{s \sim t} d_r(x_s, x_t) \tag{3.76}$$

The influence of a prior derived from a robust energy function depends much on the design of the error norm. The contribution of outlying data points should be controlled, so that they contribute only little. With the means of *influence functions*, which basically correspond to the derivative of the error norm with respect to the distance $x_s - x_t$, this can be analyzed [Hampel et al., 1986]. The influence function shows the bias of an individual measurement towards estimators if it is added to the data set.

Robust error measures can be grouped into ordinary (non-redescending) and redescending measures, which differ by their weighting for large distances (cf. Fig. 3.7):

- For *non-redescending* measures, the influence of large deviations is limited if it exceeds a certain value. That way, each manifest edge in the image contributes that number to the total energy of the image. Such a measure imposes a penalty on edges similar to that in the case of line processes (cf. Sec. 3.3.2).

  The minimax estimator [Huber, 1981] is an example of a non-redescending error measure which changes its behavior at $|x_s - x_t| = \epsilon$ to keep the influence of large outliers constant, cf. Figs. 3.7a and 3.7b:

$$d_r(x_s - x_t) = \begin{cases} \frac{(x_s - x_t)^2}{2\epsilon} + \frac{\epsilon}{2} & \text{for } |x_s - x_t| \leq \epsilon \\ |x_s - x_t| & \text{otherwise} \end{cases} \tag{3.77}$$

- For *redescending* measures, the influence of deviations becomes less the larger the deviation is, once it exceeds a certain number. Deviations that are overly large compared to that parameter tend to have smaller influence the greater they are and are thus allowed without a larger contribution to the overall energy.

A well-known example of a smooth redescending error measure is the Lorentzian estimator, cf. Figs. 3.7c and 3.7d:

$$d_r(x_s - x_t) = \log\left(1 + \frac{1}{2}\left(\frac{x_s - x_t}{\sigma}\right)^2\right) \tag{3.78}$$

The consequences that arise from one or the other choice of error norm is of course only appreciable if the deviations in an image are actually given. Let us at last consider the deviations we expect from a raw height map from white light interferometry of a plain, but rough surface:

Due to the reflection processes as discussed in Chap. 2 (especially Sec. 2.1.4) the height measurements from the surface can be expected to lie roughly within a small, restricted interval with the errors uniformly distributed over the whole height range. Then the parameter of the robust error norm should be chosen in the range of mentioned data interval. In the case of a non-redescending measure, this setting is less critical, as for a parameter too large oversmoothing would become more favorable. On the contrary, for a parameter too small more deviations would be turned into edges, but without an advantage for the overall energy. In case of a redescending measure, choosing the parameter too large would result in a similar behavior.

Also, choosing the parameter too small would likely identify more edges. At the same time, the energy contribution from large deviations would become somewhat smaller without explicitly favoring edges. The implicit penalty for edges is however reduced compared to a linear measure.

One would thus consider both ordinary and redescending error norms appropriate for our problem setting.

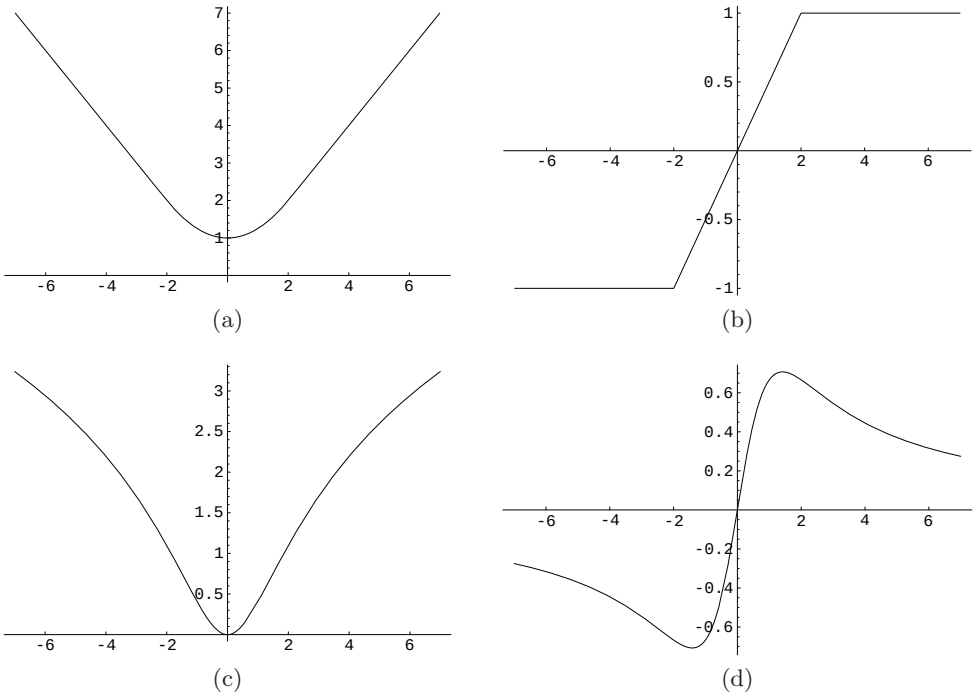


Figure 3.7: Examples of a non-redescending robust error norm, (a) Huber's minimax with (b) influence function and of a redescending robust error norm, (c) Lorentzian with (d) influence function.

4. Bayesian estimation of interferometric height maps

4.1. Overview

4.1.1. Motivation for Bayesian surface reconstruction

As we have discussed in Chap. 2.2, conventional white light interferometry data analysis spans over three consecutive steps. In the postprocessing step, erroneous, unreliable or missing height values have to be replaced by values that are the best estimates from all that can be deduced from the surroundings (cf. Fig. 2.9). If this procedure is done by some kind of spatial smoothing over the raw height map, implicitly some a priori information on the outcome is used. For example, filtering with a Gaussian mask corresponds to the assumption of normal distributed displacements for the erroneous pixels. Robust alternatives to Gaussian filtering have been considered, but their respective a priori assumptions are not as readily tractable as in the Gaussian case.

Upon this background it seems a sensible step to invert the perspective, in a way that we now choose the filtering procedure and its parameter according to available a priori information. To that end, we follow the approach of Bayesian estimation for image reconstruction (cf. Chap. 3), which we modify towards the processing and denoising of height maps from white light interferometry. With Bayesian inference, our prior knowledge can be accounted for within a tight statistical framework. From the statistics of height maps for the surface under inspection (cf. Sec. 2.1), an optimized prior for their processing can be set up. However, as very frequently with Bayesian inference, the computational load of this inverse approach can be immense. Therefore we move our focus to a class of priors, which is both computationally acceptable and well adapted for rough surfaces. That way, an approach of principal novelty for the processing of data from white light interferometry is developed.

As the data acquisition procedure in interferometry is a scanning process, the raw data are a stack of 2-D images. This can also be seen as a 3-D data set, with the third dimension representing the height axis and determined by the intensity at the respective position of the scanner. Bayesian estimation in conventional, 2-D image restoration includes the specification of a likelihood which covers the statistical uncertainty for a pixel's gray value. The likelihood can describe the random effects in the image formation process with a model for the noise and other signal deteriorations that move a pixel away from the "true" gray value. It forms a probability distribution for each pixel which can be considered as an additional data dimension.

With intensity time series from scanning interferometry, we already have this

additional data dimension available. When setting up the likelihood function, we can benefit all information in the full recorded time series for each pixel together with our external prior knowledge.

Along this path, Bayesian inference is carried over to the height estimation for white light interferometry in Sec. 4.2. While for the prior the examples presented in Sec. 3.3 can easily be carried over, the likelihood requires either the development of a statistical evaluation of the scanned time series or a practical re-interpretation of the same.

4.1.2. Scientific context

This idea for Bayesian height estimation originates from an approach developed by [Hartvig and Jensen, 2000] for the denoising of diachronous functional-MRI activation maps of functional MRI (fMRI). The authors introduce contextual information into a Bayesian approach developed by [Everitt and Bullmore, 1999]. While [Beckmann and Smith, 2003] remark to the gaining attention to spatial prior information, the procedure in [Hartvig and Jensen, 2000] so far remains solitary.

Bayesian inference is well established for the restauration of 2-D images, as described in Chap. 3, and the requirements for certain priors are the same as one can use to describe height maps from interferometry (cf. the references in Sec. 3.1). Markov random field approaches like stochastic sampling dominate the portfolio of estimation procedures.

Besides with MRI imaging, both the image formation process and the processing of the resultant 2-D image are subject of joint research with SAR[1] imaging, a high-resolution technique used for ground surveillance and classification. In particular, imaging with multichannel-SAR is interesting: Here the ground scene that should be classified is recorded by a small number of cameras (or channels with different wavelength) in parallel. While for each channel a rough mapping to ground features (soil, water, forest, etc.) exists, a more precise estimate is expected from the joint analysis of all channels. The estimation task is therefore to affix label to the pixels (cf. Sec. 3.1) for which a third data dimension is available.

In [Hjort and Mohn, 1984] a Bayesian inference procedure is described which, as a novelty, explicitly addresses the contextual information, which has not had particular attention until then. The method is very similar to the one we propose. The authors also succeed in directly calculating estimates from the a posteriori probability, albeit at the cost of length computation time and the restriction to very small images which a handful of channels. A comparison of the performance of that approach can be found in [Mohn et al., 1987]. As one can oversee, this particular direction of research was not followed any further, probably set aside with the arrival of even more powerful computers and the gaining acceptance of Markov random field methods.

[1] Synthetic Aperture Radar

4.2. Bayesian estimation

Mathematical notation We follow the convention to distinguish vectors or matrices from scalar variables by using bold script, as for an image we use $\boldsymbol{x}$, and x^j for one gray value pixel with index j. For notational simplicity, we mostly refrain from explicitly stating both spatial coordinates of an image, like for a single channel image $x^{\{\dot{x},\dot{y}\}}$. Instead, we define a straight enumeration scheme to cover the image plane and have a single-index notation, like x^j, x^{j+1}, etc. For a rows-first enumeration scheme, the correspondence from Cartesian coordinates can be found according to $j = \dot{y} \cdot \dot{x}_{\max} + \dot{x}$.

The white light interferometer provides us, for each pixel j, with a sequence of intensity values, which we denote as a vector with superscript index: $\boldsymbol{x}^j$. This vector is indexed $h = 1, \ldots, h_{\max}$. What we ultimately seek is an estimate for the height of each pixel, which we write as $\hat{h}^j$.

Let $P(h^j)$ be the a priori probability density function for some pixel's height, and $P(\boldsymbol{x}^j)$ the according function for a data sequence. The likelihood of a data sequence $\boldsymbol{x}^j$ under the condition of fixed height h^j is $f(\boldsymbol{x}^j|h^j)$, the a posteriori probability density function for a height under the condition of a data sequence will be written as $\mathbb{P}(h^j|\boldsymbol{x}^j)$. This notation is more convenient for the discussion of height maps, but does not comply with the conventions chosen in 2-D image processing (cf. Eq. (3.4), e. g.).

Bayesian estimation According to the Bayesian paradigm (cf. Sec. 3.1.2 and Eq. (3.4)), for each pixel j we can state the a posteriori probability density:

$$\mathbb{P}(h^j|\boldsymbol{x}^j) = \frac{f(\boldsymbol{x}^j|h^j)\, P(h^j)}{P(\boldsymbol{x}^j)} \tag{4.1}$$

The denominator $P(\boldsymbol{x}^j)$ is known as the Bayesian *evidence*. As discussed in Sec. 3.1.2, our aim is to find estimates of the a posteriori probability under given fixed data $\boldsymbol{x}$. We are only interested in modes of the posterior, and the evidence will remain constant, therefore $P(\boldsymbol{x}^j)$ can be left out of the remainder of the discussion:

$$\mathbb{P}(h^j|\boldsymbol{x}^j) \propto f(\boldsymbol{x}^j|h^j)\, P(h^j) \tag{4.2}$$

For the actual calculations on a computer, it can become helpful to use the logarithmized equivalent of this formula to circumvent numeric underflows with the multiplication—at the price of a numerically less stable addition:

$$\log \mathbb{P}(h^j|\boldsymbol{x}^j) \propto \log f(\boldsymbol{x}^j|h^j) + \log P(h^j) \tag{4.3}$$

Consideration of the neighborhood Bayesian estimation draws much of its power from the possibility to explicitly state prior knowledge about the surface. While of course, *global* knowledge about surface to be measured is almost never available, we often have *local* prior knowledge. Examples are the characteristics of the microscopic height profile of a surface or knowledge of its directionality

after a honing process. We describe the local knowledge with the joint statistics of a pixel and its *neighbors*. The correlation strength to the neighborhood is usually assumed to drop with increasing distance of the sites, so they are only taken into account up to a certain distance. Common settings to mimic this on the rectangular grid are the 4-neighborhood (Fig. 3.1a) covering minimum mutual, but isotropic dependency, and the 8-neighborhood (Fig. 3.1b), which additionally takes dependencies along the diagonals into account. This neighborhood setting is also common in conventional image processing, i. e. whenever a 3×3 filtering mask is used. The next larger neighborhood system, suitable for explicitly covering larger distance dependencies, is the less common 24-neighborhood or 5×5 mask (Fig. 3.1c), which also takes the second-to-next neighbors into account.

We denote the set made up of a pixel j and its neighbors (according to some neighborhood relationship which arbitrary at this point) $\mathcal{C}$. The member sites of $\mathcal{C}$ we will conveniently index $0, \ldots, k$, with the index 0 reserved for the center pixel j and $k = |\mathcal{C}| - 1$. In case of the 8-neighborhood, we would have $k = 8$, in case of the 5×5 mask's neighborhood accordingly $k = 24$.

We formulate the Bayesian statement Eq. (4.2) again for the case of a group $\mathcal{C}$ of pixels:

$$\mathbb{P}(h^{\mathcal{C}}|\boldsymbol{x}^{\mathcal{C}}) \propto f(\boldsymbol{x}^{\mathcal{C}}|h^{\mathcal{C}})\, P(h^{\mathcal{C}}) \tag{4.4}$$

After all, we are only interested in the probability distribution of the central pixel given the data sequences of the group of pixels, $\mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}})$. This expression may be found by marginalization, i. e. integrating "out" the other degrees of freedom, which here is the sum over all possible height configurations of the neighbors:

$$\mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) = \sum_{h^1} \cdots \sum_{h^k} \mathbb{P}(h^{\mathcal{C}}|\boldsymbol{x}^{\mathcal{C}}) \tag{4.5}$$

The likelihood term $f(\boldsymbol{x}^{\mathcal{C}}|h^{\mathcal{C}})$ in Eq. (4.4) expresses the probability density for all the data sequences $\boldsymbol{x}$ within the neighborhood $\mathcal{C}$ with respect to the heights $h^{\mathcal{C}}$ in that neighborhood. We can simplify this by assuming that the probability densities for each pixel are independent from each other. While this can only be the case for an ideal imaging system—we have already seen in Sec. 2.1.4 that diffraction-limited speckle do not match on rectangular image sensors—it is still a reasonable approximation: One can consider that one can regard the speckle-induced cross-talk as noise in the other pixel—now the common assumption of strictly white noise cannot hold anymore. Technically however, for current CCD sensors the active areas within a pixel are significantly smaller than the pixel itself, which also supports the independence assumption. Accepting all this, we now postulate that the data sequence acquired in one pixel does only depend on the object height in the area, which is images onto that pixel. Then the likelihoods for each pixel are independent, and the joint likelihood can be factorized:

$$f(\boldsymbol{x}^{\mathcal{C}}|h^{\mathcal{C}}) = \prod_{i=0}^{k} f(\boldsymbol{x}^i|h^i) \tag{4.6}$$

The estimation of the a posteriori probability under consideration of neighboring pixels can thus be written like:

$$\mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) \propto f(\boldsymbol{x}^0|h^0) \sum_{h^1} \cdots \sum_{h^k} \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^i)\, P(h^{\mathcal{C}}) \tag{4.7}$$

Looking at Eq. (4.7), one notices the multiply nested sums which run over small products and then each over $h_{\max}$ operations. This is altogether more than $k(h_{\max})^k$ operations for each pixel to be estimated: the a posteriori probability for all combinations of height configurations of a pixel and its neighborhood is calculated. The combinatorial task in this form bear a large computational load, in particular, if the number of height steps $h_{\max}$ is high.

4.2.1. Cost functions

The height in a pixel is calculated from the a posteriori probability distribution by an estimation function. This estimator can be chosen straight away, or by defining a cost function.

We here go along with choosing a variation of the marginal posterior mode estimator, which is a kind of local maximum-a-posteriori estimator (cf. Sec. 3.1.4. Within a local scope, the estimate is the height which locally has the highest a posteriori probability:

$$\hat{h}^0 = \arg\max_{h^0} \mathbb{P}(h|\boldsymbol{x}^{\mathcal{C}}) = \arg\max_{h^0} \sum_{h^1} \cdots \sum_{h^k} \mathbb{P}(h^{\mathcal{C}}|\boldsymbol{x}^{\mathcal{C}}) \tag{4.8}$$

As we have discussed for the general case in Sec. 3.1.4, we can deduce this estimator from a corresponding cost function:

We do so by choosing a "hard" cost function Γ for the height of the central pixel h_0, which we however define for this pixel *and* its neighbors, i. e. $h^{\mathcal{C}}$. This particularity is required for exact analogy with the estimator.

A set of height values which contains the correct ("true") value for h^0, namely ${h^0}_{\text{true}}$, is designated ${h^{\mathcal{C}}}_{(\text{true})}$. Non-zero, constant costs for a set $h^{\mathcal{C}}$ exist only when the height of the center pixels fails the correct value. Formally:

$$\Gamma(h^{\mathcal{C}}, {h^{\mathcal{C}}}_{(\text{true})}) = \begin{cases} 1, & \text{if } h^0 \neq {h^0}_{(\text{true})} \\ 0, & \text{otherwise} \end{cases} \tag{4.9}$$

Also formally, it is clear that our a posteriori probability also incorporates information about neighboring pixels, which we do not consider in our cost calculation, i. e.:

$$\Gamma(h^{C}, {h^{C}}_{(\text{true})}) = \Gamma(h^0, {h^0}_{\text{true}}) \tag{4.10}$$

The average (Bayesian) costs $\bar{\Gamma}$ are calculated by averaging over all settings of the unknown truth, ${h^{\mathcal{C}}}_{(\text{true})}$, weighted with the a posteriori probability:

$$\bar{\Gamma}(\tilde{h}^{\mathcal{C}}) = \sum_{h^0} \left(\sum_{h^1} \cdots \sum_{h^k} \Gamma(h^{\mathcal{C}}, {h^{\mathcal{C}}}_{(\text{true})}) \right) \mathbb{P}(h^{\mathcal{C}}|\boldsymbol{x}^{\mathcal{C}}) \tag{4.11}$$

$$= 1 - \sum_{h^1} \cdots \sum_{h^k} \mathbb{P}(\tilde{h}^0, h^1, \ldots, h^k|\boldsymbol{x}^{\mathcal{C}}) \tag{4.12}$$

The Bayesian estimate is an estimate of minimum costs. Looking at the last equation (4.12), we can fulfill this requirement by choosing:

$$\hat{h}^0 = \arg\max_{h^0} \sum_{h^1} \cdots \sum_{h^k} \mathbb{P}(\tilde{h}^0, h^1, \ldots, h^k | \boldsymbol{x}^{\mathcal{C}}) \tag{4.13}$$

With the choice of Eq. (4.9) we have defined costs for a configuration $\mathcal{C}$ of pixels. A fixed penalty is imposed on the center pixel if it misses the true value, but the neighboring pixels are ignored. These costs are robust as they do not increase with the distance from the true setting for the center pixel 0.

This cost function stands between a globally "hard" definition, like the maximum a posteriori estimator (cf. Eq. (3.17)), and purely locally defined "hard" costs, as we have with the marginal posterior mode estimator (cf. Eq. (3.23)).

4.2.2. Derivation of likelihood functions

The likelihood function $f(\boldsymbol{x}^j | h^j)$ describes the probability of a data sequence given a height for a single pixel. It therefore answers the question: What could the recorded data look like if the object height were h^j? In the answer to this question, and so in the likelihood function a part of the power of Bayesian inference lies: The question we ask here is the exact inverse of the original setting, i. e., what the height could be when the recorded data are like this.

We consider two major paths for the derivation of the likelihood function from the experimental settings:

1. Strict modeling of the physics of the signal formation process in white light interferometry.

2. Phenomenological derivation from the recorded data.

Both approaches have their specific advantages and drawbacks. For the first option, a good account of the physical processes, as well as of the technical imperfections within the actual interferometer setup should be taken available. It seems difficult to determine the magnitude of parameters involved. If they however could all be fixed, a full simulation of the signal generation process would be available. From that, with knowledge about the noise sources and their characteristics, a likelihood function could be generated.

Sketch of a physical model for white light interferometry The knowledge about the signal formation process for the reflection from a rough surface is only fragmentary so far, restricted to almost ideal roughness (the research baseline can be found in [Dainty, 1984]). For the ideal case, the signal of a white light interferometer is of the following form (cf. Sec. 2.2 and Eq. (2.45)):

$$I(z) = I_0 + I_1 G(z - z_0) \cos(kz + \varphi_0) \tag{4.14}$$

After reflection at an ideally rough surface the backscattered signal has a random phase shift with uniform probability distribution (cf. Sec. 2.1.3 and [Goodman, 1984]:

$$p(\varphi_0)\, d\varphi_0 = \mathcal{U}(0, 2\pi)\, d\varphi_0 \tag{4.15}$$

Also for ideally rough surfaces, the returned intensity has an exponential distribution:

$$p(I)\,dI = \frac{1}{I_0}\,e^{-I/I_0}\,dI \tag{4.16}$$

If we assume that the measurement noise on the detector $\mathcal{N}(0,\sigma^2)$ be i. i. d.[1] with variance σ^2, the likelihood can be set up as a product over the probability density function for the noise distribution Φ, evaluated for the measured intensity of each height value and given $\mathcal{K}$:

$$f(\boldsymbol{x}^j|h^j)_{\mathcal{K}} = \prod_{z=1}^{h_{\max}} \Phi(I(z)|\mathcal{K}(z|h)) \Bigg|^{\text{pixel } j} \tag{4.17}$$

The function $\mathcal{K}$ consolidates the probability distribution of an expected interference pattern with all known or unknown parameters. It is therefore an expression for the location likelihood of an interferogram: $\mathcal{K} = \mathcal{K}(z|h)_{\ldots}$. If for example the base intensity I_0 is known, and the amplitude of the inner oscillation I_1 can be calculated, they can be taken out of the estimation process. At least the height location z_0 of the interference pattern is unknown, probably also the phase shift φ_0 of the inner oscillation. When we assume a uniform distribution for the unknown properties, $\mathcal{K}$ becomes (cf. Eq. (4.14)):

$$\mathcal{K}(z|h)_{I_1,\varphi_0} = \sum_{z'=1}^{h_{\max}} \big(I_1 + G(z-h)\cos(kz' + \varphi_0)\big) \tag{4.18}$$

and

$$\mathcal{K}(z|h)_{I_1} = \int_0^{2\pi} \mathcal{K}(z|h)_{I_1,\varphi_0}\,d\varphi_0 \tag{4.19}$$

Due to the integrations along the full height range and the possible phase shifts, $\mathcal{K}$ becomes less significant as a "pattern matching" tool for the likelihood in Eq. (4.17).

In practice, the proper modeling of even more subtle effects like the intensity variation due to mixing of polarizations makes this approach less efficient and only increases the computational burden.

Phenomenological derivation The phenomenological approach is the empirical route to a likelihood function. The characteristics of a properly derived likelihood function are mimicked with the transformed raw data series.

An ideal likelihood function would feature a strong response around where the true height value is hidden in the time series. Where noise dominates the data, it should return only a weak response. Actually, these requirements are fulfilled already to a high degree with the methods developed for interferogram preprocessing (cf. Sec. 2.2.1 and references there).

[1] Independent, identically distributed—a common assumption in statistics, sometimes also abbreviated 'iid'

For the results following and the comparison in Sec. 4.3, this second path is chosen. We use a quasi-likelihood that is derived from the sliding average algorithm (cf. Sec. 2.2.1).

4.2.3. Choice of prior and direct a posteriori estimation

Direct estimation of the a posteriori probability Before we go into calculations with the a posteriori probability distribution, we prove the following equation [Hartvig and Jensen, 2000], useful to rearrange the nested sums that appear in the marginalization of the posterior probability:

$$\sum_{h^1} \cdots \sum_{h^k} \prod_{j=1}^{k} f(j, h^j) = \prod_{j=1}^{k} \sum_{h} f(j, h) \tag{4.20}$$

Proof:
By induction over k; we see for $k = 1$ the immediate equivalence. For $k+1$, we have

$$\sum_{h^1} \cdots \sum_{k^{k+1}} \prod_{j=1}^{k+1} f(j, h^j) \tag{4.21}$$

$$= \sum_{h^1} \cdots \sum_{h^k} \sum_{h^{k+1}} \prod_{j=1}^{k} f(j, h^j) f(k+1, h^{k+1}) \tag{4.22}$$

$$= \sum_{h^{k+1}} \left[\sum_{h^1} \cdots \sum_{h^k} \prod_{j=1}^{k} f(j, h^j) \right] f(k+1, h^{k+1}) \tag{4.23}$$

$$= \sum_{h^{k+1}} \left[\prod_{j=1}^{k} \sum_{h} f(j, h) \right] f(k+1, h^{k+1}) \tag{4.24}$$

$$= \prod_{j=1}^{k} \sum_{h} f(j, h) \sum_{h^{k+1}} f(k+1, h^{k+1}) \tag{4.25}$$

$$= \prod_{j=1}^{k+1} \sum_{h} f(j, h) \tag{4.26}$$

□

This rearrangement can significantly reduce the computational effort for the calculation of a posteriori probabilities. If we have $h_{\max}$ values for each of the nested sums on the left side of Eq. (4.20), then the computation requires $k\,(h_{\max})^k$ operations. The right side only accounts for $k\,h_{\max}$ operations.

Typical numbers in practice could be $h_{\max} = 1000$ and $k = 8$. For the left side, this adds up to $8 \cdot 10^{24}$ operations, but for the right side only 8000 operations.

Choice of prior and direct a posteriori estimation The third component still remaining for the Bayesian reconstruction process is the prior probability distribution, $P(h^{\mathcal{C}})$, describing the probability for the occurrence of a certain height configuration within the neighborhood (cf. Sec. 4.1). In the prior, we can cast our general knowledge about the surface under test, in particular on the possible height configurations, as well as technical restrictions, especially the maximum achievable resolution for white light interferometry with rough surfaces.

By thoughtful choice of the prior, we have seen that the exact estimation of the posterior probability can be made feasible [Hartvig and Jensen, 2000]. So far, estimates from the a posteriori probability have in practice always required the use of simulation methods.

δ-prior We first look at a quite simple prior, which, as we will learn later, is part of a certain class of priors for direct a posteriori estimation. The prior should favor smoothness, and it should be robust. With these two ingredients, we can reasonably well cover the global properties of rough, planar surfaces with edges.

A very simple approach is to define a prior having only two outcomes, each for a favorable and a unfavorable configuration:

$$P(h^{\mathcal{C}}) = \begin{cases} q_1 & \text{if } h^1, \ldots, h^k = h^0, \\ q_0 & \text{otherwise} \end{cases} \tag{4.27}$$

This can also be written like:

$$P(h^{\mathcal{C}}) = q_0 + (q_1 - q_0) \prod_k \delta(h^k - h^0) \tag{4.28}$$

The prior compares the height of a neighborhood pixel k, named h^k with that of the central pixel, h_0. If they are unequal, the δ-function returns a zero. Due to the product over all neighbors, the second term of the prior only becomes non-zero when all neighbors have the same height as the center pixel, i. e., the surface is locally completely smooth. This is a favorable configuration, and the prior then becomes $P(h^{\mathcal{C}}) = q_1$. For other configurations, the prior is $P(h^{\mathcal{C}}) = q_0$.

The prior is a valid probability distribution if we choose $q_0 > 0$ and $q_1 \geq 0$. With $q_1 > q_0$, the desired smoothing effect is present in Bayesian inference. This prior is also very robust, as the probability for the non-smooth case does not depend on the actual distance of the h^k from h^0. The terms q_0, q_1 are parameters which can be used for weighting the favorable over the unfavorable configurations and for normalizing the whole expression.

With this—somewhat unusual, but still valid—prior, we now calculate the a posteriori probability according to Eq. (4.7):

$$\mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) \propto f(\boldsymbol{x}^0|h^0) \sum_{h^1} \cdots \sum_{h^k} \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^i) \left(q_0 + (q_1 - q_0) \prod_k \delta(h_k - h_0) \right)$$

(4.29)

Out of the nested sum, the product of δ-functions eliminates all contributions except for the ones with $h^1 = \cdots = h^k = h^0$:

$$\mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) \propto f(\boldsymbol{x}^0|h^0) \left(q_0 \sum_{h^1} \cdots \sum_{h^k} \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^i) + (q_1 - q_0) \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^0) \right) \tag{4.30}$$

Using Eq. (4.20), we rearrange the sums and products to get:

$$\mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) \propto f(\boldsymbol{x}^0|h^0) \left(q_0 \prod_{i=1}^{k} \sum_{h} f(\boldsymbol{x}^i|h) + (q_1 - q_0) \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^0) \right) \tag{4.31}$$

In this form, the a posteriori probability distribution is rather easy to calculate. To go a step further, we look into the particular sum

$$\sum_{h} f(\boldsymbol{x}^i|h) =: \mathcal{S}_{f_i} \tag{4.32}$$

which describes the overall likelihood for a data sequence $\boldsymbol{x}$ in pixel i, taken over all possible height values. If we use the phenomenological derivation of likelihood values, this sum is directly related to the average visibility or signal-to-noise of the interference signal in the full data series of a pixel.

For some a posteriori estimators, only the maximum value of the posterior probability is needed. If the visibility sum $\mathcal{S}_{f_i}$ is available from preprocessing, the required calculations can be done very fast on ordinary computers:

$$\arg\max_{h^0} \mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) = \arg\max_{h^0} f(\boldsymbol{x}^0|h^0) \left(q_0 \prod_{i=1}^{k} \mathcal{S}_{f_i} + (q_1 - q_0) \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^0) \right) \tag{4.33}$$

The outcome of the Bayesian estimation has a large dependency on the constants q_0 and q_1. Generally, the values of the (normalized) likelihood function have a similar order of magnitude, therefore a balanced weighting between the center and the neighboring pixels would require $(q_1 - q_0) \approx {q_0}^k$.

Let us consider two extrema for this weighting:

- $q_0/(q_1 - q_0)^{-k} \gg 1$: In this case, the parameter $(q_1 - q_0)$ could be neglected. The a posteriori probability becomes proportional to the likelihood of the central pixel, which leads to a local maximum likelihood estimator. In this case no smoothing is present, and the resultant height map is the same as we could obtain when only a preprocessing that corresponds to the likelihood is applied (cf. Sec. 4.2.2).

- $q_0/(q_1 - q_0)^{-k} \ll 1$: Now, the parameter q_0 could be neglected. The a posteriori probability is determined by the likelihoods of all pixels within

the neighborhood and the central pixel plays only a minor role. Therefore a multiplicative filtering within the neighborhood takes place, and the height map is estimated as the maximum likelihood for the filtered data. This situation also has some resemblance to a special case of channel smoothing (cf. Sec. 5.2).

Rectangle prior The central assumption underlying the δ-prior discussed in the last paragraphs is complete local smoothness. This is represented by the height values of neighboring surface points being equal within the resolution allowed by the discrete height stepping. We can relax this constraint in a way that we allow for somewhat larger height variation within a neighborhood, so that we arrive naturally at the *rectangle* prior:

$$P(h^{\mathcal{C}}) = \begin{cases} q_1 & \text{if } h^1, \ldots, h^k \in [-\Lambda + h_0,\, \Lambda + h^0] \\ q_0 & \text{otherwise} \end{cases} \tag{4.34}$$

This can be written in form of a single-line function as:

$$P(h^{\mathcal{C}}) = q_0 + (q_1 - q_0) \prod_{i=1}^{k} W_\Lambda(h^i - h^0) \tag{4.35}$$

The function W_Λ that replaces the δ-peak from Eq. (4.28) is a rectangle window of width Λ defined like:

$$W_\Lambda(h) = \begin{cases} 1 & \text{for } -\Lambda \leq h \leq \Lambda \\ 0 & \text{otherwise} \end{cases} \tag{4.36}$$

The second term of this prior only becomes non-zero in a configuration where the height values of all neighbors of the central pixel lie within the Λ-range around the central height, h_0. The prior then becomes $P(h^{\mathcal{C}}) = q_1$. If only a single pixel is outside this range, the product becomes zero. Such a configuration we consider as unfavorable, the prior here becomes $P(h^{\mathcal{C}}) = q_0$. As before, the restrictions $q_0 > 0$ and $q_1 \geq 0$ ensure that Eq. (4.35) is a valid probability distribution. The rectangle prior shares its robustness with the δ-prior, as also here the co-domain is strictly bounded to only two values, whatever large the differences between the h^i and h^0 gets.

With this prior, we concede variation of the height within a small range around each pixel still as a favorable configuration. Due to the local definition, with this prior the global properties cannot be controlled. Gradual changes, overall tilt and large-scale deformations such as waviness are therefore also considered favorable according to this prior.

The breadth of the rectangle W_Λ should be chosen in coarse correspondence to the range of height variation of the rough surface: As discussed in Sects. 2.1.4 and 2.2, the phase of the reflected light is arbitrary and the detected height only a random value between the limits of each speckle. Although this is strictly true only for perfectly rough surfaces, it seems plausible to choose Λ so that the prior allows (favors) height variation within the presumed range of roughness.

We now utilize this prior to calculate the a posteriori probability with the same approach as we have demonstrated for the δ-prior. The posterior probability from Eq. (4.7) then becomes:

$$\mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) \propto f(\boldsymbol{x}^0|h^0) \sum_{h^1} \cdots \sum_{h^k} \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^i) \left(q_0 + (q_1 - q_0) \prod_{i=1}^{k} W_\Lambda(h^i - h^0) \right) \tag{4.37}$$

For the nested sum, the prior W_λ makes all contributions zero that lie outside the λ-range around the height of the central pixel. Therefore we can write:

$$\begin{aligned} \mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) \propto f(\boldsymbol{x}^0|h^0) \Bigg(& q_0 \sum_{h^1} \cdots \sum_{h^k} \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^i) \\ & + (q_1 - q_0) \sum_{h^1=h^0-\Lambda}^{h^0+\Lambda} \cdots \sum_{h^k=h^0-\Lambda}^{h^0+\Lambda} \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^i) \Bigg) \end{aligned} \tag{4.38}$$

We again make use of Eq. (4.20) to exchange the sums and products and get:

$$\begin{aligned} \mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) \propto f(\boldsymbol{x}^0|h^0) \Bigg(& q_0 \prod_{i=1}^{k} \sum_{h} f(\boldsymbol{x}^i|h) \\ & + (q_1 - q_0) \prod_{i=1}^{k} \sum_{h=h^0-\Lambda}^{h+\Lambda} f(\boldsymbol{x}^i|h) \Bigg) \end{aligned} \tag{4.39}$$

For practical purposes, a fast calculation of the maximum of the a posteriori probability distribution is helpful. Here we can make use of pre-calculated values for the S_{f_i} sums Eq. (4.32) and for the sums

$$\sum_{h=h^0-\Lambda}^{h+\Lambda} f(\boldsymbol{x}^i|h) := \tilde{f}_\Lambda(\boldsymbol{x}^i|h) \tag{4.40}$$

which basically represents a moving average. This gives us finally:

$$\arg\max_{h^0} \mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) = \arg\max_{h^0} f(\boldsymbol{x}^0|h^0) \left(q_0 \prod_{i=1}^{k} \mathcal{S}_{f_i} + (q_1 - q_0) \prod_{i=1}^{k} \tilde{f}_\Lambda(\boldsymbol{x}^i|h) \right) \tag{4.41}$$

The exchange operation from Eq. (4.20) is the central step to make direct calculation of estimates from the a posteriori probability possible:

For a typical application, we have to consider in the order of 1000 height steps. Let us further assume an 8-neighborhood ($k = 8$) and an interval of favorable height variation of 11 steps ($\Lambda = 5$). To calculate the a posteriori probability for one pixel from Eq. (4.38), the calculation of the second summand is required.

This makes up roughly 1000 (height range) calculations of $10^k = 10^8$ additions of 1000 products, in total about 10^{14} operations *per pixel*.

On the other side, for Eq. (4.39), the second summand accounts for 1000 (height range) calculations of 8 products of 10 sums, or totally $8 \cdot 10^5$ operations, which is fairly bearable for average computers nowadays.

Fig. 4.1 is a quick demonstration of the performance of Bayesian surface estimation with a rectangle prior. The measurement object is the turned steel piece photographed in Fig. 4.2. The primary height map obtained from preprocessing (Fig. 4.1a) bears a large number of outliers, which is an indication of the very poor quality of the raw data. The reconstruction with the Bayesian estimation procedure (Fig. 4.1b) is slightly patchy, but with sharp edges and contains only few outliers (for a detailed discussion cf. Sec. 4.3.5).

Normalization of a prior In the last paragraphs, we skipped the question of the properties and mutual relationship of the parameters q_0, q_1 and Λ.

The normalization condition for the prior $P(h^{\mathcal{C}})$ leads to a constraint on the ratio q_0/q_1. The condition is such that

$$1 = \sum_{h^{\mathcal{C}}} P(h^{\mathcal{C}}) = \sum_{h^1} \cdots \sum_{h^k} P(h^{\mathcal{C}}) \tag{4.42}$$

i. e., the sum over the probabilities of all possible height configurations of the set $\mathcal{C}$ should be equal one.

We now specify this for the priors considered so far:

- For the δ-prior, one favorable configuration in terms of smoothness exists, this is when all neighboring heights are equal to the central height value. The latter can be anywhere in the range $1, \ldots, h_{\max}$. That is, $h_{\max}$ configurations exist with a probability of q_1:

  $$P_{\text{favourable}} = q_1 \, h_{\max} \tag{4.43}$$

  All the other configurations have the specific probability q_0. We can calculate their quantity by subtracting the number of favorable configurations from the total number, which is $(h_{\max})^{k+1}$:

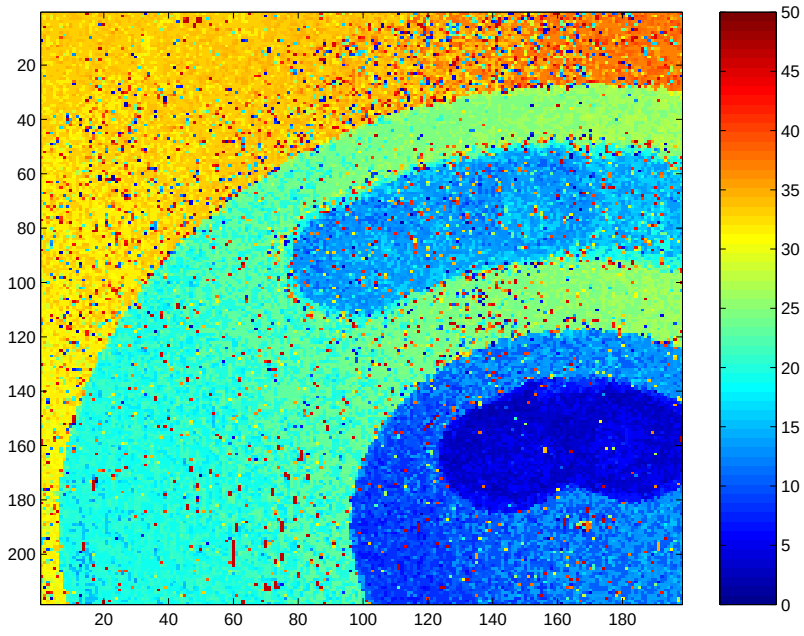
  $$P_{\text{unfavourable}} = q_0 \left((h_{\max})^{k+1} - h_{\max} \right) \tag{4.44}$$

  Altogether, this gives us the normalization condition:

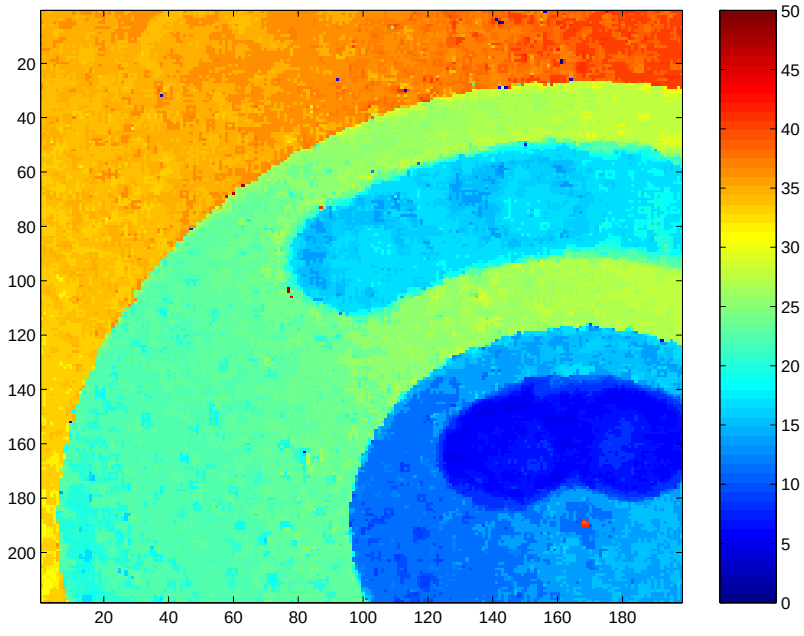
  $$1 = q_1 \, h_{\max} + q_0 \left((h_{\max})^{k+1} - h_{\max} \right) \tag{4.45}$$

  For the δ-prior, we have the freedom to choose the ratio q_0/q_1. As an example, let us look into the case of only $h_{\max} = 100$ height steps and an 8-neighborhood. We choose $q_0/q_1 = 100$, that is, a 100 times larger a priori probability for a smooth neighborhood than a rough setting. Then we find

  $$q_0 \approx 10^{-18} \qquad \text{and} \qquad q_1 \approx 10^{-20} \tag{4.46}$$



(a)



(b)

Figure 4.1: Reconstructions for data recorded at 84 μm/s. (a) shows the height map obtained from preprocessing only, (b) the height map from Bayesian surface estimation. Scale is 1.68 μm per frame.

- For the rectangle prior, we can use a similar approach. The favorable configurations cover the range $h^0 - \Lambda \leq h^i \leq h^0 + \Lambda$, i. e. $2\Lambda + 1$ height steps. This range we take into account for each neighboring pixel, in addition to the freedom to settle the central pixel over the full range. That way we neglect minor boundary effects for the central pixel being near the limits of the height range, in particular $h^0 < \Lambda$ and $h^0 > h_{\max} - \Lambda$. Thus we have:

$$P_{\text{favourable}} = q_1 \, (2\Lambda + 1)^k \cdot h_{\max} \tag{4.47}$$

All other configurations are unfavorable concerning smoothness, and we find their number by subtracting the favorable ones from the complete number:

$$P_{\text{unfavourable}} = q_0 \left((h_{\max})^{k+1} - (2\Lambda + 1)^k \cdot h_{\max} \right) \tag{4.48}$$

Finally, this leads to the normalization condition:

$$1 = q_1 \, (2\Lambda + 1)^k \cdot h_{\max} + q_0 \left((h_{\max})^{k+1} - (2\Lambda + 1)^k \cdot h_{\max} \right) \tag{4.49}$$

We here have the freedom to choose two out of the three parameters. This can be ratio q_0/q_1 to tune the smoothing enforcement of the prior and Λ for the favorable height range. The best ratio for q_0/q_1 can in practice be found by screening through of a small range of values typical of the particular kind of surface. We suggest choosing Λ to be in the same order as the surface roughness. The prior will then ignore the height variation due to roughness, but suppress gross errors such as outliers.

With the same numbers used in the δ-prior example and additionally setting the parameter $\Lambda = 10$, we here find:

$$q_0 \approx 10^{-18} \qquad \text{and} \qquad q_1 \approx 10^{-20} \tag{4.50}$$

One can see from the numbers for q_0 and q_1 that the implementation of this estimation procedure has to take particular care for the numerical stability of the calculations, cf. Sec. 4.3.1.

4.3. Application and assessment

In this section, we describe a pilot application and introduce methods how to quantify the power of denoising approaches for white light interferometry. We then prove the feasibility of the devised approach and discuss and compare the results in detail.

4.3.1. Examples of application

Measurement object For our examples, we use the prior Eq. (4.35), which is well adapted to the reconstruction of rough technical surfaces. As such,

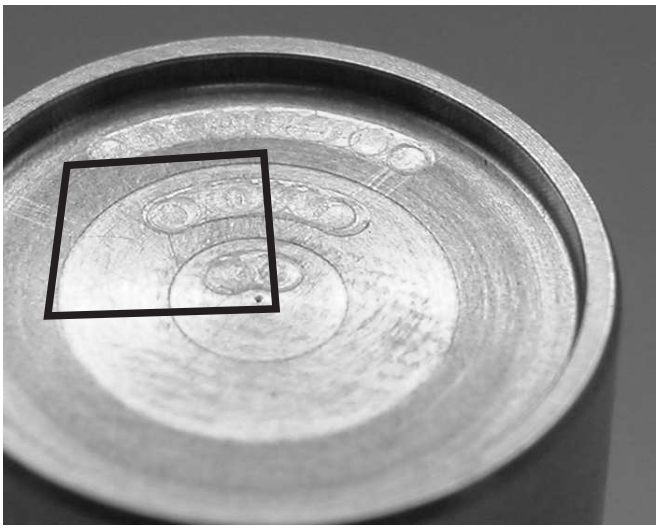


Figure 4.2: Photograph of a turned steel piece (diameter: 19.5 mm) used in the examples. The black frame indicates the approximate area of height measurements, slightly skewed in this perspective.

we present a turned steel piece (see photograph in Fig. 4.2). This sample piece is 19.5 mm in diameter. Is carries a circular outer rim of 1 mm width and 1 mm depth. In the inside region, it additionally features three circular areas of increasing depth towards the center. These areas are separated by steps of 20 μm height. Over a quarter sector of the piece three tracks are visible, one on each of the three areas. The latter features were added by using a worn turning tool, which leads to minor quality and presumably higher roughness and less distinguished, even rounded edges. In contrast, the three circular areas mentioned first were obtained with a new high-precision turning tool, such that their borders have precise edges. All these curved features represent nice challenges for the operations defined on a fixed rectangular grid. In the rotational center of the whole piece, a tip that remained from the turning process is barely visible. It could serve us as the experimentum crucis for an algorithm's discrimination power between singular surface features and outliers.

By our experience, the features exhibited by the surface of the sample piece can be considered typical of the challenges white light interferometry faces in industrial precision manufacturing.

Algorithmic realization The Bayesian estimation approach is implemented in C++ on a standard PC. In its core, the algorithm contains only two nested loops, where the a posteriori probability is calculated by multifold multiplications of small numbers.

In order to prevent underflows of variables and loss of numerical precision for mixed summations and multiplications, several measures have been taken.

scanning speed	feed per frame	phase feed per frame	10%-width of envelope
$2\,\mu$m/s	$0.04\,\mu$m	$0.09\,\pi$	397 frames
$14\,\mu$m/s	$0.28\,\mu$m	$0.61\,\pi$	57 frames
$28\,\mu$m/s	$0.56\,\mu$m	$1.36\,\pi$	28 frames
$56\,\mu$m/s	$1.12\,\mu$m	$2.72\,\pi$	14 frames
$84\,\mu$m/s	$1.68\,\mu$m	$4.07\,\pi$	9 frames
$112\,\mu$m/s	$2.24\,\mu$m	$5.43\,\pi$	7 frames

Table 4.1: Correspondences of technical details for measurements at different scanning speeds with a 50 Hz camera and a light source of mean wavelength $\bar{\lambda} = 825\,$nm (see text).

First of all, one can perform an overall rescaling, which is possible as we are only interested in modes of the posterior probability. Some groups of critical numerical operations are encapsulated and then rescaled or changed in execution ordered to group operations with variables of the same magnitude. Particular products of probabilities that have the tendency to get quickly too small are logarithmized.

In spite of this tuning, for the measurements presented in this section, on a 1.2 GHz P IIIm machine, the processing time varies between 31 s for 14 μm/s and 2.8 s for 112 μm/s, including the input of raw data from hard disk.

Data acquisition The algorithm makes use of a full 3-D set of raw intensity measurements. These data were obtained using a white light interferometer system, incorporating a near infrared LED light source, and produced by the University of Erlangen and 3D Shape GmbH. It is a prototypic realization of the approach presented in [Dresel et al., 1992] with the ability to output the full raw data acquired during a measurement scan.

The data were obtained working with several scanning speeds, ranging from $14\,\mu$m/s (which, at a frame rate of 50 Hz, corresponds to the Nyquist frequency of the interferograms inner oscillation for this light source), to $112\,\mu$m/s (which corresponds to an 8-fold subsampling of the inner oscillation). Table 4.1 gives some technical correspondences of this parameter.

It is of additional interest to see over how many frames the envelope of the interference oscillation spans at a certain scanning speed. This number can give us a rough hint about the chance to detect the interference and so about the reliability of extracting the correct height from the data sequence. Therefore we give the 10%-width of a Gaussian envelope, calculated with $l_c = 14.81\,\mu$m and $\sigma = l_c/4$ (from [Restle, 2003]) in Table 4.1. This value has no theoretical foundation and could be somewhat optimistic. The quite common $1/e$-width would correspond to 66% of the 10%-width.

For data acquisition, the system was set up under real-world conditions, on a vibration-isolation table in the normal laboratory environment. The room temperature remained constant within less than 1 K, but vibrations to the system

were present due to normal business traffic in the surrounding building. The white light interferometer was optimized in terms of its optical alignment and a sturdy fixation of the system and our sample. However, the illumination was chosen lower and less homogeneous than optimal in order to further challenge the processing algorithms.

4.3.2. Methods for quantitative comparison

In this section we discuss how to compare reconstructed height maps and how to find measures for the quality of a particular reconstruction.

Our objective is to state quantitatively, which of two height map reconstructions is better, and which algorithm performs better on a certain problem. That is, we wish to answer the question, which height map reconstruction $\hat{\boldsymbol{h}}_a$ or $\hat{\boldsymbol{h}}_b$ is nearer to the true height distribution.

In this context, we try to refine the notion of a "true" height distribution and discuss different measures for comparison.

A ground truth for height reconstruction Our first goal is to provide the "truth" to which to compare our measurements. Ideally, these data should be independent from a measurement technique, reproducible and commonly accepted. Apart from white light interferometry, a height map of comparable resolution for a rough surface could also be obtained with confocal microscopy and techniques of significantly higher resolution, like atomic force microscopy (AFM) or raster tunnel microscopy (RTM). For rough surfaces laser interferometry is generally not an option, as is requires prior knowledge for unambiguous reconstruction. There have been first efforts to establish a correspondence between results from white light interferometry and tactile devices, which are very widely spread in industry (cf. Sec. 2.4 and [Windecker and Tiziani, 1999]). This investigation is however founded on measurements of a standardized surface with highly parallel grooves, which makes it intrinsically one-dimensional. This allows for a comparison between 2-D maps from interferometry and 1-D profiles from the tactile device, but on the other hand, these results cannot be expected to carry over to real-world surfaces of stochastic variations in 2-D. For the other techniques mentioned, so far rather simple structures like binary gratings have been investigated and "correspondences" [Recknagel et al., 1998] for certain scalar roughness parameters have been established.

We assume that white light interferometry and confocal microscopy could deliver comparable results, as both approaches are based on light scattering and both share the same restrictions to lateral resolution due by the optical imaging system. This topic could however not be investigated and must be left to further research.

With the tools of AFM and RTM, the microstructure of technically worked surfaces can be resolved far beyond the limits of optical methods. With the use of tools that simulate the rough surface reflection of white light [Ettl, 2001], it is in principle possible to calculate the signal arising from the lower resolution of the interferometer's optical system. But as the basic interaction process

(induced static electrical fields in AFM and tunnel current in RTM) is different to white light interferometry, it generally cannot be expected that height maps were comparable in a straightforward manner. As before, this question should be further investigated before making use of these methods.

We conclude this excursion with the finding that any "true" height map cannot be reliably obtained at the current state of research. Instead, we choose to revert to interferometry itself to prepare a *reference height map*.

Reference height map for white light interferometry The reference height map for our interferometer has to be measured with the same device as the data later used for comparison to cancel influences as discussed above. However, this means that the results will be based on a partially recursive argument. We try to decimate this issue by the following construction for the reference map:

1. Fix optimal environmental conditions (illumination, vibrations, temperature, air flow, and others) for the data acquisition process.

2. Perform a series von N measurements at a slow scanning speed in a tight succession.

3. Calculate N height maps by a reliable preprocessing algorithm.

4. Remove overall shifts.

5. Calculate a robust average height map and perform a final outlier-detection and removal procedure.

There are of course alternatives to this specific method. However, the proposal appears to be the best what can be done without resorting to a complex post-processing procedure. This ought to be avoided as this could introduces new deviations or conceal artifacts from the surface.

For the construction of the reference height map, measurements were recorded at a slow scanning speed of $2\,\mu$m/s. The data of one recording (no. 19) was obviously spoilt due to some error during image capture, which manifested in form of a gap in the recorded time series; it had to be removed, thus $N = 24$. The remaining time series were ensured to be free of deterioration, by visual inspection of the raw data.

We applied the *sliding average* algorithm (cf. Sec. 2.2.1), which is a recommended linear filtering algorithm [Schraud, 2000]. Overall shifts of the height maps were detected with an image-wise median filtering operation and successively canceled by shifting the height maps to a common reference height. Higher order errors such as tilts or large-scale deformations were not accounted for. From the reconstructed maps $\hat{\boldsymbol{h}}_i$ a common average map $\boldsymbol{h}_{\text{med}}$ was constructed with the median operator, applied pixel-wise over all height maps:

$$\boldsymbol{h}_{\text{med}} = \text{med}\{\hat{\boldsymbol{h}}_1, \ldots, \hat{\boldsymbol{h}}_N\} \tag{4.51}$$

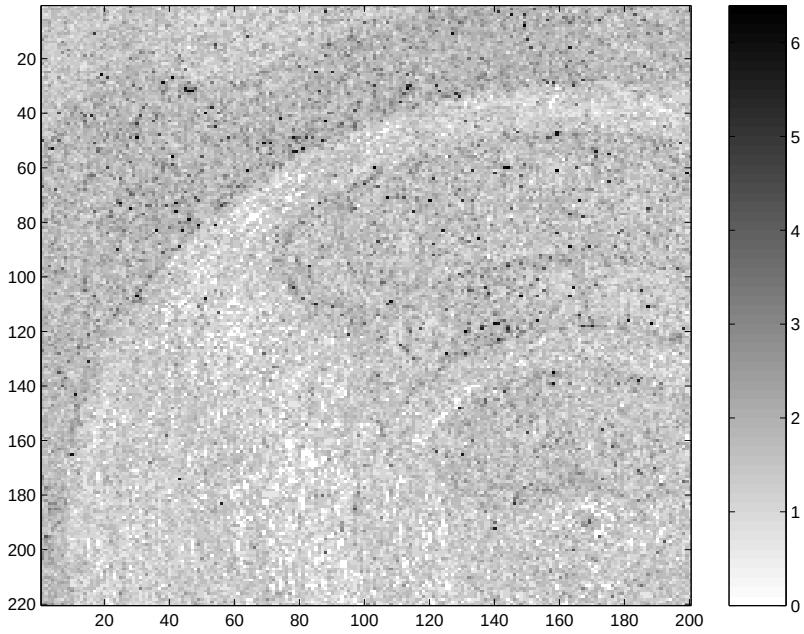


Figure 4.3: Spatial distribution of logarithmized $\mathrm{MAD}^{(\mathrm{ref})}$-values (Eq. (4.52)) for the data used to create the reference height map.

This map exhibits a small number of invalid pixels, as can be expected for a rough surface with a static speckle image. By visual inspection, these sites stand out by their large variation of height values reported from preprocessing. This observation gives a heuristic means to identify these outliers.

As a variation measure, we use the robust MAD operator (median absolute deviation), with the following definition[1]:

$$\mathrm{MAD}^{(\mathrm{ref})}\{\hat{\boldsymbol{h}}_1,\ldots,\hat{\boldsymbol{h}}_N\} = \mathrm{med}\{|\hat{\boldsymbol{h}}_1 - \boldsymbol{h}_{\mathrm{med}}|,\ldots,|\hat{\boldsymbol{h}}_N - \boldsymbol{h}_{\mathrm{med}}|\} \tag{4.52}$$

Fig. 4.3 shows the spatial distribution of $\mathrm{MAD}^{(\mathrm{ref})}$-values for the data sequence used to create the reference height map. To enhance visibility, the values are logarithmized and the gray value encoding is the darker the larger a pixel's value is—an approach we will use frequently in the course of this chapter. The $\mathrm{MAD}^{(\mathrm{ref})}$ data contain some zero values, for which the logarithm cannot be evaluated. In that case the values not available are replaced by zero again. Therefore values of zero fall together with those next to zero in a logarithmized figures, provoking a little error in these visual representations.

Those pixels for which the MAD exceeds a threshold k we mark as defective in order to replace them later. This threshold can only be chosen heuristically, this is where the recursive nature of our bootstrapping strategy becomes apparent. Fig. 4.4 shows the distribution of MAD values for $\boldsymbol{h}_{\mathrm{med}}$. If one assumes the histogram were made up from a highly populated distribution peaking around

[1] Note that a variety of definitions for a MAD exists, and sometimes (notably with some versions of MATLAB [The Mathworks, Inc., 2002]) an additional factor is introduced to normalize the respective definition with the standard deviation of a normal distribution.

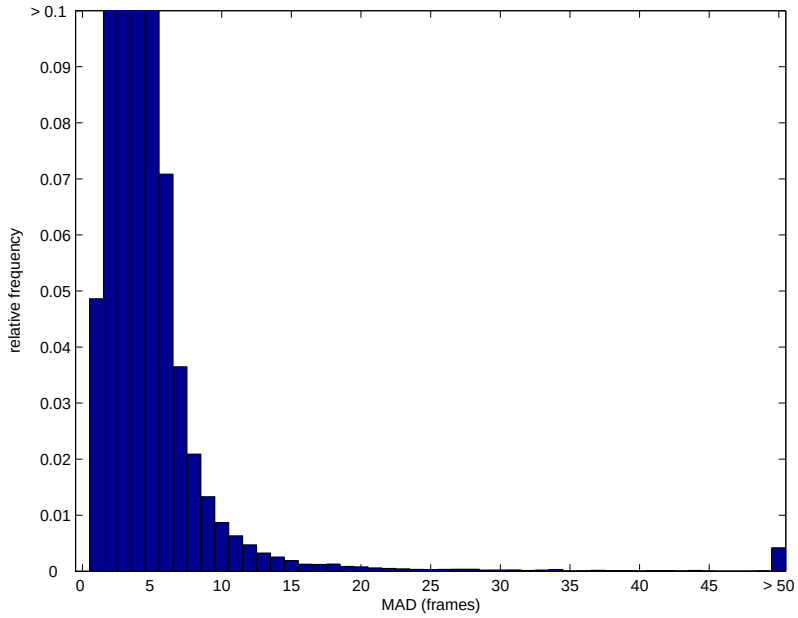


Figure 4.4: Relative histogram of $\mathrm{MAD}^{(\mathrm{ref})}$ in the reference height map. The last entry shows the cumulated sum of all larger entries, the ordinate is cut at 0.1.

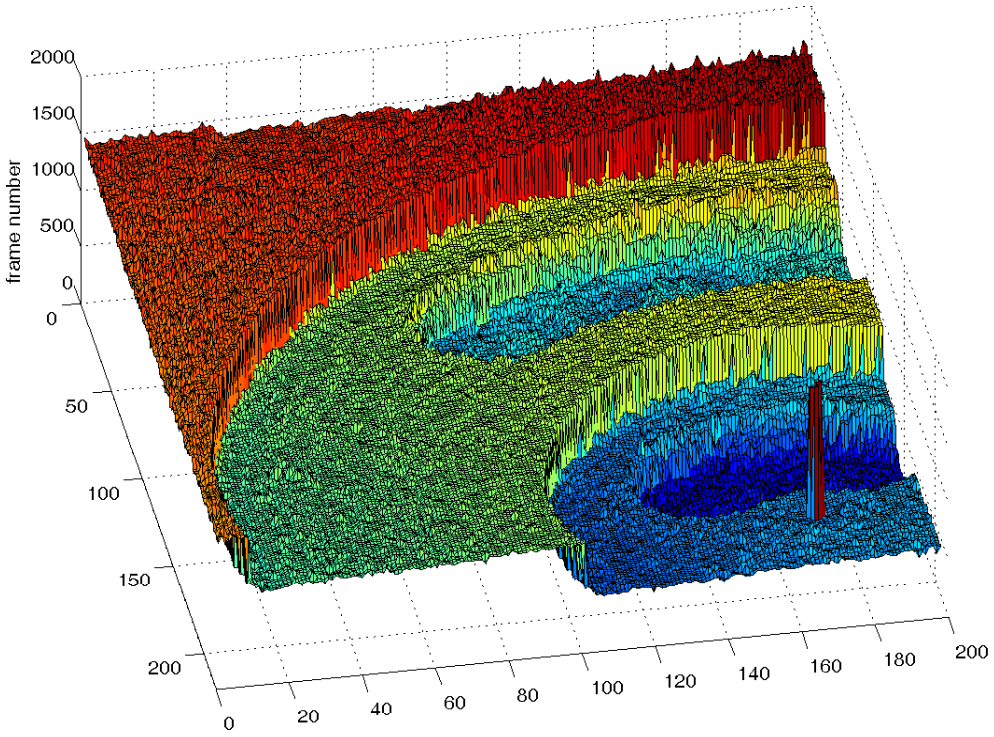
$3-5$ frames, dropping exponentially to the right, and a somewhat uniform distribution due to invalid pixels, the borderline can be drawn somewhere around $25-35$ frames. Thus we choose $k = 30$ frames. Again, this argumentation has to be taken cum grano salis: The possible alternative approach would be a statistically sound estimation of the threshold with the tools of decision theory. While not knowing the functional description for the tail of the left distribution, this seems disproportionate in the light of the remaining uncertainties.

The pixels marked as defective are finally replaced with the median of their spatial neighborhood within $\boldsymbol{h}_{\mathrm{med}}$. In our experiments, the invalid pixels are so sparse that the 3×3 filter $\mathrm{med}_{3\times 3}$ seems sufficient. Thus the reference height map is complete:

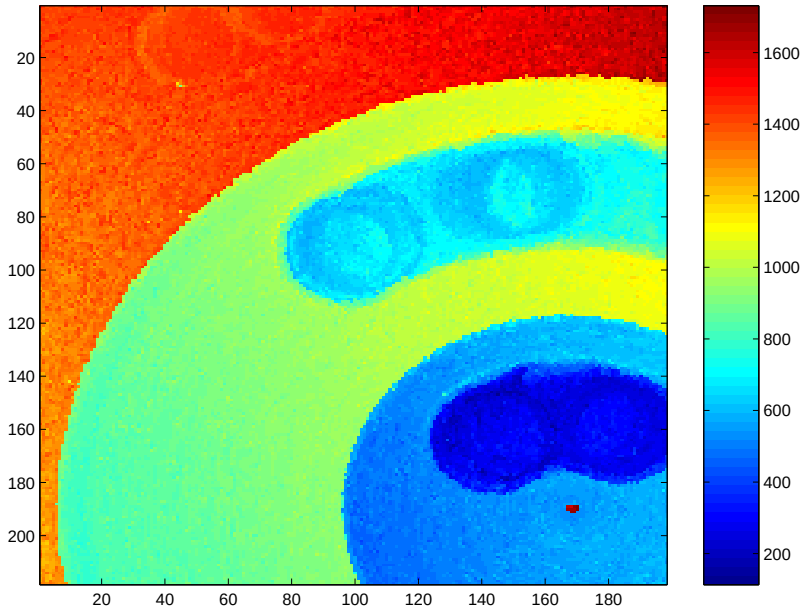
$$\boldsymbol{h}_{\mathrm{ref}} = \begin{cases} \mathrm{med}_{3\times 3}\, \boldsymbol{h}_{\mathrm{med}} & \text{for } \mathrm{MAD}\{\boldsymbol{h}_{\mathrm{med}}\} > k \\ \boldsymbol{h}_{\mathrm{med}} & \text{otherwise} \end{cases} \tag{4.53}$$

In the end, we have a reference height map $\boldsymbol{h}_{\mathrm{ref}}$ which, without additional knowledge about the true surface, we consider free of errors (see Fig. 4.5).

Removal of overall shifts, rescaling and border effects The interferometer setup used in our experiments has no absolute scale for axial movement measurements. The height values calculated are therefore based on an arbitrary starting condition, usually the height at the start of a scanning procedure is declared as the zero height level. The experiments are usually conducted in tight sequence to avoid the influence of gradual shifts in the environment. However,



(a)



(b)

Figure 4.5: The reference height map. (a) shows a bird view, (b) a color encoded map of the same scene. The scale of the height axis is 0.04 μm per frame.

it can happen that the recording system randomly misses out a height step indicated from the movement stage, which is one technical reason for deterioration of the interference signal, in this case for the full image.

As this effect is not reproducible, we at least try to compensate for overall shifts within a set of reconstructed height map. This is done by shifting each map so that the map-wide height median is zero:

$$\boldsymbol{h}_l' = \boldsymbol{h}_l - \mathrm{med}_j(h_l)^j \qquad \text{shift correction} \tag{4.54}$$

For comparisons with a reference map, the investigated height map is shifted to match the median height of the reference map, respectively.

In order to reduce discretization effects and rounding errors, the data obtained at higher scanning speeds where they are allocated to their coarser height scale are rescaled to match the scanning speed of the reference height map for the comparison. We will mostly use the plain integer unit "frames" in our comparison, with the correspondence ranging from $0.04\,\mu$m height difference per frame at $2\,\mu$m/s scanning speed, to $2.24\,\mu$m per frame at $112\,\mu$m/s.

Next the borders of an image, spatial filters require special measures, as the filter mask spans over non-existing parts outside the image. For this comparison, generally 3×3 pixel filter masks, respective neighborhood relations are used. Then the results for a borderline of one pixel width around each map are removed from evaluation, also the plotted figures are cut down accordingly.

Simple comparison For an overview comparison, a detailed error map is less practical. Instead, we compute a scalar error estimate which can be used to both provide a coarse quality measure and to tune the respective parameters of compared algorithms. For N measurements of each $|\mathcal{S}|$ pixel and a distance measure ρ, the average error per pixel ("pp") is:

$$\bar{\mathcal{E}}_{\mathrm{pp}} = \frac{1}{N} \cdot \frac{1}{|\mathcal{S}|} \cdot \sum_{l=1}^{N} \sum_{j=1}^{|\mathcal{S}|} \rho\left((h_{\mathrm{ref}})^j, (\hat{h}_l)^j\right) \tag{4.55}$$

The measure $\rho(\cdot,\cdot)$ can be adjusted to reflect the costs of an erroneous height estimation in an actual application. Without, we refrain to linear costs for our experiments:

$$\rho\left((h_{\mathrm{ref}})^j, (\hat{h}_l)^j\right) = \left|(h_{\mathrm{ref}})^j - (\hat{h}_l)^j\right| \qquad \text{for pixel } j \tag{4.56}$$

and just plainly use the symbol $\bar{\mathcal{E}}_{\mathrm{pp}}$ for the error from definition Eq. (4.55) calculated with this measure, i. e., the average *absolute* error per pixel. This figure is not a robust estimator, so it strictly penalizes left-over outliers compared to the reference map.

Extended comparison An analysis reaching beyond the average absolute error can possibly give some insight into the question what reconstruction approach to follow for a given data quality. It is therefore sensible to take the original data

quality into consideration when comparing with the performance of different reconstruction algorithms.

We thus choose classify the data by two features:

- The magnitude of the absolute error $\mathcal{E}_{\text{abs}}(h^j)$, taken as the absolute deviation of the estimated height from the height fixed by the reference map:

$$\mathcal{E}_{\text{abs}}(h^j) = \left|(h_{\text{ref}})^j - (\hat{h})^j\right| \qquad \text{for pixel } j \tag{4.57}$$

- The median absolute deviation $\text{MAD}(h^j)$, as an overall measure of the difficulty of obtaining a correct estimate of a pixel's height by an algorithm—its "measurability".

Surface "measurability" assessment The latter feature should help in a more detailed analysis of the power of an approach in particularly handling poorer data quality caused by the surface under measure. To that end, the quality of the raw data is assessed with the difficulty of obtaining a correct estimate for a pixel. For each pixel site of a test surface, the differences between repeated height reconstructions gives base to a measure of how reliable this site is to reconstruct.

We propose to rather use a simple and less robust approach for height reconstruction, which shows a sensitive reaction to low data quality. Our choice in these experiments is the established sliding average algorithm. The calculated average absolute errors displayed in Fig. 4.25 support this choice: The larger the scanning speed during data acquisition is, the fewer frames contain the interference signal (cf. Table 4.1). Consequently, one expects a strong correlation between sample variance in height values and scanning speed for non-robust reconstruction.

For measuring the differences, we use the reference height map (Fig. 4.5) as a gold standard and compare pixel-wise each map reconstructed with the sliding average algorithm against it. The differences are measured as the absolute deviation. Out of these numbers, the median value over the set of reconstructions is calculated and can be visualized as a 2-D map. This is done for each scanning speed anew, so that we obtain for each speed a map of "measurability" values. On one hand, this makes comparisons between different scanning speeds much less reliable, but on the other hand, we can ignore the drastic changes in the qualitative nature of the raw data across different scanning speeds. So we focus on what can be obtained inherently from the data at each acquisition speed and only compare to one reference height map.

Of course, a number variations to this approach are possible and viable. Starting with the difference measure, our use of the absolute deviation suggests linear costs underlying, which is merely ad hoc and could be inappropriate for other actual applications. We chose to compare against the reference height map, recorded outside of the current measurement series and in some patches postprocessed using spatial filters. As an alternative, one could choose a map

synthesized for each scanning speed as the pixel-wise mean or median as a reference. In that case however, we would be blind to systematic errors arising within a set of reconstructions, as could be a height "needle", then completely missing due to poor measurability.

For assessing the surface "measurability", we then arrive at using a formula very similar to the one used to mark defective pixels in the reference map build-up, cf. Eq. (4.52):

$$\mathrm{MAD}\{\hat{\boldsymbol{h}}_1, \ldots, \hat{\boldsymbol{h}}_N\} = \mathrm{med}\{|\hat{\boldsymbol{h}}_1 - \boldsymbol{h}_{\mathrm{ref}}|, \ldots, |\hat{\boldsymbol{h}}_N - \boldsymbol{h}_{\mathrm{ref}}|\} \tag{4.58}$$

Fig. 4.6 shows maps of the MAD values obtained from reconstructions of $N = 25$ measurements taken at 14 μm/s, the highest speed possible for Nyquist-sampling of the interferogram's inner oscillation. In the linear scale diagram, only a few points in the map call to attention: prominently in the lower right area around to the needle (left over from the turning process), which suggests that the measurements were spatially not perfectly aligned. Among the points scattered all over the map a higher density can be made out along the edges of the turned piece, which is not surprising either. From the logarithmic scale plot, one better make out the overall level of variability in the height reconstruction. The lower right turned area and the traces brought in by a worn turning tool exhibit a slightly higher level of variability. For the latter feature, this of course comes not unexpected. For the higher variability of the lower right area, the reasons are not as clear and we assume a combination of focus shift and locally poorer illumination.

Fig. 4.7 shows the same data, now accumulated for the equal number of measurements taken at 84 μm/s. This speed yields data of particularly poor quality, as the 1:4-correspondence between sampling rate and periodicity of the interferogram leads to strong suppression of the signal. One basically sees the same features as in Fig. 4.6, now on a much larger scale, visible even though the MAD-values are logarithmized. In addition, in the middle/upper right area groups of very bright colored pixels poke out. These have MAD-values of around 1000 frames. Upon closer inspection of the raw data one sees that the heights of these pixels have collectively been misdetected to similar values far off the reference height. The reason could be a short shaking or some other transient irregularity of the full setup. The horizontal elongation of these occurrences can be attributed to the directionality of the turned piece under inspection, and gives a good impression on how delicate the effects here are.

4.3.3. Settings for assessment

Interpretation of 2-D histograms With the two pieces of information available for each pixel, the magnitude of the absolute estimation error and the median absolute deviation as a signal quality measure, we can perform an extended comparison of the different approaches for height reconstruction.

For visualization we choose 2-D histograms, in which the information is aggregated over all pixels and encoded in gray scale density for each bin. For further comparison, plots of difference of two histograms are created which

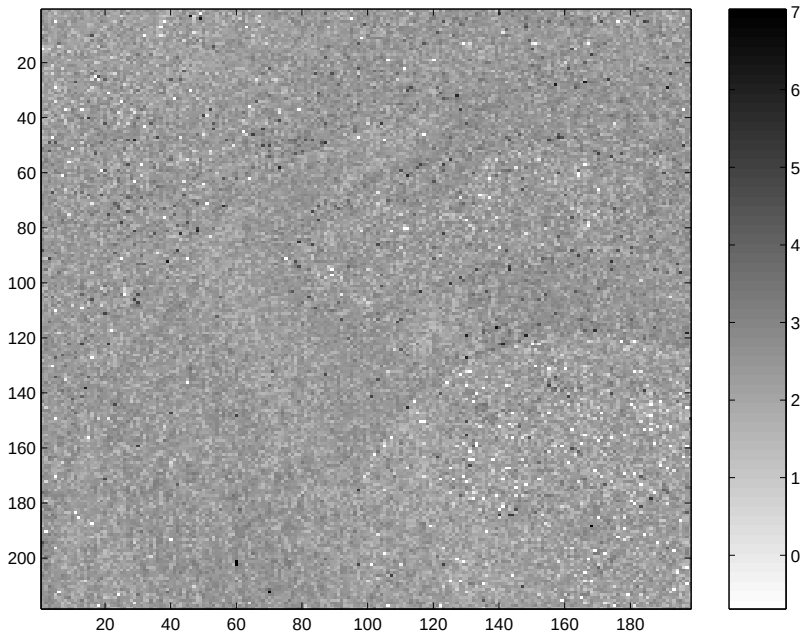


Figure 4.6: Spatial distribution of the pixel-wise logarithm of MAD-values against the reference height map for a series of $N = 25$ reconstructions with only preprocessing applied. Scanning speed is $14\,\mu$m/s, scale is $0.04\,\mu$m per frame.

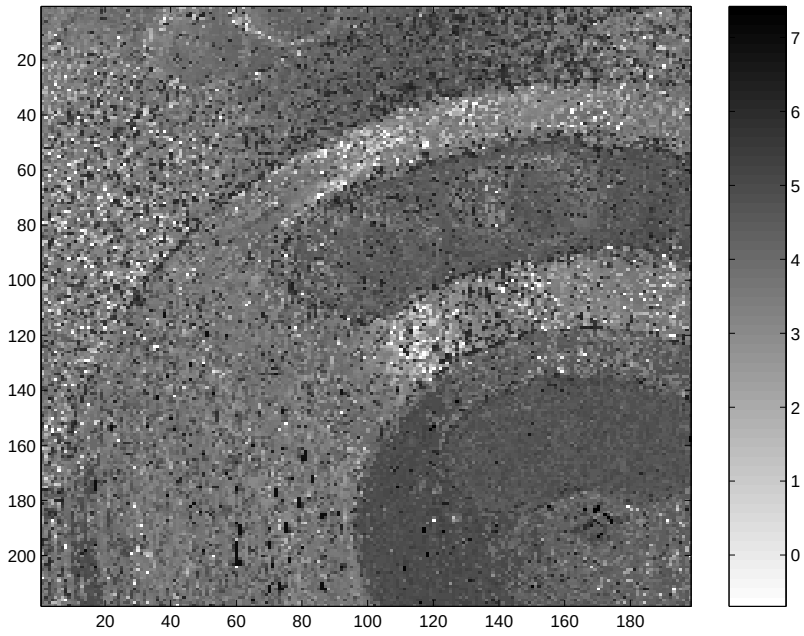


Figure 4.7: Same as Fig. 4.6, but now with a scanning speed of $84\,\mu$m/s, scale is $0.04\,\mu$m per frame.

show the power of an approach relative to the power of another approach. In such a figure it is of course not visible how the magnitude of error is shifted between individual pixels. As the MAD-value is a fixed property of the pixels for one scanning speed, the density between two histograms compared can only be shifted along the estimation error axis. thus the sum of occupancies for one MAD-value bin remains constant and is zero for difference histograms.

Above all of this we should be aware that the data going into comparison does not cover all possible combinations of features and are only sparse in some areas—therefore the interpretations are of varying accuracy.

Algorithms under comparison For all experiments, we use raw data that was previously recorded in $N = 25$ sequences for each of the recording speeds given in Table 4.1, apart from the $2\,\mu$m/s data, which is used to build up the reference map. The data is read from hard disk, the processing times we state therefore only cover the pre- and postprocessing stages but not the acquisition effort.

Preprocessing only We have used height maps provided with the sliding-average preprocessing algorithm (cf. Sec. 2.2.1) to prepare the reference height map (cf. Sec. 4.3.2). For rough surfaces, this pixel-wise approach is prone to errors from instable reflection conditions. Especially for higher scanning speeds, the height map obtained by this algorithm is heavily contaminated with outliers that ought to be removed by some kind of postprocessing, otherwise future analysis is difficult. Hence we include these results primarily to allow for an unbiased comparison of the other methods.

For the algorithms under comparison, Table 4.2 contains the best parameter settings for the different scanning speeds. For the method of Bayesian estimation and the approaches involving spatial filter masks, we always use a 3×3 pixel square mask for the neighborhood definition.

Median filter mask As we have seen in Sec. 2.3.2 filtering with the median filter mask is a simple and common image processing approach, and as such it is included in the comparison. The median operation provides maximum robustness [Donoho and Huber, 1983] which makes it particularly suitable tor remove outliers while preserving edges in the height map. Other than by the mask size, which we set to 3×3 for all algorithms, this filter cannot be parameterized. We use a standard MATLAB implementation of the median filter mask which processes a height map in about 0.27 s.

Adaptive median filter Given the main drawback of the median filter, its tendency to oversmooth valid information we include this approach into the comparison. Only those pixels which are suspected outliers due to the larger-than-normal MAD in their spatial surrounding are subjected to filtering, the others are left intact (cf. Eq. (2.75)). The threshold (parameter c in the equation) for the filtering decision is chosen so that for each scanning speed the average absolute error per pixel $\bar{\mathcal{E}}_{\mathrm{p}}$ (cf. Eq. (4.55)) of the data sequence is minimized.

scanning speed	size of dataset	preprocessing window size	adaptive median threshold c
14 μm/s	289 frames	9 frames	15.4
28 μm/s	144 frames	5 frames	10.7
56 μm/s	72 frames	3 frames	10.5
84 μm/s	49 frames	2 frames	0
112 μm/s	37 frames	2 frames	0

scanning speed	nonparametric smoothing limit	Bayesian estimation parameters Λ	q_0/q_1
14 μm/s	4.9 (frames)^2	6 frames	10^{-2}
28 μm/s	2.9 (frames)^2	4 frames	10^{-2}
56 μm/s	1.6 (frames)^2	4 frames	10^{-4}
84 μm/s	2.8 (frames)^2	5 frames	10^{-4}
112 μm/s	1.3 (frames)^2	3 frames	10^{-4}

Table 4.2: $\bar{\mathcal{E}}_{\mathrm{pp}}$-optimum parameters for the different algorithms when applied on the turned piece data, found by empirical screening of a small range typical for this surface.

Nonparametric smoothing In Sec. 2.3.3 we have presented a recently developed postprocessing algorithm which uses confidence information to minimize information loss in a linear smoothing approach by adapting and weighting the variable width filter mask. The core algorithm does not require setting of parameters, but it uses a mapping of the confidence values as calculated by the interferometer setup to a pre-estimated ensemble variance of the preprocessed height map. In addition, for the recursive implementation we use the spatial variance of the resulting height map is an adjustable termination condition. It is selected such that a minimum average absolute error $\bar{\mathcal{E}}_{\mathrm{pp}}$ (cf. Eq. (4.55)) is achieved.

The nonparametric smoothing algorithm can be significantly impaired by a poor correlation of confidence values to the variance, cf. to [Restle et al., 2004] for a discussion. For this benchmark, we have calculated the variance of the preprocessed ensemble separately and fed it directly into the postprocessing algorithm. Therefore we can avoid this issue by not using the confidence values at all.

Bayesian estimation We use the Bayesian estimation with a rectangle prior, cf. Eq. (4.35). Its width Λ and the quotient q_0/q_1 are set by empirically screening a small range of values that have been found typical of the surface to be reconstructed. The optimum parameter set is the one with the least average absolute error $\bar{\mathcal{E}}_{\mathrm{pp}}$.

4.3.4. Detailed comparison

We discuss the $N = 25$ measurements taken at a scanning speed of $28\,\mu$m/s, which is double the speed that corresponds to the Nyquist rate for the setup we used, $14\,\mu$m/s. While the speed-up is not that large, at this scanning speed we can nicely demonstrate the benefits and drawbacks of the different approaches under comparison.

Estimation with only preprocessing applied With only preprocessing, the first measurement is reconstructed as depicted in Fig. 4.8. One can see numerous peaks, spiking both to the lower and upper end of the height range. These are outliers, which could not be assigned a proper height value due to invalid raw data. Depending on details of the algorithm implementation and feeble variations of the signal, the outliers most often come up with height values at either limits of the output range.

From the lower figure, it seems as if the outliers are somewhat more frequent in the upper left area and the lower middle field.

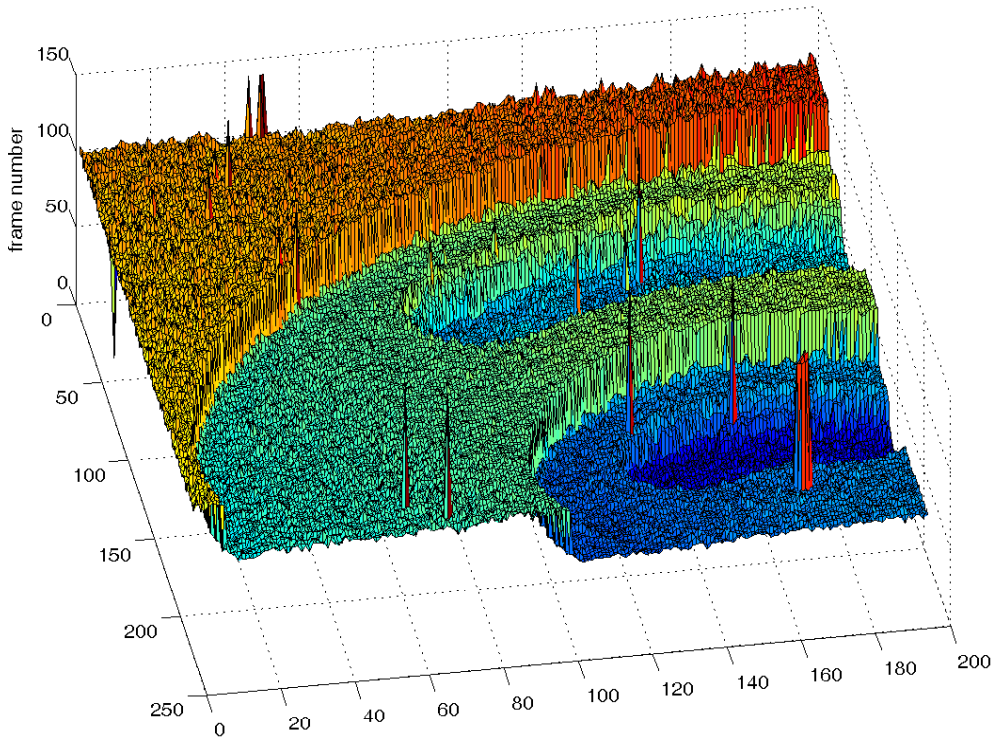
The tip left-over from the piece's manufacture is a prominent feature in the lower right area, it is slightly larger in area than the surrounding outlier peaks.

As the preprocessing does not work with any spatial or neighborhood information, the form of all edges of the test piece should be reconstructed very close to the optimum. One can make out that the edges brought in by the rework using a worn turning tool are significantly more rounded than the sharp edges left over from the fresh tool.

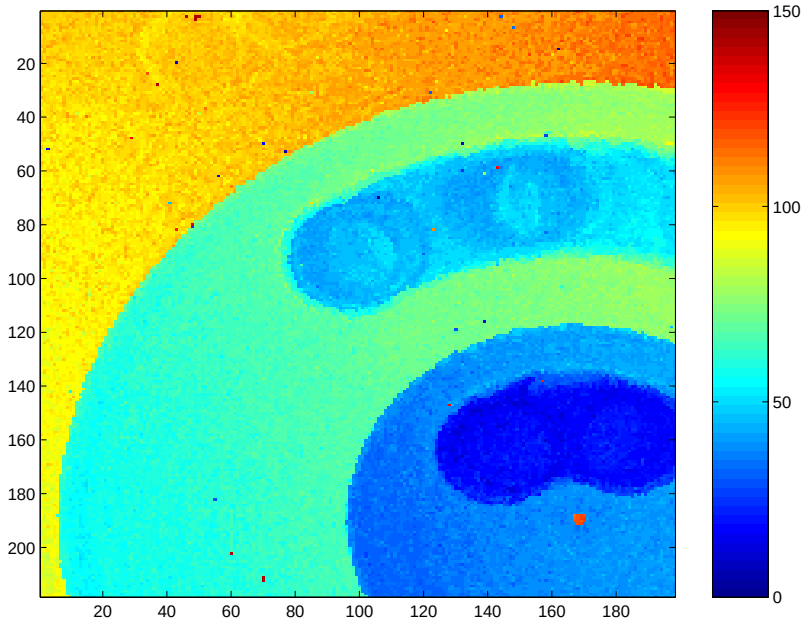
The average deviation of the height maps obtained after only preprocessing against the reference height map are registered with the respective MAD-values. Fig. 4.9 shows the MAD-values for each pixel of the sequence of $N = 25$ recordings taken at the same scanning speed. A logarithmized scale is used to enhance visibility of the plot. One can see that reconstructions are mostly of low MAD-values, thus each reconstruction is reliable in general. The surface structure of the test piece can still be made out. Outliers with a large MAD-value are marked in dark and are scattered all over the surface. They occur more often, but still stochastically scattered, near edges and next to the tip-like manufacture artifact in the lower right corner.

The histogram in Fig. 4.10 shows the frequencies of MAD-values for this sequence of reconstructions. For the ordinate, a logarithmic scale is chosen to better visualize the occupancies for larger MAD-values. However, the vast majority of pixels have a small variability in the height estimate of around 1 to 7 frames, corresponding to a height variation of 0.56 to $3.92\,\mu$m. This lies near the upper limit of the surface roughness range we expect for metal piece measured. One sees that the minimum MAD-value is not the most frequent outcome.

For a more detailed analysis, we correlate for each pixel the absolute estimation error throughout the sequence to the difficulty to obtain a correct result, as measured by the pixel-wise spread of the estimates, i. e., MAD-value as given in Fig. 4.9. The result is presented in a 2-D histogram in Fig. 4.11, with the



(a)



(b)

Figure 4.8: Reconstruction with preprocessing only for data obtained at 28 μm/s. (a) shows a bird view, (b) a color encoded map of the same scene. The scale of the height axis is 0.56 μm per frame.

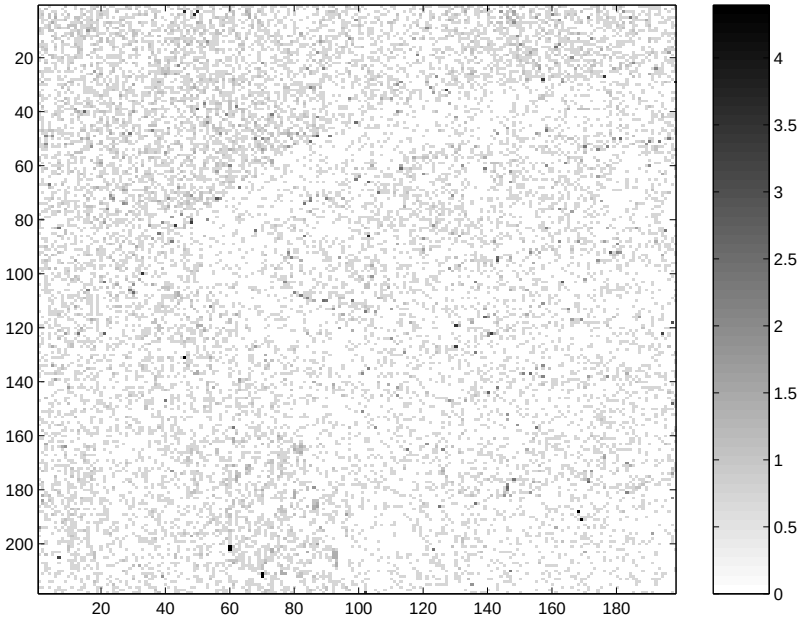


Figure 4.9: Robust measurement of the average reconstruction error for a series of $N = 25$ measurements with only preprocessing applied: spatial distribution of the pixel-wise logarithm of MAD-values (measured in frames) against the reference height map. Scanning speed is $28\,\mu$m/s, scale is $0.56\,\mu$m per frame.

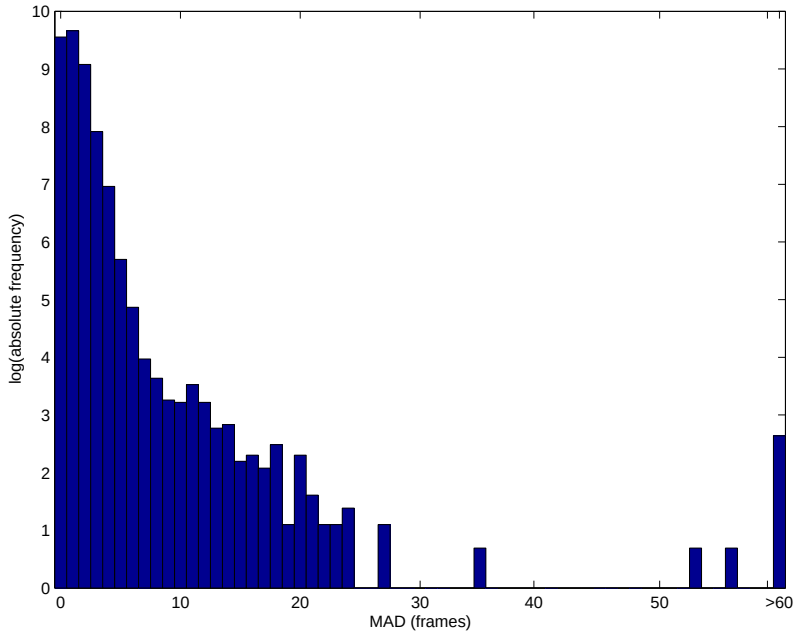


Figure 4.10: Logarithm of the absolute frequency of MAD-values for the sequence of height maps obtained after only preprocessing. Scanning speed is $28\,\mu$m/s, scale is $0.56\,\mu$m per frame.

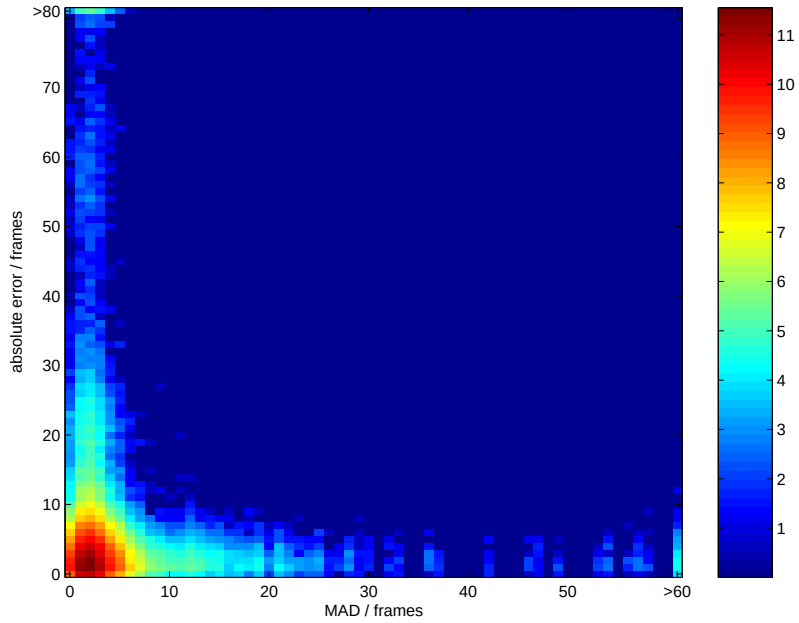


Figure 4.11: Distribution of errors in the sequence of height maps obtained after only preprocessing. The occupancies of the 2-D histogram are logarithmized and plotted as a function of the absolute estimation error and the robust variability (MAD-value). Scanning speed is 28 μm/s, scale is 0.56 μm per frame.

occupancies for each bin (in 2-D, this corresponds to a box) logarithmized for the visual representation. That is, the histogram Fig. 4.10 is split up in a second dimension according to the absolute error of each pixel.

One can see that the data gathers in two ridges along the axes, with a very strong concentration in the lower left corner. Along the horizontal axis, most of the entries have a MAD-value of 7 frames or lower, and beyond of about 30 frames, the entries become sparse. The vertical ridge is located in a region of 7 frames or lower for the MAD. The most significant contribution of entries is estimated with an absolute error of 10 frames or smaller, the majority bears an error of 1 to 5 frames. The entries on the vertical beyond that stem from pixels which, at least in some measurements, show off as outliers with an overly large estimation error. In general, we would expect entries of large absolute error also for the high-MAD region[1]. These are missing, which seems to be mainly a result of a selection effect, i. e., the data of high-MAD regions is too sparse and the estimation of those pixels probably too volatile—and sometimes even accurate. In Fig. 4.12 the horizontal and vertical axes are scaled logarithmically, used to visually enhance (zoom in) the high-occupancy area in the lower left of Fig. 4.11.

[1] Note that the MAD-value is calculated against the reference height (Eq. (4.58)), not intrinsic to the data set as with $\mathrm{MAD}^{(\mathrm{ref})}$ (Eq. (4.52)). Therefore absolute deviations smaller than the MAD-value are possible.

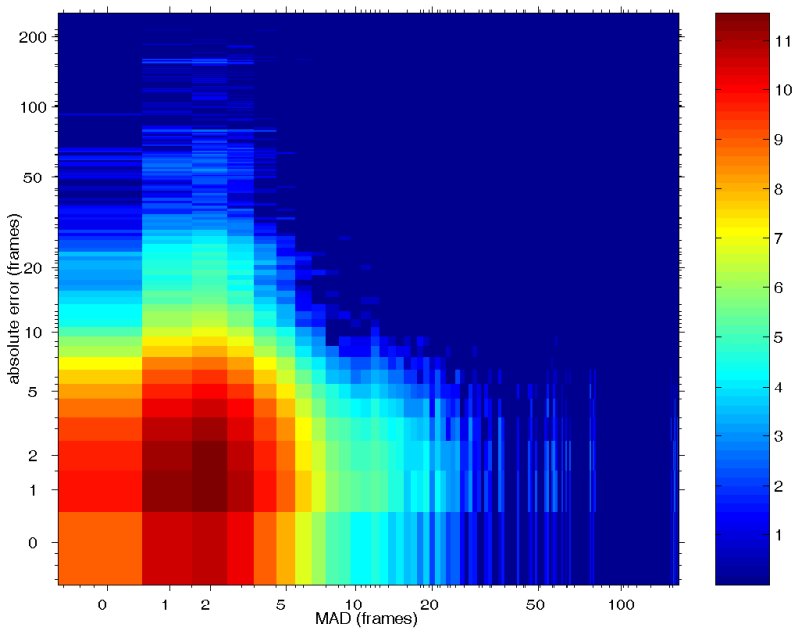
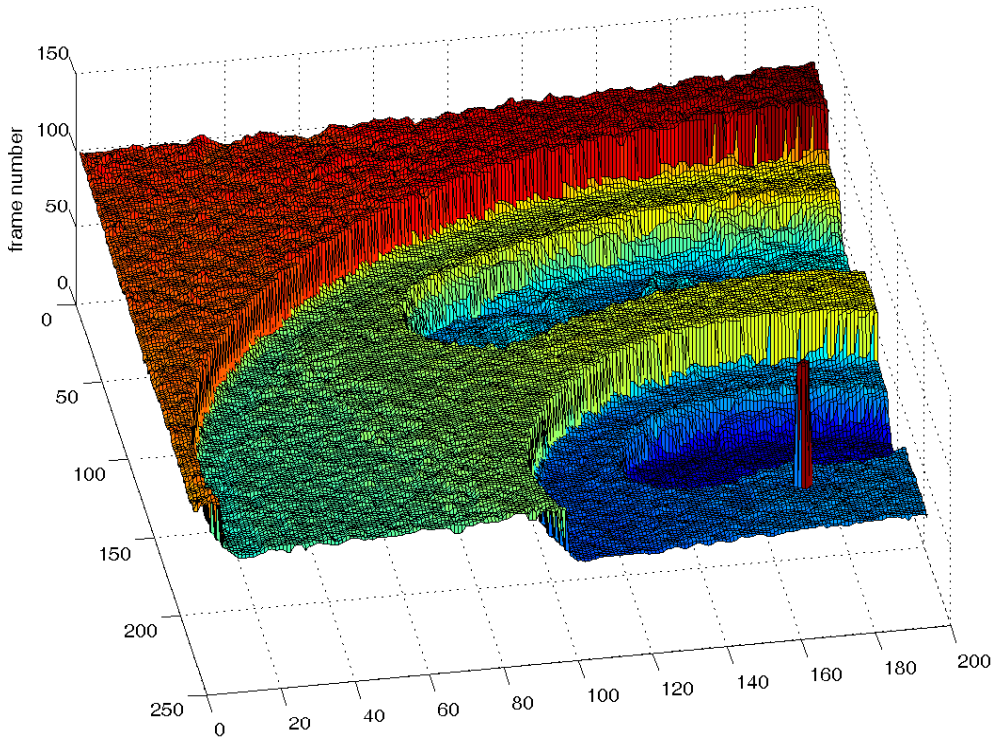


Figure 4.12: Same as Fig. 4.11, but with a logarithmic scale on both axes for visual enhancement of the high-occupancy area.

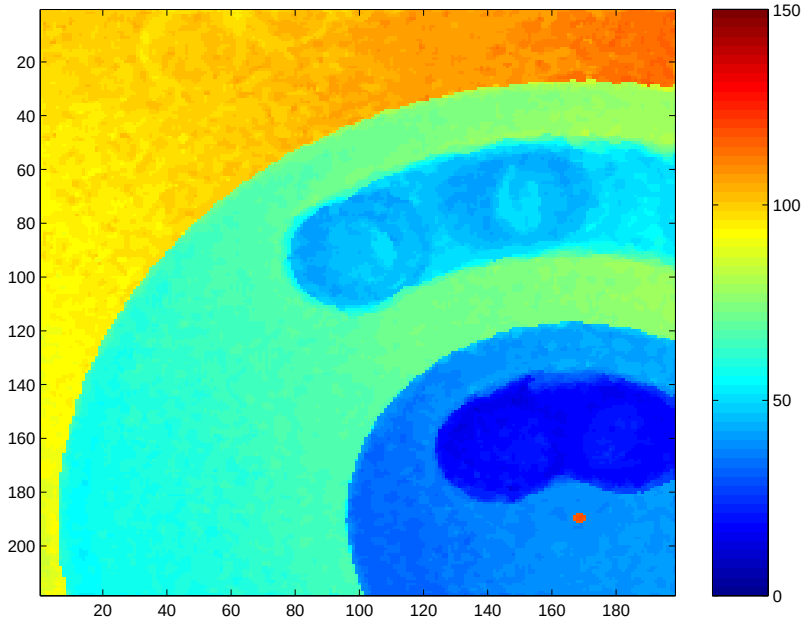
Median-filter postprocessing With a 3×3 median filter mask applied as postprocessing, the first measurement of the sequence is estimated as shown in Fig. 4.13. As one can see, the outliers are removed and replaced by a value from their neighborhood. The structure of the surface on small scale, especially visible in Fig. 4.13a, is less coarse than without median filtering (Fig. 4.8a) and appears more smooth and wavy. While the steep edges made with the new turning tool are as crisp as in Fig. 4.8b, the edges obtained from the worn tool are more rounded and less irregular (Fig. 4.13b).

For a more detailed discussion we look into the absolute deviation achieved with the filtering operation. The MAD-values obtained with only preprocessing applied (Fig. 4.9) serve us as an estimate of the "measurability" if form of the variability in repeated recordings. Similar to the preprocessing-only case, figs. 4.14 and 4.15 show 2-D histograms of occupancies for combinations of absolute deviation after postprocessing and related MAD-value. Upon direct comparison, only subtle differences can be made out. The effects of the median filter, evident in the height map, such as removal of outliers and blurring of sharp edges, find no obvious correspondence in the error histograms just mentioned. However, the strong accumulation of pixels in an error range of 1 to 5 frames in Fig. 4.15 gives again hint at the surface roughness of the piece measured. We can expect it in the order of 0.5 to 3 μm on the scale of the 3×3 pixel patches, that is a side length of about 100 μm.

To better assess the performances, we instead inspect the differences in occupancy of the 2-D histograms for postprocessing by median filtering and preprocessing-only. This is done in Fig. 4.16 for the important area of highest



(a)



(b)

Figure 4.13: Reconstruction with median filter postprocessing applied for data obtained at 28 μm/s (cf. Fig. 4.8). (a) shows a bird view, (b) a color encoded map of the same scene. The scale of the height axis is 0.56 μm per frame.

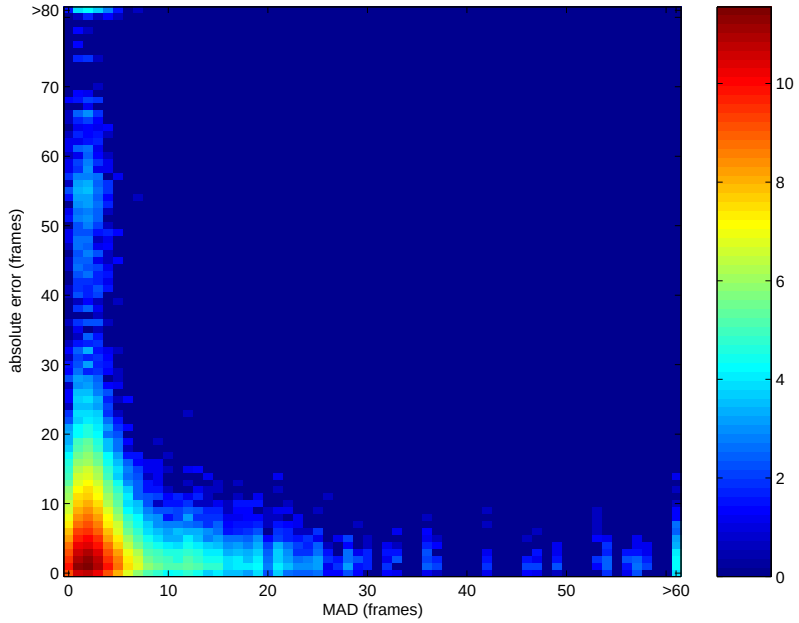


Figure 4.14: Distribution of errors in the sequence of height maps obtained after postprocessing with a 3×3 median filter. Settings of the figure cf. Fig. 4.11.

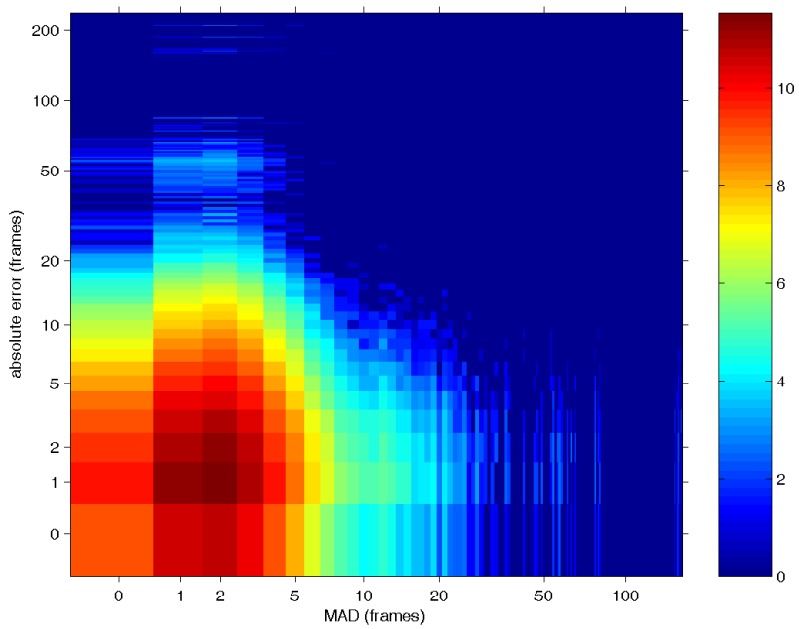


Figure 4.15: Same as Fig. 4.14, but with a logarithmic scale on both axes for visual enhancement of the high-occupancy area.

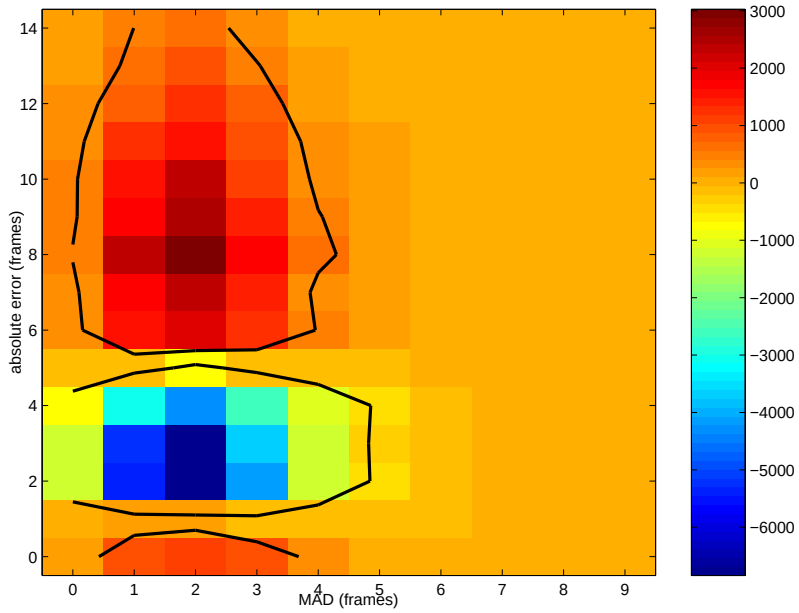


Figure 4.16: Error histogram for 3×3 median filter processing minus the error histogram for preprocessing only. Scale is $0.56\,\mu$m per frame, the black contours are drawn at $\pm 20\,N$, with $N = 25$ scans.

occupancy in the lower left corner. Here, the effect of the median filter on the bulk of ordinary pixels can be studied, the outliers do not fall in this region. The black contour is drawn rather arbitrarily at the occupancy levels of $\pm 20\,N$ (N the number of scans in the data taking session) to visually emphasize the differences. For the interpretation one has to keep in mind that the MAD-value is fixed for each pixel, therefore the different algorithms lead to reallocations only within columns of the same MAD.

The blue region is the error range of about 2 to 4 frames, where less pixel are found after the median filtering compared to the case of only preprocessing. This mass is "moved" primarily towards a larger absolute deviation, with an error of about 6 to 12 frames. Additionally, a smaller surplus for the median filter can be found with very small errors of 1 or zero frames. These findings show that the height variation within a spatial neighborhood—where the median filter acts—is usually larger than the uncertainty within a single pixel, as we have discussed above. The median filter in most cases introduces values from mismatching probability distributions where the spatial and ensemble statistics disagree.

In summary, we can state that the median filter not only efficiently removes outliers, but also deteriorates the bulk of regular preprocessed height estimates, leading to larger errors than before. The reason is excessive smoothing, performed also in regions of the height map which do not require it. Linked to this is an obliteration of the surface microstructure, as can be seen vividly in comparison of Fig. 4.13a with Fig. 4.5a and even with Fig. 4.8a.

Adaptive median postprocessing The adaptive median filter is particularly suited to reduce the oversmoothing tendency of the classic median-filter. As described in Sec. 2.3.2, the smoothing is applied only to the outliers which are found by a Hampel detector. One can see in the reconstruction Fig. 4.17.

With very few exceptions, this postprocessing removes all outliers from the map. In addition, the microstructure of the surface is not as smooth as after the simple median filter. Instead it more resembles the reconstructions obtained after preprocessing-only (Fig. 4.8) or the reference height map (Fig. 4.5). Both the edges worked of the new and the worn turning tool are reconstructed sharply, the blur that is partly present after application of the simple median filter cannot be observed.

The filter threshold (Eq. (2.75) has been chosen so as to minimize the average absolute error $\bar{\mathcal{E}}_{\mathrm{pp}}$. This error measure puts a linear penalty on the distance of outliers. A quadratic measure which would penalize outliers stronger would find its minimum at a smaller threshold. This would inhibit outliers even better, but also more regular pixels would be identified as outliers and filtered with the median operator. The price to pay for stronger suppression of outliers is the oversmoothing in the bulk of pixels.

The 2-D difference histogram comparing the residual errors after the adaptive median filter to that after only preprocessing is presented in Fig. 4.18.

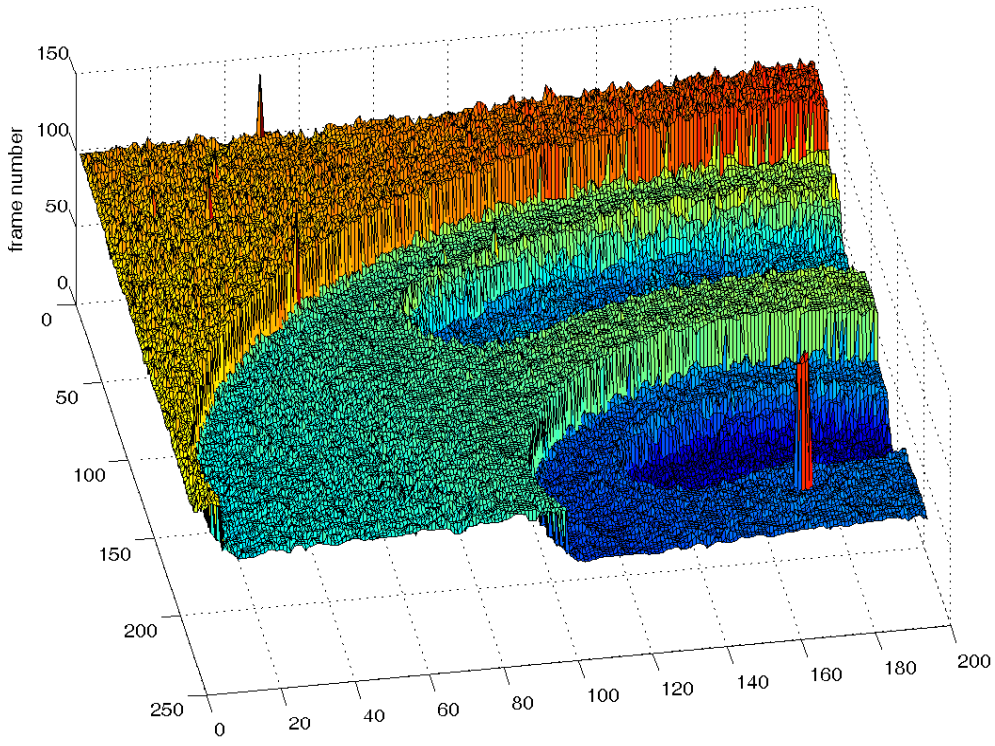
Note that the color scale and the color of zero difference are individual for each difference histogram. Comparing this figure to the case of a simple median filter (Fig. 4.16), the overall magnitude of differences is much smaller: the adaptive median filter changes less height values than the median filter. However, similar to the median filter, the adaptive median still leads to larger absolute errors for the mass of pixels with small MAD-values.

Postprocessing with nonparametric smoothing The nonparametric smoothing, with its target variance chosen to minimize $\bar{\mathcal{E}}_{\mathrm{pp}}$, yields a crisp reconstruction with only few outliers, shown in Fig. 4.19.

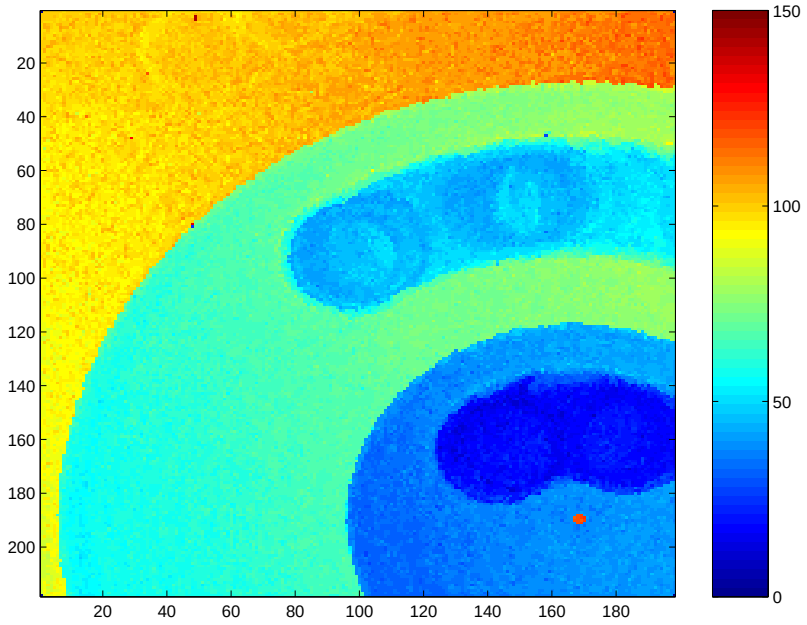
The height map resembles the map obtained by the adaptive median filter. In Fig. 4.19a a clear representation of surface microstructure is perceptible. The edges and the tip-like machining residue in the lower right corner are well represented. As one can see in Fig. 4.19b, even the sharp edges introduced by the new turning tool are crisp without blurring, which is remarkable for a postprocessing based on linear filtering.

The 2-D histogram of occupancy differences in Fig. 4.20 (nonparametric smoothing minus preprocessing only) is different from the diagrams presented up to now. Here, two clear bulges are visible, but now with their signs exchanged, compared to before: The postprocessing is able to reduce the absolute error for the pixels in the low-MAD region, which is the most densely occupied region on the MAD-scale. A larger number pixels are found to have an error of 4 or less frames, while a field with an error of 5 to 10 frames—a little more diffuse to circumscribe—is depleted in compensation.

We can therefore expect height maps from nonparametric smoothing to be more precise for pixels representing the surface microstructure. With the me-



(a)



(b)

Figure 4.17: Reconstruction with the adaptive median postprocessing applied for data obtained at 28 μm/s (cf. Fig. 4.8). (a) shows a bird view, (b) a color encoded map of the same scene. The scale of the height axis is 0.56 μm per frame.

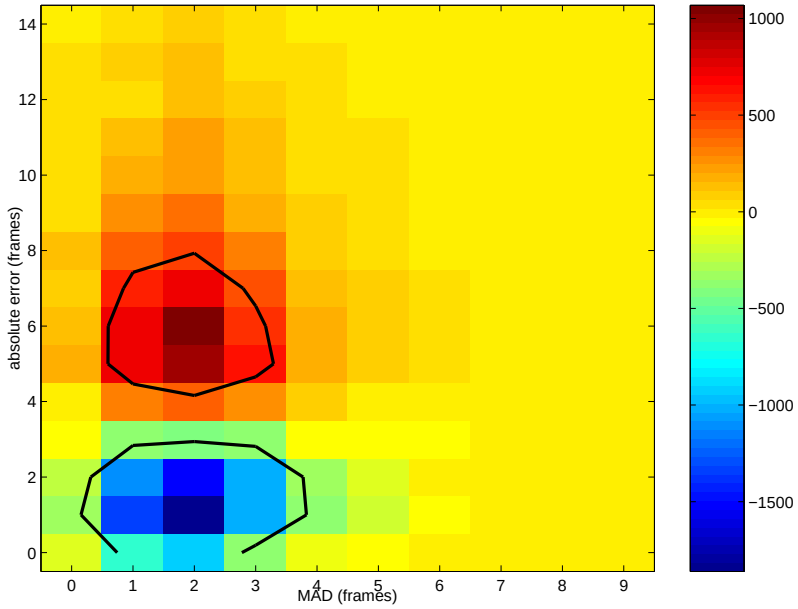


Figure 4.18: Error histogram for the adaptive median filter minus the error histogram for preprocessing only. As before, the scale is $0.56\,\mu$m per frame, the black contours are drawn at $\pm 20\,N$, with $N = 25$ scans.

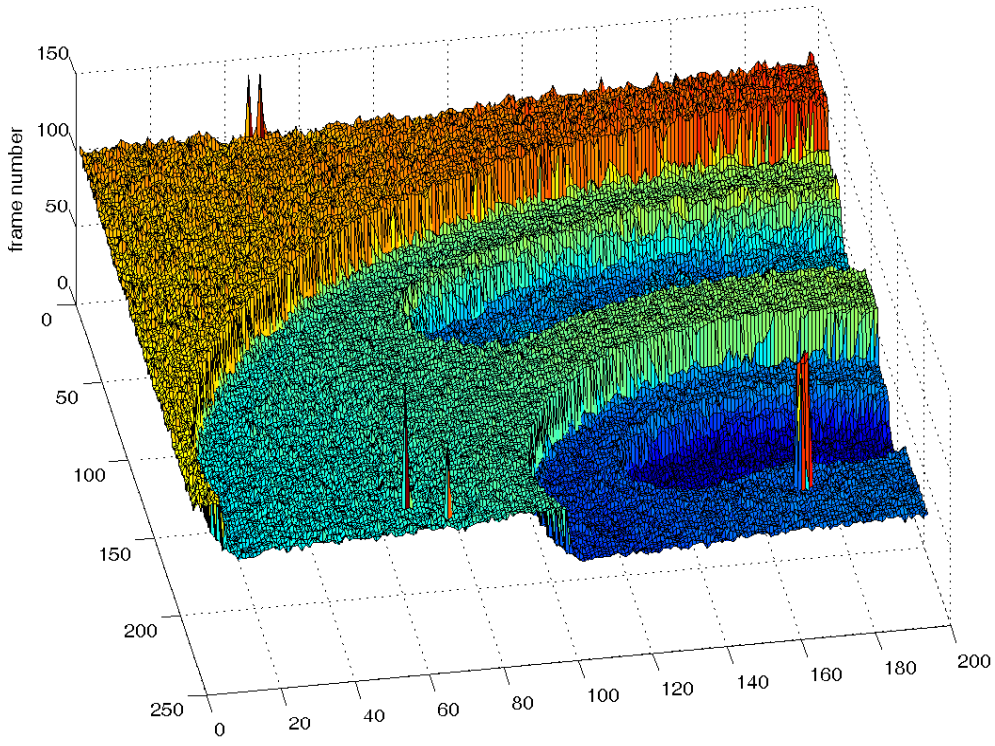
dian filter based approaches, for the removal of outliers the price of a higher error with these pixels had to be paid.

Bayesian surface estimation The Bayesian approach introduced in Sec. 4.2 estimates the height map as shown in Fig. 4.21. The map here has a similar appearance as the map obtained with the adaptive median filter and the nonparametric smoothing. The roughness and small-scale variance of the surface microstructure is close to the reference height map. The clear-cut edges worked with the new turning tool as well as the are somewhat fuzzy edges introduced with the worn tool are accurately represented. However, a small number of outliers is still apparent.

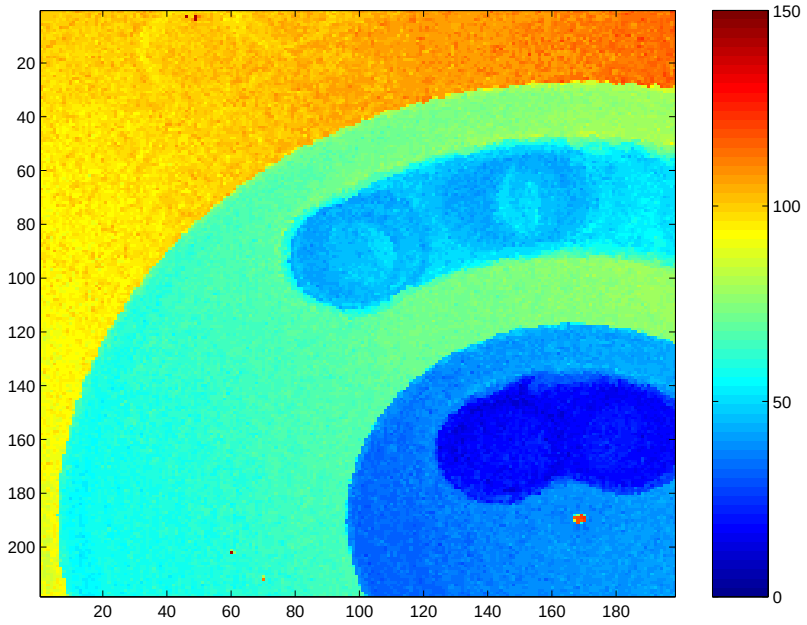
It is possible to remove these residues as well by a different choice of the parameters (the q_0/q_1-ratio). The parameters of the algorithm are optimal according to the $\bar{\mathcal{E}}_{\mathrm{pp}}$ error measure (cf. Table 4.2), and similar to the case of the adaptive median filter such a setting would have an inferior performance for the bulk of the whole surface.

The 2-D histogram of occupancies for the Bayesian estimation minus preprocessing only in Fig. 4.22 differs remarkably from the plots shown for the first three approaches, but resembles the plot obtained for the nonparametric smoothing algorithm.

The positive bulge at very low absolute errors of 0 to 1 frame show that more pixels are estimated with minimum error than with the preprocessing-only approach, which is different from all other postprocessing algorithms considered.



(a)



(b)

Figure 4.19: Reconstruction with postprocessing by nonparametric smoothing for data obtained at 28 μm/s (cf. Fig. 4.8). (a) shows a bird view, (b) a color encoded map of the same scene. The scale of the height axis is 0.56 μm per frame.

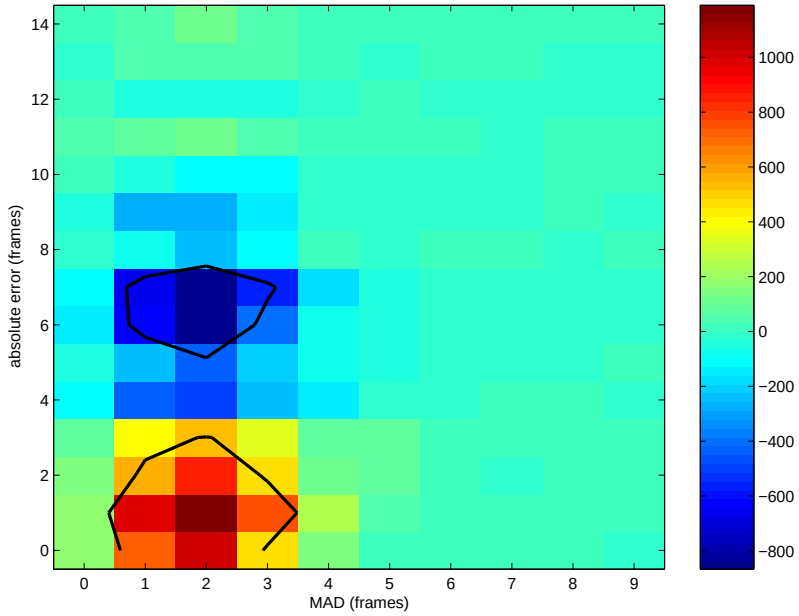


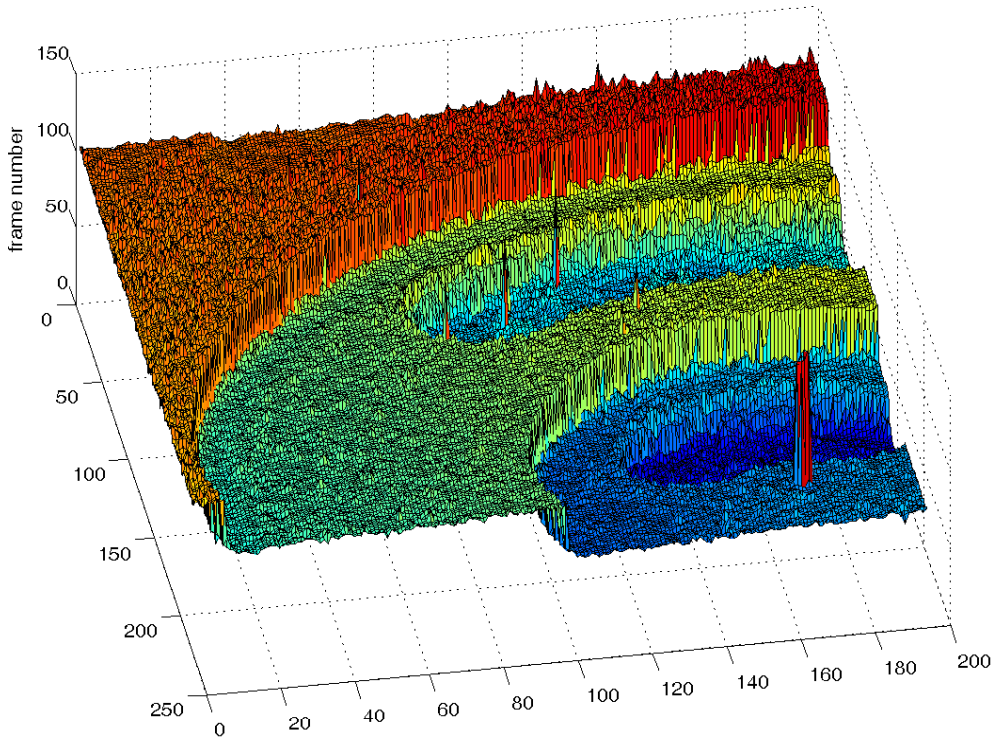
Figure 4.20: Error histogram for nonparametric smoothing minus the error histogram for preprocessing only. Scale is $0.56\,\mu$m per frame, the black contours are drawn at $\pm 20\,N$, with $N = 25$ scans.

The area where most pixels of the original reconstruction reside (cf. Figs. 4.14 and 4.15) correspondingly has, around 2 to 4 frames of absolute error, a significantly lower occupancy than after preprocessing-only.

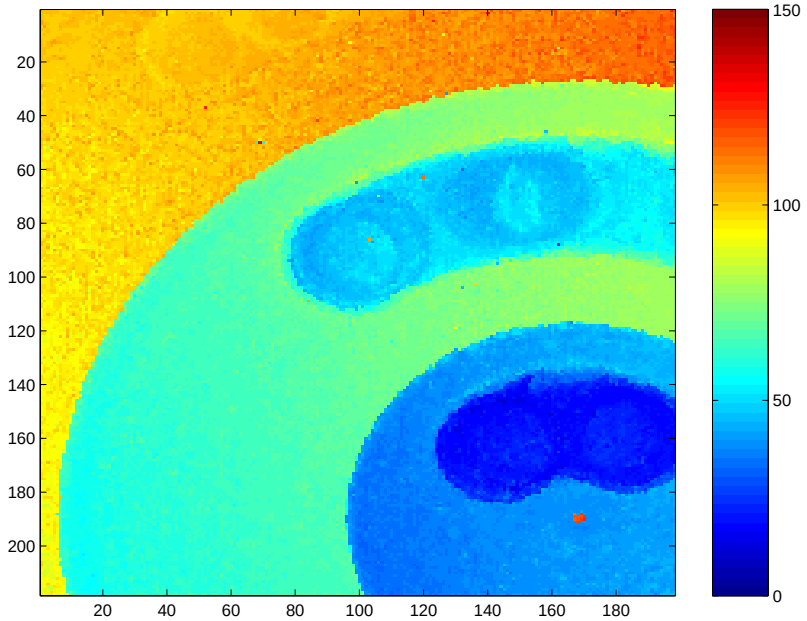
The data basis for large MAD-values is quite sparse, therefore only a summarized discussion of the error distribution is reasonable. The occupancy differences of Bayesian estimation minus preprocessing-only are displayed in Fig. 4.23a, now taken for the sum of MAD-values > 20 frames. In addition, in Fig. 4.23b the cumulative sum of this graph is displayed. One can consider these plots as originating from a line-wise summation of Fig. 4.22 for the larger MAD-values not shown in that figure. For the high-MAD fraction, more pixels are estimated with the Bayesian approach towards a very small absolute error of 0 to 1 frame than after preprocessing only. For higher absolute error, the occupancy difference is oscillating, with a bias to the negative side, this is the region depleted after Bayesian estimation. From Fig. 4.23b we can see that before reaching an absolute error of about 20 frames, the differences are essentially completely leveled. The Bayesian estimation reduces the errors also for the region with a higher variability of pixels difficult to estimate at all.

When we compare the analysis and the 2-D histograms for nonparametric smoothing and this approach, the two algorithms obviously behave similar in nature. In addition, according to the overall error measure $\bar{\mathcal{E}}_{\mathrm{pp}}$ both approaches perform equally well, and somewhat better than all other approaches for this scanning speed (cf. Table 4.3).

It is interesting to see the differences in detail. In the 2-D histogram Fig. 4.24, the difference in occupancy for the Bayesian estimation minus nonparametric



(a)



(b)

Figure 4.21: Reconstruction with Bayesian surface estimation obtained for raw data recorded at 28 μm/s. (a) shows a bird view, (b) a color encoded map of the same scene. The scale of the height axis is 0.56 μm per frame.

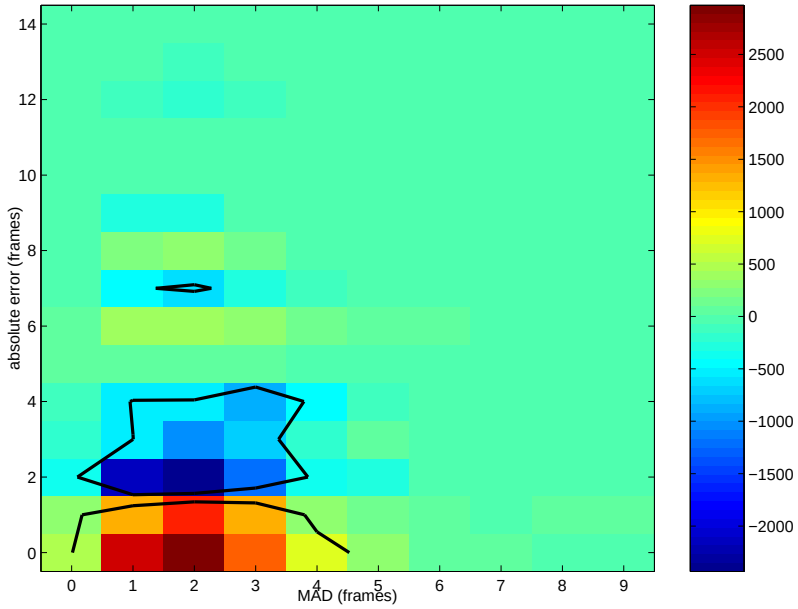
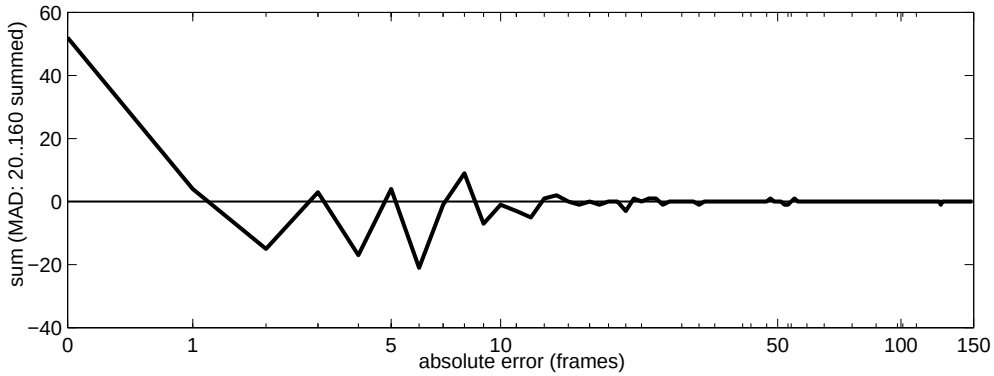


Figure 4.22: Error histogram for Bayesian estimation minus the error histogram for preprocessing only. Scale is 0.56 μm per frame, the black contours are drawn at $\pm 20\,N$, with $N = 25$ scans.

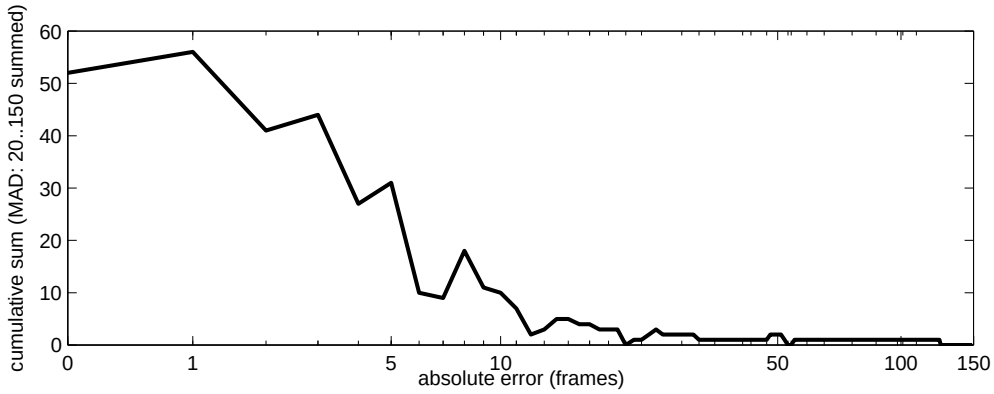
smoothing is plotted. The Bayesian approach shows higher occupancy for the area of very small absolute error (0 to 1 frames), as well, to a smaller degree, an area of about 6 frames. In contrast, the area of 2 to 4 frames is depleted. The Bayesian estimation therefore leads often to a smaller absolute error than the nonparametric smoothing, as well a less often to a larger absolute error. While both approaches achieve almost the same average absolute error, the Bayesian approach often succeeds in reconstructing the height with the minimum possible error—at the price of sometimes reconstructing with an error that is larger than the average error remaining after postprocessing with nonparametric smoothing. Looking back at Fig. 4.22, the infrequent tendency of the Bayesian approach to fall short of the true height by some frames can be anticipated only faintly from the plot.

Summary: Postprocessing for 28 μm/s measurements This scanning speed provides us with preprocessed data that, on one side, carries a significant number of outliers one likes to remove, and on the other side that suffers noticeably from conventional postprocessing, here in form of the median filter. The deterioration of the surface microstructure caused by the median filter leads to an actually larger average absolute error $\bar{\mathcal{E}}_{\mathrm{pp}}$ then for the preprocessed only data, in spite of its outliers. The adaptive median algorithm can circumvent this, as is displaces height values from the bulk of pixels to a smaller degree only.

Both the nonparametric smoothing and the Bayesian estimation approach succeed in delivering an even smaller average absolute error. They keep the residual error small for the bulk of pixels—those that make up the bigger part



(a)



(b)

Figure 4.23: Domain of MAD > 20 frames for the histogram in Fig. 4.22: (a) shows the horizontal sum and (b) the corresponding cumulative sum of the occupancy difference for Bayesian estimation minus preprocessing only for MAD > 20 frames. Scale is $0.56\,\mu$m per frame.

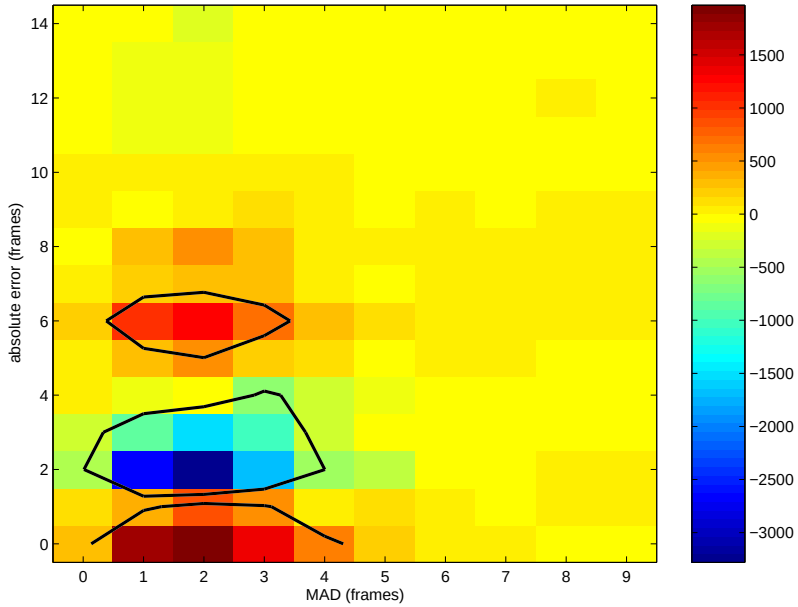


Figure 4.24: Error histogram for Bayesian estimation minus the error histogram for nonparametric smoothing. Scale is $0.56\,\mu$m per frame, the black contours are drawn at $\pm 20\,N$, with $N = 25$ scans.

of the surface, but also miss slightly more outliers. Finally, all approaches leave the tip-like artifact in the lower right corner of the surface intact.

For the tunable algorithms in this comparison, the parameters as optimized with the $\bar{\mathcal{E}}_{\mathrm{pp}}$-measure yield settings with do not completely abandon outliers, although this can be made possible with other settings for all algorithms. If, in practice, outliers are critical, one could therefore either modify the parameterization further towards a stronger smoothing, or on could use a measure for parameter optimization that penalizes outliers harder.

Scanning speed	Preprocessing only	Nonparametric smoothing	Median filter	Adaptive median	Bayesian estimation
14 μm/s	0.55 μm	0.53 μm	0.69 μm	0.56 μm	0.45 μm
28 μm/s	0.71 μm	0.68 μm	0.78 μm	0.72 μm	0.68 μm
56 μm/s	1.27 μm	1.15 μm	1.02 μm	1.02 μm	0.89 μm
84 μm/s	4.84 μm	2.64 μm	2.43 μm	2.43 μm	1.59 μm
112 μm/s	2.69 μm	1.91 μm	1.64 μm	1.64 μm	1.17 μm

Table 4.3: Absolute error per pixel ($\bar{\mathcal{E}}_{\text{pp}}$), cf. Eq. (4.55), for the algorithms under comparison, cf. also Fig. 4.25. The measurements for 84 μm/s suffer from the particularly low signal-to-noise ratio for this scanning speed (cf. Sec. 4.3.2).

4.3.5. Further results

The analysis that has been detailed in Sec. 4.3.4 for the scanning speed of 28 μm/s can be performed in the same manner for the other scanning speeds. In this section, we only discuss the central results. One can see from Table 4.3 and Fig. 4.25 that the absolute error grows the larger the scanning speed becomes. The main reason for this increase is that the interference signal becomes shorter for higher speeds, but also the discretization of the calculated height values, which gets proportionally larger with increasing speed (cf. Table 4.1) contributes to the error. An interpolation of the raw data or, in case of Bayesian estimation, of the a posteriori probability along the height axis could reduce the impact of discretization.

For better readability, the figures with reconstructed height maps are all moved to Appx. A. Similar to the discussion of the 28 μm/s measurements in Sec. 4.3.4, an analysis of the distribution of residual error versus data quality, i. e. the 2-D histograms, leads to similar findings for other scanning speeds. In that way, the 28 μm/s measurements are fairly representative for the whole set of scanning speeds, and we therefore omit a detailed discussion of the 2-D histograms in a move to save paper.

Measurements with 14 μm/s scanning speed For these measurements, the absolute errors for all four approaches lie closely together. The Bayesian estimation approach shows the minimum error of $\bar{\mathcal{E}}_{\text{pp}} = 0.45\,\mu$m, followed by the results obtained from the nonparametric smoothing. With the preprocessing only approach, a barely larger average error can be achieved, and only slightly worse than this the adaptive median performs. Last is the median filter, whose performance has a larger gap to that of the other approaches. This is however not surprising as the $\bar{\mathcal{E}}_{\text{pp}}$ measure rewards a better representation of the bulk of pixels higher than the removal of—at this scanning speed only few—outliers.

For data recorded at 28 μm/s, the absolute errors obtained with the different algorithms in our comparison take the same order as for 14 μm/s and differ only little in the absolute magnitude. As we have seen in Sec. 4.3.4 for

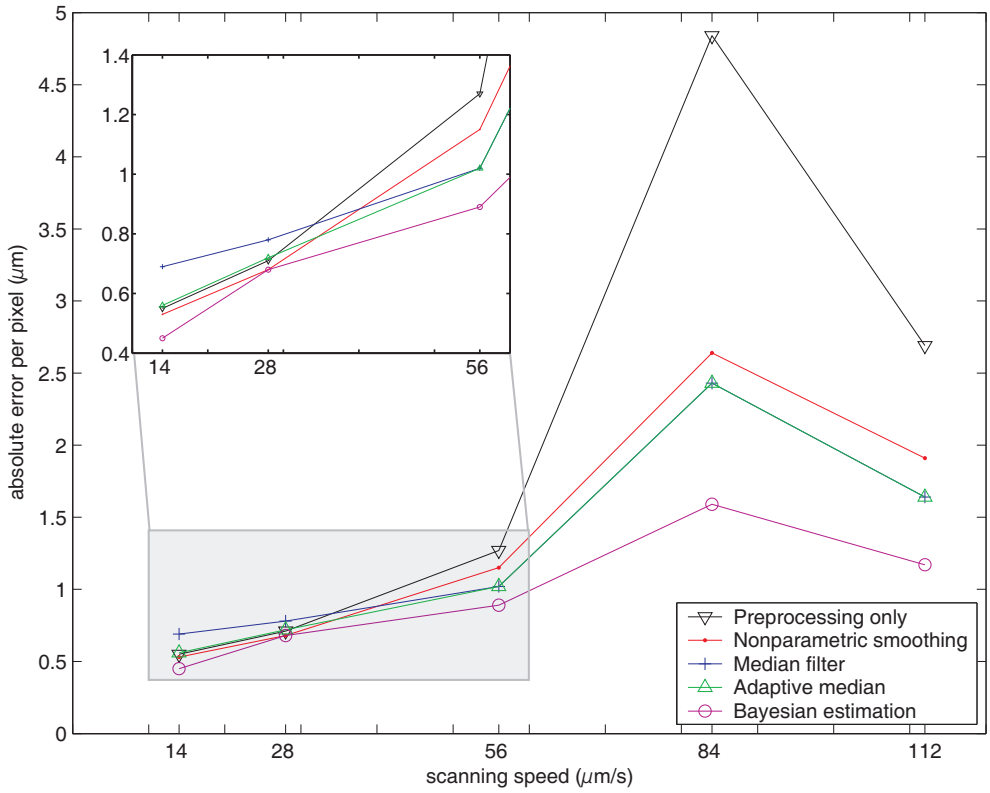


Figure 4.25: Graphical representation of the results from Table 4.3. The connecting lines are drawn only for better visual discrimination.

28 μm/s, the reconstructions still have qualitative differences, particularly for the microstructure of the surface. The Bayesian estimation, and to a smaller degree, also the nonparametric smoothing are more precise with the surface representation, while it lets more outliers slip through, other than the median filter approaches.

Measurements with 56 μm/s scanning speed The Bayesian estimation approach has the least average error in this setting. It is now followed by the results from the median filter and the adaptive median with practically equal absolute errors[1], while the nonparametric smoothing has a larger error, followed by at last the result obtained with the postprocessing only approach, which suffers from the significantly increased number of outliers, clearly perceptible in Fig. A.2.

A closer inspection reveals that the nonparametric smoothing leads to residual errors of about 2 frames for the bulk of pixels, similar to the 28 μm/s-measurements. The median filters however leave the postprocessed data with one or zero frames absolute error, and therefore obtain a smaller $\bar{\mathcal{E}}_{\mathrm{pp}}$ value.

The optimum setting for the parameter c of the adaptive median is very close to zero (cf. Table 4.2), so that nearly all pixels are considered outliers and thus the spatial median filter is rarely skipped. As a consequence, the errors $\bar{\mathcal{E}}_{\mathrm{pp}}$ of the median based approaches differ only marginally, and the reconstructions show fluctuations merely on the level of the discretization noise, with the exception of fewer outliers left over from the adaptive median filter.

Measurements with 84 μm/s scanning speed These measurements are significantly affected by the a resonance phenomenon between sampling rate and periodicity of the interferogram's inner oscillation, which leads to a very poor signal-to-noise ratio (cf. Sec. 4.3.2). This leads to a steep rise in the number of outliers after preprocessing (cf. Fig. A.3) and subsequently in the average error $\bar{\mathcal{E}}_{\mathrm{pp}}$. Without postprocessing, the acquired height maps are not usable.

Both median filter based approaches approximately halve the absolute error. The optimum setting of the adaptive median's parameter c is zero, therefore both algorithms yield equal results and no further gain is possible by restricting the spatial median filter to fewer pixels.

The nonparametric smoothing algorithm leads to a slightly larger average error than the median filter also for this scanning speed. A high density of outliers impairs the performance of the nonparametric smoothing: The weighted filter approach ignores outliers and therefore higher weights are put on the remaining pixels. In addition, the target variance can only be reached by reiterating the filtering operation with a larger mask size. All this leads to a smoother reconstruction, but at the same time reduces the reproducibility of the original height value.

The Bayesian estimation copes best with the large number of outliers, as the integration of pre- and postprocessing eliminates the need to assign a height

[1] The numbers in Table 4.3 are rounded to the second digit.

value to spoilt data: Data series that in the conventional approach would result in outliers show a flat or fuzzy likelihood that contributes little to the Bayesian inference.

Measurements with 112 μm/s scanning speed This is the fastest scanning speed used in this series. At this speed, the processed time series contain only few samples of the actual interference pattern, therefore misdetections are rather frequent once the signal-to-noise ratio is too low. The number of outliers in the preprocessed data (cf. Fig. A.4 is however smaller than with 84 μm/s, but larger than with 56 μm/s.

Also at this scanning speed, the Bayesian estimation performs best in terms of $\bar{\mathcal{E}}_{\mathrm{pp}}$, followed by the median filter approaches. The height maps obtained from nonparametric smoothing have a slightly larger absolute error. These measurements do not differ essentially from the 56 μm/s or 84 μm/s data.

4.3.6. Conclusions and hints for application

For the measurements of the turned steel piece (Fig. 4.2), we have seen that for all scanning speeds evaluated, out of all tested postprocessing approaches the Bayesian estimation yields the height maps of minimum error $\bar{\mathcal{E}}_{\mathrm{pp}}$. The difference to the other methods is more significant for lower signal-to-noise ratios and higher scanning speeds. For the 14 μm/s and 28 μm/s measurements, the improvement over the adaptive median filter is smaller.

All these results are strictly valid only for the test piece measured. For other metallic surfaces of similar roughness, our findings should be transferable, as single tests have shown. This may not be the case for surfaces of notably different structure and local characteristics, like rubber, plastics or ceramics.

The computational cost for the Bayesian estimation is approximately linearly correlated to the size of a scanned data sequence (number of frames). In contrast, the other postprocessing filters require a constant processing time.

In summary, for measurements at fast scanning speeds, particularly low data quality and if the surface microstructure should be represented more precisely, the Bayesian estimation is the best choice. If the measurements have a high quality and a fast and less sophisticated postprocessing focusing around the removal of outliers is desired, the adaptive median filter is a good alternative.

5. Comparison with Bayesian approaches in image processing

Overview In this chapter we discuss conceptual similarities or junctures of the Bayesian estimation approach developed in Chap. 3 to methods of image restauration and functional image processing.

Removal of noise, outliers and other unwanted artifacts is the common task of what sometimes is called "early" image processing, put here in contrast to "later" stages as image classification or image understanding. As common as this task is, the field of restauration methods is vast. Consequently, we can only pick out very few approaches. These either make use of Bayesian methods, or they extend the 2-D domain of images into the third dimension, and along this way come near to the Bayesian estimation of Chap. 3.

5.1. Relation to Gibbs field methods

Background Markov random fields are an important concept for the functional image restauration (cf. Sec. 3.2.1). Image properties that can be expressed in form of a Markov field local characteristics or, equivalently, a Gibbsian energy operator (Hamiltonian) allow for a fast sampling of the a posteriori probability distribution of a Bayesian inference problem, by means of the Gibbs sampler, stochastic cooling or related methods, like ICM (cf. Sec. 3.2.2).

The Bayesian approach from Chap. 4 stands close to the Gibbs-Markov approach for the inference problem, and a mathematical link can be formed regarding the "dimensionality" of the data: In the Bayesian inference for interferometry, a 2-D plane is to be embedded into a 3-D data set, with the third dimension formed by the raw data along a height scale, which are used to calculate the (quasi-) likelihood (cf. Sec. 4.2.2). Within image restauration, often a noise model is set up, which describes the likelihood of a certain gray value given a hypothesis on the true gray value. While this model usually is assumed for the whole image, we can augment each pixel with its gray value probability distribution according to this likelihood. That way the analogy to interferometric data can be completed: The incoming 2-D image is embedded as a (fairly uneven) plane into the 3-D cube, with the third dimension formed by the gray value scale. The task of image restauration is now to "play around" with the plane by the help of a sampling algorithm to prepare an output image that is both denoised and close to the original data.

In Chap. 4 we demonstrated how, thanks to the choice of suitable priors (Sec. 4.2.3), the a posteriori probability can be calculated directly. In this section, we now study if these or similar priors can describe a Markov random

field. If this would be the case, we could elegantly compare different Markov random field samplers with the performance of the direct Bayesian estimation.

For a correct description of a Markov random field which makes efficient sampling strategies possible, either the energy function has to stated explicitly in terms of cliques [Besag, 1986], [Winkler, 2003], or is has to be proven that the local characteristics are consistent, e. g. as shown in [Besag, 1974].

Rectangle prior The prior introduced in Eq. (4.35) is characterized in particular by its non-linearity: Its output, the results of AND operations, takes only two values, and so is independent from the exact number of neighbors, as well as many different configurations in the neighborhood are mapped into one value. These characteristics allow for the highly efficient calculation of the a posteriori probability. But they also inhibit this prior from a description compliant to a Markov random field.

Slightly simplifying the notation from Eq. (4.35), we write the prior like:

$$P(h^{\mathcal{C}}) = a + b \prod_{i \in \partial_{h^0}} W(h^i, h^0) \tag{5.1}$$

∂_{h^0} denotes the neighborhood of pixel h_0. With the requirement $a > 0$ for the offset a we ensure positivity for the expression, as required for a random field.

For the Gibbs field representation the energy function must be made up from summed contributions for each clique, cf. Eq. (3.41). The energy function U then goes into the Gibbs formula Eq. (3.39). With that equation, we can permissively ignore the potentiation with e for the moment, having a simple proportionality between U and Π, or can even allow a logarithm in that place. All these variation could be accommodated for, in particular with estimators locating the maximum of the a posteriori probability.

For our prior however, the crucial point is the product, which is determined by all neighbors of a pixel together only. This can be carried over to a Gibbsian formalism only if these neighbors together form a clique and we can define a corresponding neighborhood relationship. This requires of course that all other properties of cliques also be fulfilled (cf. Sec. 3.2.1). In particular the mutual neighborhood relationship requirement which implies a kind of "translation invariance" for the neighborhood relation would let us end up with a neighborhood relation of infinite extent. This *could* be a Markov random field, but with correlations spanning over the whole image, any option for efficient sampling is forfeit.

Along a less formal argumentation one could find this illustration: The mixture of additions and multiplications in the prior definition prevents us from setting up an ordinary Gibbs potential. Such a function should be breakable down into (additive or multiplicative) contributions from small cliques and conventional neighborhoods of few pixels, which is not possible for Eq. (5.1).

Variations of the rectangle prior If one neglects the parameter a in Eq. (5.1), the way to a simple Gibbsian energy functions is opened up: For estimators that

are independent from logarithmized probabilities, the following neighborhood potentials could be used:

$$V_{\{h^0\}} = 0 \tag{5.2a}$$
$$V_{\{h^0,h^i\}} = \log W(h^i, h^0) \tag{5.2b}$$
$$V_{\{\text{higher order cliques}\}} = 0 \tag{5.2c}$$

Of course, setting $a = 0$ has a significant impact on the Bayesian estimation approach and is problematic in general. Without the constant offset, the prior can become zero outside the rectangle and while it is still a probability distribution, it completely disables high edges in the image. Also, positivity is required by the Hammersley-Clifford theorem (cf. Sec. 3.2.1).

On the other side, using this modification and additionally choosing a particular setting for the window function:

$$W(h^i, h^0) = \delta(h^i, h^0), \tag{5.3}$$

we get to the potential of the Potts model (Eq. (3.56) in Sec. 3.2.1). Along this route also the more general Gaussian and auto-logistic models could be considered with an appropriate choice of the window function W, but all is based on the somewhat inapt setting $a = 0$.

Linearization of the rectangle prior The behavior of the rectangle prior can, at least to a certain degree, imitated by a linear replacement. "Linear" here suggests that the prior should not decide between only two outcomes, but should react proportionally to the number of neighbors that fall within the window limit. This would make a representation by additive neighborhood potentials possible. Such a prior cannot benefit from the efficient posterior estimation developed in Chap. 4 exactly because it has too many outcomes.

A possible realization could use a smoothed replacement for the rectangle window which never reaches zero:

$$W_\mu(\Delta_h) = k\left(\tanh\left(\mu(\Delta_h + \Lambda)\right) - \tanh(\mu(\Delta_h - \Lambda))\right) \tag{5.4}$$

The parameter μ tunes the steepness of the transitions, Λ is the width parameter of the window, $\Delta_h = h^i - h^0$. In the limit $\mu \to \infty$ it becomes the original rectangle window:

$$\lim_{\mu\to\infty} W_\mu(h^0 - h^i) = W(h^0, h^i) \tag{5.5}$$

As there always holds $W_\mu > 0$ the prior with W_μ yields a valid probability distribution.

With this modified window function, the settings laid out in the above paragraph (Eqs. (5.2) could be used in a mathematically sound way. Starting with Eq. (3.47) we the get the following local characteristics of the Markov random

field:

$$\Pi(\boldsymbol{h}|\partial_{h^0}) = \frac{1}{z}\exp\left(-U(\boldsymbol{h}) + U(\partial(h^0))\right) \tag{5.6}$$

$$\propto \exp\left(-\sum_{c\in C} V_c(\boldsymbol{h})\right) \tag{5.7}$$

$$= \exp\left(-\sum_{i \text{ with } h^0\sim h^i} \log W_\mu(h^i - h^0)\right) \tag{5.8}$$

with $\boldsymbol{h} = \{h^0, \partial(h^0)\}$ the set of height values for a center pixel and it's neighborhood. The modified window function offers therefore additional possibilities to approximate the original rectangle prior.

Conclusion The rectangle prior Eq. (5.1) cannot be represented by a Gibbsian energy function or a Markov random field. The reason is that the nonlinear behavior of the prior with only two output values cannot be represented with the means of sums or products of neighborhood interaction potentials. Two approaches, which diverge from the original prior but which can be represented by a Gibbs energy function are developed and discussed.

5.2. Relation to channel smoothing

Background Channel smoothing is a rather novel approach to image denoising by averaging a higher-dimensional representation of the original image [Scharr et al., 2003] . It comprises the following steps:

1. Preparation of the *channel representation* of the original data,
2. linear filtering in the domain of the channel representation,
3. reconstruction of the smoothed image from the channel representation.

Channel smoothing has some relations to robust estimation [Chu et al., 1998], diffusion approaches [Felsberg et al., 2002] and then indirectly to bilateral filtering [Barash, 2001], which makes it interesting to study.

Channel representation The channel representation is a general concept of mapping a signal into a space of higher dimension. It was invented as a tool for the projection of non-trivial invariances into a linear and easier accessible space [Nordberg et al., 1994]. This is done via an encoding function B, which gives a local, smooth function in the channel space and centered around zero. For computational efficiency, a limited support encoding can be used which is shifted towards the signal value. The encoding for functional value f is given by the corresponding channels c_n, which form a vector of length N with non-zero values only in the vicinity of f:

$$c_n(f) = B(f - n) \tag{5.9}$$

Accordingly, a signal $f(x)$ on an M-dimensional carrier ($M = 2$ for an image) is encoded into a N-dimensional vector functional or, equivalently, a function on $M \times \{1, \ldots, N\}$.

A central idea of this approach is the sparseness of representation. E. g., a signal of discrete 8-bit range could be encoded into only 5 to 10 channels. To avoid loss of information, these channels are real-valued. In the decoding step, the channels are interpolated, according to the encoding function used.

Channel smoothing Channel smoothing is achieved by convoluting a smoothing filter G with the channel representation of the image, that is, for each of the n channels $c_n(x)$:

$$c'_n(x) = G(x) * c_n(x) \quad \forall\, n \tag{5.10}$$

Each channel is therefore smoothed separate from the others. The resultant image x' is obtained after decoding the smoothed channel representation.

$$x' = \sum_n B_n(c'_n(x)) \tag{5.11}$$

In the channel representation, we have a stack of matrices rather than a single image, one for each channel. To obtain smoothing in the channel representation, common linear filter masks are applied separately to each matrix. The gray value of single pixel is found encoded as a column perpendicular through the stack.

In [Scharr et al., 2003] the use of a B_2-spline encoding is proposed, which has the advantages of a linear decoding, fast computation due to its limited support and others more. The decoding step involves only a linear interpolation of few terms, unless smoothing leads to the representation of several gray values within one channel ("metamery"). In these cases, the reconstruction is centered around the largest channel value, which, from a statistical viewpoint, can be considered as the most reliable value.

Channel smoothing is able to preserve edges in the image, i. e., to act like a robust estimator by removing little differences of data with small variance and outliers of large variance, but retaining larger deviations that are supported by a small variance. Mathematically, this is supported by the influence function one can derive (cf. Sec. 3.3.3), which in case of Scharr's B_2-spline encoding corresponds to that of a robust error norm: It is linear for small deviations, reaches a maximum and falls smoothly to zero for larger differences [Felsberg et al., 2002].

Channel smoothing and Bayesian estimation The channel representation of an image uses a data structure that shows some similarities to that of Bayesian inference methods, in particular with the interpretation of Sec. 5.1. Technically, the channel encoding approach can then be compared to the noise modeling with a likelihood function. Together with the discussion in Sec. 5.1, this correspondence technically extends towards interferometric 3-D data. However, the processing strategies and their underlying theoretical foundation differ much:

Central to channel smoothing is that each channel is smoothed separate from the others, cf. Eq. (5.10). Within the nonlinear decoding step, the environment can be taken into account. On the other side, the a posteriori of the Bayesian estimation is calculated by directly consolidating information spread across the neighborhood and the height scale. Only in the particular case of the the δ-prior (cf. Sec. 4.2.3) with no offset the estimation is limited to information within one height level. We then get for Eq. (4.28):

$$P(h^{\mathcal{C}}) = b \prod_{i=1}^{k} \delta(h^i - h^0) \tag{5.12}$$

As discussed in Sec. 5.1, removing the offset ($a = 0$) is disfavorable. The a posteriori probability then becomes:

$$\mathbb{P}(h^0|\boldsymbol{x}^{\mathcal{C}}) \propto \prod_{i=1}^{k} f(\boldsymbol{x}^i|h^i)|_{h^i=h^0} \tag{5.13}$$

Using estimators of maximum modes of the a posteriori probability, the above equation can be logarithmized:

$$\mathbb{P}'(h^0|\boldsymbol{x}^{\mathcal{C}}) = \sum_{i=1}^{k} \log f(\boldsymbol{x}^i|h^i)|_{h^i=h^0} \tag{5.14}$$

This equation is the same that can be obtained by applying a channel smoothing approach with a special smoothing filter mask. This mask is zero for the central pixel, i. e. for a 3×3 neighborhood:

$$G(x) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{5.15}$$

The final decoding step that reconstructs the smoothed image from the channel representation is then replaced by a simple maximum mode estimator.

Conclusion Akin to the case of Gibbs field methods in Sec. 5.1, a bridge between channel smoothing and the Bayesian estimation for interferometric data can only be built with grave concessions. The unique characteristics of the interferometry approach to work within a 3-D domain must be cut down by a δ-type prior. This, on the other hand, does not do justice to the peculiarities of channel smoothing, namely the freedom to choose an encoder that balances smoothing with robustness. The 3-D dataset from interferometry has a fixed physical foundation that should be represented with the likelihood function and thus it leaves little room for variations that are physically still justifiable.

5.3. Relation to robust estimation

Background Robust estimation—in its simplest form—seeks to minimize the overall image gradient. Its "robustness" stems from the use of a robust error measure to weight the magnitude of the image gradient.

Generally, using a robust error measure reduces the influence of outliers on any optimization procedure. Edges in the image often feature large gradient and can as well be regarded as outliers [Black and Rangarajan, 1996], [Black and Sapiro, 1999]. If only small error costs are imposed, the blurring of edges is avoided. Such a degradation model favors images with smooth regions, separated by a number of steps, similar to the situation we experience with height maps from white light interferometry. The exact outcome of the minimization procedure depends on the approximation to the spatial gradient in the image [Jähne, 2002] and the form of the error measure used, respectively its associated influence function (cf. Sec. 3.3.3).

The following estimation is therefore a valid criterion for edge-preserving denoising [Black and Sapiro, 1999]:

$$\hat{\boldsymbol{x}} = \arg\min \int_{\Omega} \rho(||\nabla \boldsymbol{x}||) d\omega \tag{5.16}$$

Here Ω denotes the image domain, $\boldsymbol{x}$ the intensity image and $\nabla \boldsymbol{x}$ the gradient image.

Link to Bayesian estimation A discretization of Eq. (5.16) for the domain and gradient could look like:

$$\hat{\boldsymbol{x}} = \arg\min_{x} \sum_{s \in \boldsymbol{x}} \sum_{p \in \partial_s} \rho_r(x_p - x_s) \tag{5.17}$$

with ρ_r a robust error measure. While we take the smoothness assumption for height maps into account with the prior, it is here embedded in the minimum-gradient requirement and the form of the error norm.

If we restrict the domain of the estimation in Eq. (5.17) to just the neighborhood of each pixel s, we get:

$$\hat{x}_s = \arg\min_{x} \sum_{p \in \partial_s} \rho_r(x_p - x_s) \tag{5.18}$$

With help of a simple Bayesian cost function (cf. Sec. 3.1.4, this minimum-error problem is turn into a local maximum a posteriori estimation (MPM):

$$\hat{x}_s = \arg\max_{x_s \in X_s} P(x_s | y^{\mathcal{C}}) \tag{5.19}$$

with $\mathcal{C} = \partial(x_s)$. Its a posteriori probability is then:

$$\mathbb{P}(x_s | y^{\mathcal{C}}) = \frac{1}{Z} \exp\left\{ - \sum_{p \in \partial(x_s)} \rho_r(x_p - x_s) \right\} \tag{5.20}$$

$$= \frac{1}{Z} \prod_{p \in \partial(x_s)} \exp\{-\rho_r(x_p - x_s)\} \tag{5.21}$$

As with channel smoothing in Sec. 5.2, we look at the special case of the δ-prior with no offset (cf. Sec. 4.2.3), which limits the estimation to within the

current height level. It gives the a posteriori probability in Eq. (5.13), which we use to adapt the error measure for "our" robust estimation:

$$\rho_r(x_p - x_s) = -\log f(y_p|x_p = x_s) \tag{5.22}$$

which is the negative of the log-likelihood.

We have discussed modeling the likelihood for white light interferometry in Sec. 4.2.2. If we integrate the phase shift of the inner oscillation, the ideal likelihood should look similar to the following, with G_σ the (Gaussian) envelope and a notation closer to image processing:

$$\begin{aligned} f(y|x) &= a + b * G_\sigma(y - x) && (5.23)\\ &= a + b * \exp\left\{-\frac{(y-x)^2}{\sigma^2}\right\} && (5.24) \end{aligned}$$

Then we obtain for the error measure:

$$\rho(z) = -\log\left(a + b\exp\left\{-\frac{z^2}{\sigma^2}\right\}\right), \tag{5.25}$$

where we set $z = y - x$ for convenience.

Discussion of the error measure Eq. (5.25) is the error measure that corresponds to Bayesian surface estimation in a special restricted case, due to the choice of the prior. Let us now see what the characteristics of this error measure are, in the scope of robust estimation.

It is important not to ignore the summand a, which actually adds robustness. With $a > 0$, Eq. (5.23) gives a residual probability even for large differences, far away from $y \approx x$, where the exponential term alone drops quickly to zero.

Leaving a out, we would obtain the ordinary, non-robust quadratic error norm. This aspect can be seen with the series expansion of Eq. (5.25):

$$\rho(z) = -\log(a+b) + \frac{b}{a+b}\frac{z^2}{\sigma^2} + \left\{\left(\frac{b}{a+b}\right)^2 - \frac{b}{a+b}\right\}\frac{z^4}{2\sigma^4} + \mathcal{O}(z^6) \tag{5.26}$$

As the expression in curly braces is always < 0, the error function is dampened with $\mathcal{O}(z^4)$ for large z. The offset $-\log(a+b)$ can safely be ignored now. The robustness can be assessed by evaluating the influence function, the derivative of $\rho(z)$:

$$\psi(z) = \frac{b}{a+b}\frac{2z}{\sigma^2} + \left\{\left(\frac{b}{a+b}\right)^2 - \frac{b}{a+b}\right\}\frac{2z^3}{\sigma^4} + \mathcal{O}(z^5) \tag{5.27}$$

The error function above is similar to the robust estimator proposed by Leclerc (cf. [Leclerc, 1989]). It also shows a redescending influence function—by comparison of its series expansion to Eq. (5.26), the similarity to our error function becomes obvious:

$$\begin{aligned} \rho(z) &= 1 - \exp\{-z^2/\sigma^2\} = \frac{x^2}{\sigma^2} - \frac{x^4}{\sigma^4} + \frac{x^6}{\sigma^6} + \mathcal{O}(x^8) && (5.28)\\ \psi(z) &= \frac{2x}{\sigma^2}\exp\{-z^2/\sigma^2\} && (5.29) \end{aligned}$$

Conclusion From the calculations one can see that Bayesian surface estimation is very distantly related to robust estimation for 2-D images with an edge-preserving error norm. In particular we had to accept major qualitative changes to the Bayesian prior in order to adapt it to the framework of robust estimation.

Hints towards anisotropic diffusion The concept of anisotropic diffusion [Perona and Malik, 1990] has widely been established as an image denoising tool that is able to preserve edges. Based on the physical concept of diffusion [Jähne, 2002], in an iterative algorithm of discrete time the gray values are modified according to the diffusion equation:

$$\frac{\partial}{\partial t}\,\boldsymbol{x} = \operatorname{div}\left[\mathcal{D}\nabla\boldsymbol{x}\right] \tag{5.30}$$

The diffusion tensor $\mathcal{D}$ is chosen as a function of the current image. If one chooses it inverse to the structure tensor of the image, one can suppress smoothing across image features like edges, and enforce smoothing in regions where untextured noise dominates [Weickert, 1998].

It has been shown that robust estimation is closely related to anisotropic diffusion [Black et al., 1998]. Among others, the authors prove that the classical weighting term introduced for diffusion [Perona and Malik, 1990],

$$g(z) = \exp\left\{-\frac{z^2}{k^2}\right\} \tag{5.31}$$

is in fact equal to robust estimation using Leclerc's error norm:

By calculus of variations, the integral optimality criterion Eq. (5.16) can be transformed into the general diffusion equation

$$\frac{\partial I(x,y,t)}{\partial t} = \operatorname{div}\left[g(||\nabla I||)\nabla I\right] \tag{5.32}$$

The link is here given by defining the gradient's weighting function as

$$g(z) = \frac{\rho'(z)}{z} \tag{5.33}$$

We have seen that Bayesian surface estimation can be considered as a particular kind of robust estimation for a certain setting of the prior. Subsequently, with these findings the loose link can be stretched towards anisotropic diffusion.

6. Summary

Overview In this thesis, we have discussed the processing and reconstruction of height maps that are obtained from scanning white light interferometry. Primarily, a novel estimation approach based on Bayesian inference has been developed, which is more accurate than the procedures we have compared it to.

White light interferometry has become an important measurement tool for the inspection of surfaces. In particular, the focus of applications lies with rough, but precisely worked surfaces, as those continuously gain importance in manufacturing industry, but at the same time can almost not be measured with more established approaches, here to mention laser interferometry and tactile mechanical test devices. White light interferometry fills an important gap, and we have presented the optical foundations and the physics of rough surface reflection.

It is almost certain that with measurement or testing tasks for white light interferometry, there coincides the requirement that the measured data be as precise as possible and free of errors. Interferometry with rough surfaces is however naturally error-prone, in that measurement artifacts arise. This makes the postprocessing and denoising approaches discussed here an integral part of a white light interferometry measurement setup.

Preprocessing for white light interferometry The height map is not a direct outcome of the interferometric measurement, but a result from the preprocessing step. The aim here is to calculate a preliminary height map from the raw data that is fast, economic and as precise as necessary. While robust postprocessing is able to remove gross errors and outliers if they do not cluster in large groups, it cannot recognize slight shifts in height values. Therefore we have only the intention to provide a precise height estimate for pixels with sufficient signal-to-noise ratio.

We have presented known preprocessing algorithms and the practical evaluation of a novel approach based on a wavelet analysis. As for our test objects it does not perform better, and sometimes even less precise than the established algorithms, so far we cannot recommend it, based on these measurements.

Postprocessing of height maps The height map obtained from preprocessing usually contains a number of outliers and other measurement artifacts, which ought to be removed during postprocessing. At the same time however, the height map could contain similar structure which are image features and must persist, as well as the surface topology. This task is usually subsumed as "smoothing" or "denoising" of the surface.

We have touched the theory of linear and robust filtering and have given an account of postprocessing algorithms, both derived from conventional (2-D) image processing, and specialized for white light interferometry by taking the confidence measure into account. In addition, we present a novel modification to the median filter which improves its accuracy by adapting it to the local surface quality.

Bayesian inference in white light interferometry For the conventional image processing, Bayesian methods have turned out to be powerful tools for the reconstruction of deteriorated images. But only with the development of the Gibbs sampler and related methods the "inverse" problem of image restauration has been made solvable with a reasonable effort of computation time.

We have discussed the theoretical background of Bayesian inference with a focus on image processing and again on height map estimation. In image processing, particular attention has been turned to robust statistics and priors that can retain edges.

For height map estimation, we have presented a novel approach that unifies pre- and postprocessing in a single estimation procedure. It abolishes the conventional pipeline structure that reduces the raw data to a 2-D image before postprocessing in favor of locating the optimum embedded surface in the full 3-D data set. This method is an adaptation from the Bayesian estimation for images. By restricting the choices for prior probability distributions to a certain class with few outcomes, we have been able to to directly compute the a posteriori probability of the inference problem.

With the Bayesian surface estimation, it is therefore not necessary to resort to stochastic sampling or simulation methods to cope with the vast size of the a posteriori configuration space. The optimum configuration with respect to the marginal posterior mode estimator can be calculated analytically.

This outcome is interesting also because with such a prior, we are able to obtain height reconstructions that have a smaller average absolute error than what conventional postprocessing can achieve.

Comparison of postprocessing algorithms In a series of multifold recordings with different scanning speeds, height measurement raw data have been taken of a turned metallic piece, which we hope is a sufficiently representative sample for at least part of the white light interferometer applications in manufacturing industry. For the comparison of different postprocessing algorithms, a procedure for setting up a reference height map has been devised.

The postprocessing algorithms have been compared according to the average absolute error against the reference height map and also classified by the MAD-value as a measure of the difficulty to obtain a height estimate. The differences in overall accuracy between the tested algorithms is larger in case of low signal-to-noise ratio and, related to this, a fast scanning speed. For higher data quality the differences are less pronounced. With the detailed analysis it has become clear that the Bayesian estimation is particularly precise for the bulk of pixels that represent the flat surface where it reproduces the surface microstructure

best possible. The median filter based algorithms are less precise with the microstructure, but are generally more reliable with the removal of outliers.

In summary, we would recommend the adaptive median filter for measurement tasks with a high raw data quality and where the postprocessing should be less elaborate. If a more accurate reproduction of the surface topology is desired, or if the raw data are of low quality, the Bayesian surface estimation appears a better choice.

Outlook The developments started in this thesis leave a number of old and new questions open and offer some possibilities to continue and expand the work started.

The Bayesian surface estimation developed has yet only been tested with a handful of surfaces, a detailed comparison with other postprocessing approaches exists for only one sample piece. More practical experience would be desirable. The parameter setting still has to balance between the best possible reconstruction of the surface topology. With the current optimization measure, the average absolute error, this balance cannot be directly adapted to different applications. It could also be evaluated if and when other priors that provide more more levels for the result could perform better.

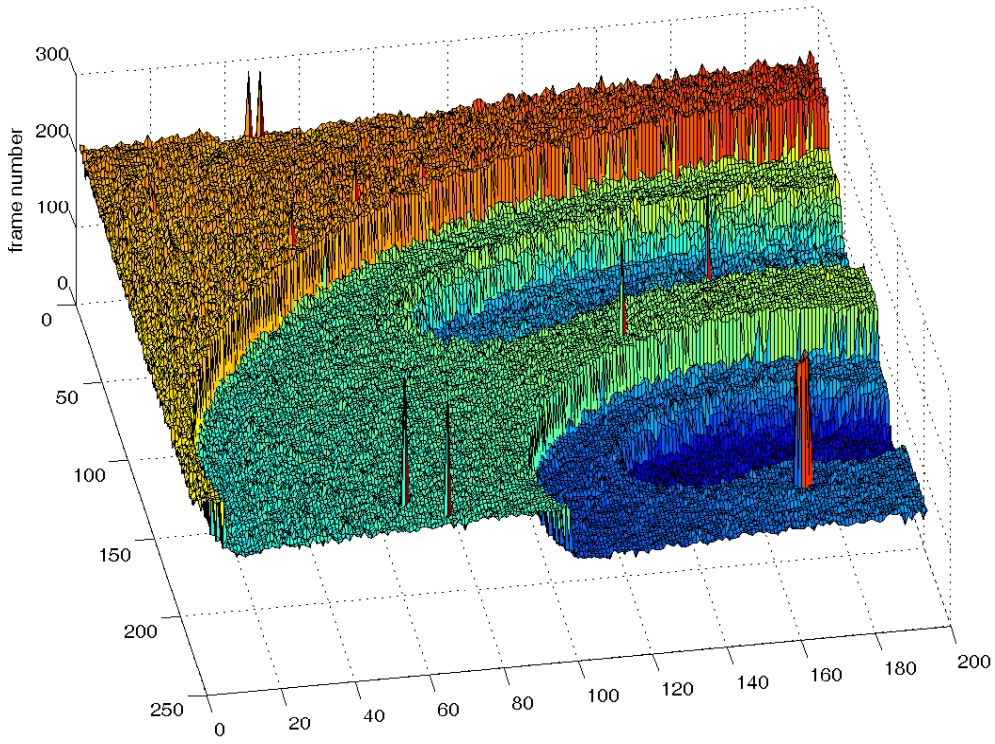
For high scanning speeds, the Bayesian surface estimation is particularly well suited. Currently however, the data processing is limited to the height levels once fixed with the interferometer setup (the scanning speed together with the frame rate of the imaging system). As a consequence, the obtained height maps share the same, coarse step size, which is probably not always desired. This issue could be solved with interpolation, either for the raw data, or in the course of the height map estimation.

With the ambition to find new connections, we have briefly looked into emerging methods of image restauration, i. e., Gibbs sampling and channel smoothing. So far, any two sides have been proven unique by their own kind, and a transfer is only possible by putting up significant constraints. Still, if we could find bridges, not only it could be possible to adapt these powerful image processing methods for height map estimation, but also to bring the Bayesian surface estimation closer to 2-D image processing. Some labeling problems, such as the depth estimation from stereo disparity data, are very similar to white light interferometry height estimation. Unfortunately, up to now we have not been able to obtain proper results for the depth reconstruction from disparity maps.

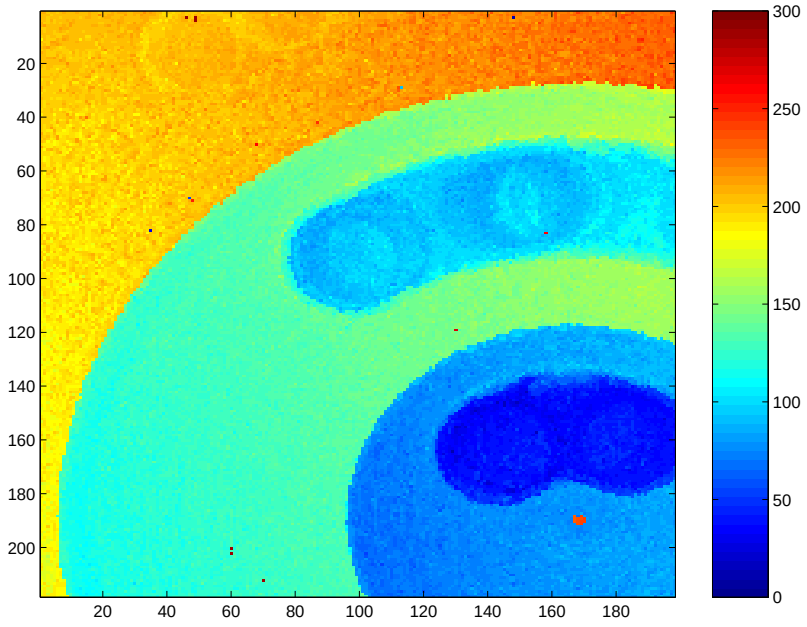
Nevertheless, further work in all these directions has a great potential of mutual benefit.

A. Additional height map reconstructions

In this appendix, we gather figures of height map reconstructions for scanning speeds other than 28 μm/s. These figures are referred to in Chap. 4, in particular in Sec. 4.3.5.

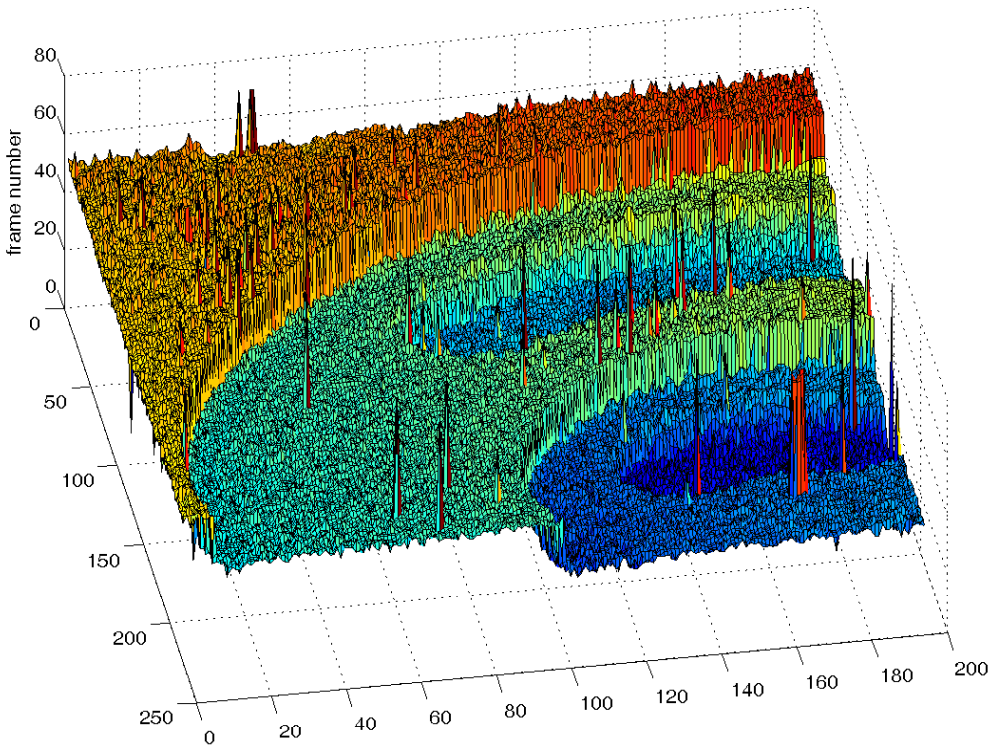


(a)

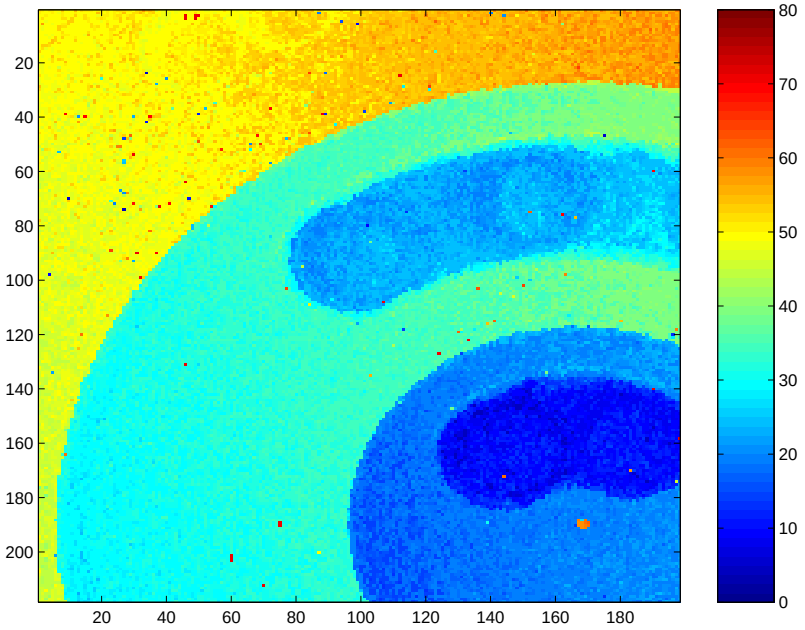


(b)

Figure A.1: Reconstruction with preprocessing only. The scanning speed is 14 μm/s, the scale of the height axis 0.28 μm per frame.

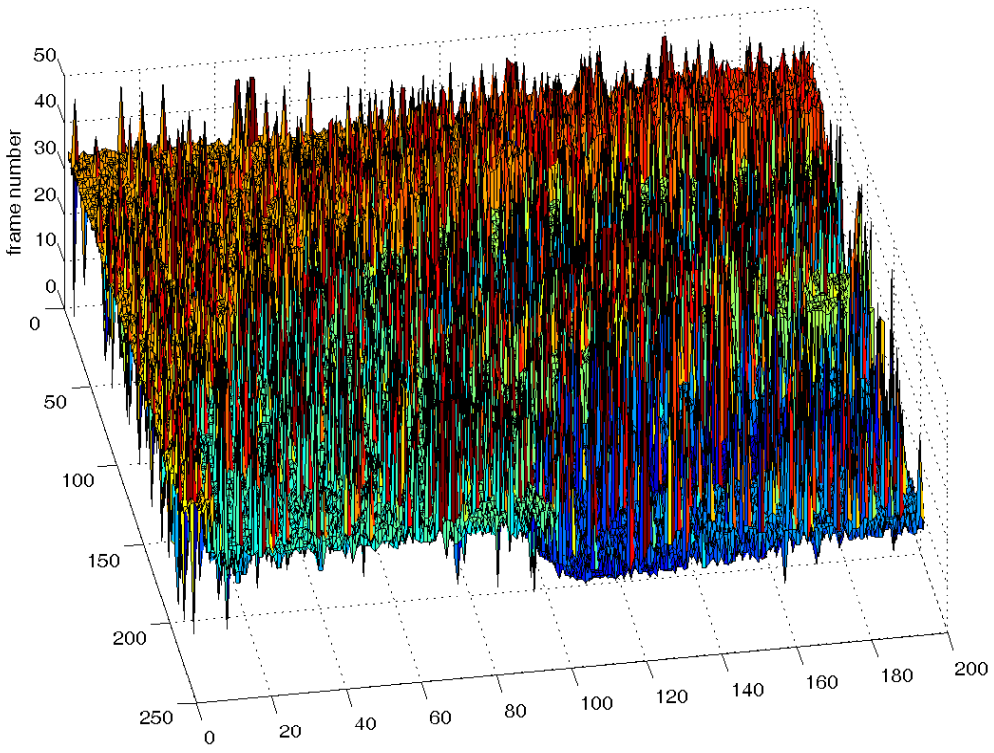


(a)

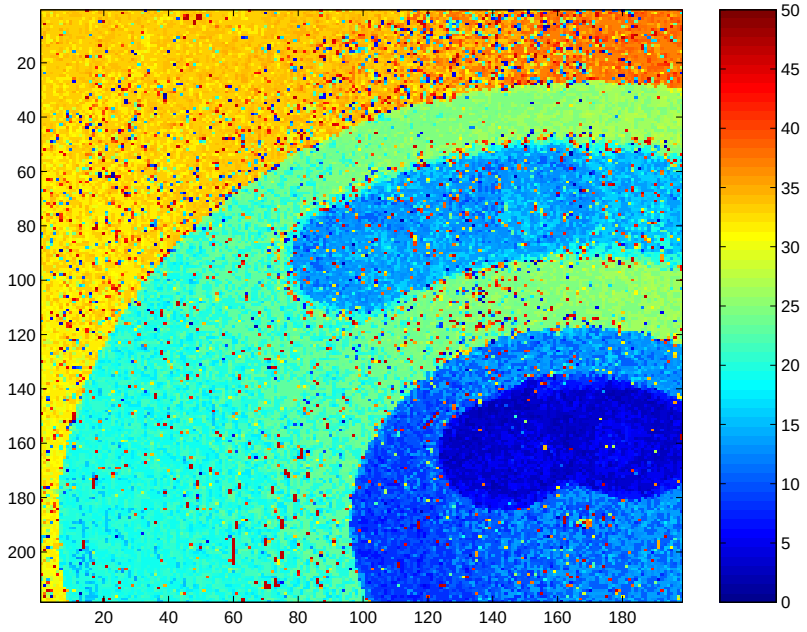


(b)

Figure A.2: Reconstruction with preprocessing only. The scanning speed is 56 μm/s, the scale of the height axis 1.12 μm per frame.

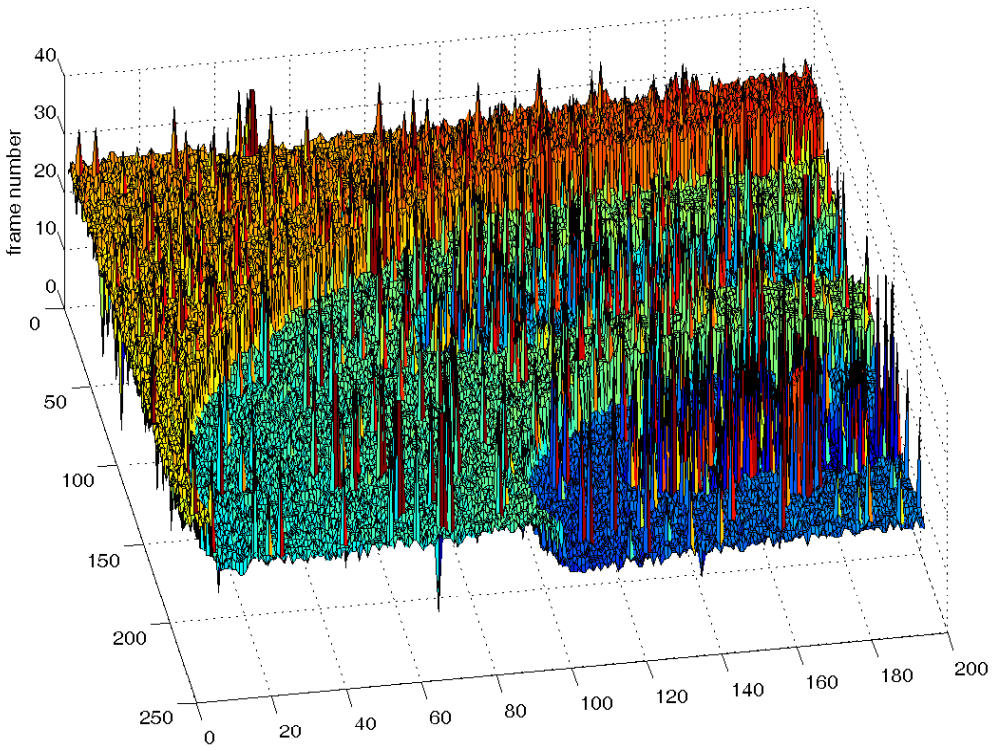


(a)

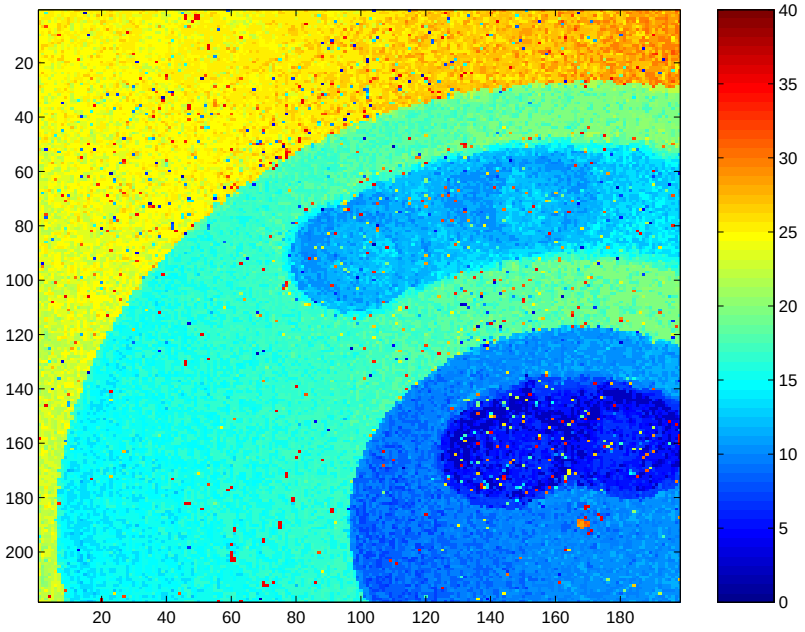


(b)

Figure A.3: Reconstruction with preprocessing only. The scanning speed is 84 μm/s, the scale of the height axis 1.68 μm per frame.



(a)



(b)

Figure A.4: Reconstruction with preprocessing only. The scanning speed is 112 μm/s, the scale of the height axis 2.24 μm per frame.

List of Figures

List of Tables

Bibliography

[Ammon et al., 1997] Ammon, G., Andretzky, P., Blossey, S., Bohn, G., Ettl, P., Habermeier, H. P., Harand, B., and Häusler, G. (1997). 'coherence radar' – new modifications of white-light interferometry for large object shape acquisition. In *Proceedings of the EOS Topical Meeting on Optoeletronic Distance Measurements and Applications.*

[Averintsev, 1972] Averintsev, M. B. (1972). Description of Markov random fields using Gibbs conditional probabilities. *Theory Probab. Appl.*, 17(1):21–35. In Russian.

[Aziz, 1998] Aziz, D. J. (1998). Interferometric measurement of surface roughness in engine cylinder walls. *Opt. Eng.*, 37(5):1429–1434.

[Barash, 2001] Barash, D. (2001). Bilateral filtering and anisotropic diffusion: Towards a unified viewpoint. In *Proceedings, Scale-Space 2001*, Lecture Notes in Computer Science vol. 2106, pages 273–280, Berlin, Heidelberg, New York.

[Bayes, 1763] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.*, 53:370–418.

[Beckmann and Smith, 2003] Beckmann, C. F. and Smith, S. M. (2003). Probabilistic independent component analysis for functional magnetic resonance imaging. Technical Report TR02CB1, Oxford Center for Functional Magnetic Resonance Imaging of the Brain (FMRIB).

[Bergmann and Schaefer, 2004] Bergmann, L. and Schaefer, C. (2004). *Lehrbuch der Experimentalphysik: Bd. 3 Optik.* de Gruyter, Berlin, 10th edition.

[Besag, 1974] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Statist. Soc. B*, 36(2):192–236.

[Besag, 1986] Besag, J. (1986). On the statistical analysis of dirty pictures. *J. R. Statist. Soc. B*, 48(3):259–302.

[Besag et al., 1995] Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statist. Sci.*, 10(1):3–66.

[Bickel et al., 1983] Bickel, P. J., Doksum, K., and Hodges Jr., J. L., editors (1983). *A Festschrift for Erich L. Lehmann.* Wadsworth Intl., Belmont, Cal.

[Black and Rangarajan, 1996] Black, M. J. and Rangarajan, A. (1996). On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Int. J. Comput. Vis.*, 19(1):57–91.

[Black and Sapiro, 1999] Black, M. J. and Sapiro, G. (1999). Edges as outliers: Anisotropic smoothing using local image statistics. In *Scale-Space Theories in Computer Vision, Second International Conference, Space-Space '99*, Lecture Notes in Computer Science vol. 1682, pages 259–270, Berlin, Heidelberg, New York.

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

[Blake and Zisserman, 1987] Blake, A. and Zisserman, A. (1987). *Visual Reconstruction*. MIT Press, London.

[Bohn, 2000] Bohn, G. (2000). *Hardware-implementierte Algorithmen zur Optimierung des Meßprinzips Kohärenzradar*. PhD thesis, Univ. Erlangen-Nürnberg.

[Born and Wolf, 1999] Born, M. and Wolf, E. (1999). *Principles of Optics*. University Press, Cambridge, 7th edition.

[Bouman and Sauer, 1993] Bouman, C. and Sauer, K. (1993). A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. Image Processing*, 2(3):296–310.

[Caber, 1993] Caber, P. J. (1993). Interferometric profiler for rough surfaces. *Appl. Opt.*, 32(19):3438–3441.

[Chalmond, 2003] Chalmond, B. (2003). *Modeling and Inverse Problems in Image Analysis*, volume 155 of *Applied Mathematical Sciences*. Springer, Berlin, Heidelberg, New York.

[Chib and Greenberg, 1995] Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Amer. Statist.*, 49(4):327–335.

[Chu et al., 1998] Chu, C. K., Glad, I. K., Godtliebsen, F., and Marron, J. S. (1998). Edge-preserving smoothers for image processing. *J. Amer. Statist. Assoc.*, 93(442):526–541.

[Dainty, 1984] Dainty, J. C., editor (1984). *Laser Speckle and Related Phenomena*. Springer, Berlin, Heidelberg, New York, 2nd edition.

[Dändliker et al., 1995] Dändliker, R., R., H., Poltich, J., and Zimmermann, E. (1995). High accuracy distance measurement with multiple-wavelength interferometry. *Opt. Eng.*, 34:2407–2412.

[Daubechies, 1992] Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NFS Regional Conference Series in Applied Mathematics 61. SIAM.

[Davidson et al., 1987] Davidson, M., Kaufman, K., Mazor, I., and Cohen, F. (1987). An application of interference microscopy to integrated circuit inspection and metrology. In *SPIE Proceedings Vol. 775: Integrated Circuit, Metrology, Inspection and Process Control*, pages 233–247.

[Davies and Gather, 1993] Davies, L. and Gather, U. (1993). The identification of multiple outliers. *J. Amer. Statist. Assoc.*, 88(423):782–801.

[de Groot and Colonna de Lega, 2003] de Groot, P. and Colonna de Lega, X. (2003). Valve cone measurement using white light interference microscopy in a spherical measurement geometry. *Opt. Eng.*, 42(5):1232–1237.

[de Groot et al., 2002] de Groot, P., Colonna de Lega, X., Kramer, J., and Turzhitsky, M. (2002). Determination of fringe order in white-light interference microscopy. *Appl. Opt.*, 41(22):4571–4578.

[de Groot and Deck, 1995] de Groot, P. and Deck, L. (1995). Surface profiling by analysis of white-light interferograms in the spatial frequency domain. *J .Mod. Opt.*, 42(2):389–401.

[Deck and de Groot, 1994] Deck, L. and de Groot, P. (1994). High-speed noncontact profiler based on scanning white-light interferometry. *Appl. Opt.*, 33(31):7334–7338.

[DIN, 1990] DIN (1990). *DIN Standards No. 4760–4776*. DIN, Berlin.

[Dobruschin, 1968] Dobruschin, P. L. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Probab. Appl.*, 13(2):197–224.

[Donoho and Huber, 1983] Donoho, D. L. and Huber, P. J. (1983). *The Notion of Breakdown Point*, pages 157–184. In [Bickel et al., 1983].

[Dresel et al., 1992] Dresel, T., Häusler, G., and Venzke, H. (1992). Three-dimensional sensing of rough surfaces by coherence radar. *Appl. Opt.*, 31(7):919–925.

[Eberle, 2005] Eberle, M. (2005). *Untersuchungen zur hochgenauen Fertigungsmesstechnik mit dem Kohärenzradar*. PhD thesis, Univ. Erlangen-Nürnberg.

[Ettl, 2001] Ettl, P. (2001). *Über die Signalentstehung bei Weißlichtinterferometrie*. PhD thesis, Univ. Erlangen-Nürnberg.

[Ettl, 2002] Ettl, P. (2002). Definition of a confidence measure. Personal communication.

[Everitt and Bullmore, 1999] Everitt, B. S. and Bullmore, E. T. (1999). Mixture model mapping of brain activation in functional magnetic resonance images. *Hum. Brain Mapping*, 7:1–14.

[Felsberg et al., 2002] Felsberg, M., Scharr, H., and Forssén, P.-E. (2002). The B-spline channel representation: Channel algebra and channel based diffusion filtering. Technical report, Computer Vision Lab., Linköping Univ.

[Fercher et al., 1985] Fercher, A. F., Hu, H. Z., and Vry, U. (1985). Rough surface interferometry with a two-wavelength heterodyne speckle interferometer. *Appl. Opt.*, 24(14):2181–2188.

[Flach, 2002] Flach, B. (2002). *Strukturelle Bilderkennung*. Habil. thesis, Univ. Dresden.

[Fleischer et al., 2001] Fleischer, M., Windecker, R., and J., T. H. (2001). Theoretical limits of scanning white-light interferometry signal evaluation algorithms. *Appl. Opt.*, 40(17):2815–2820.

[Fleischer et al., 2000] Fleischer, M., Windecker, R., and Tiziani, H. J. (2000). Fast algorithms for the data reduction in modern optical 3-D profile measurement systems using MMX technology. *Appl. Opt.*, 39(8):1290–1297.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6(6):721–741.

[Goodman, 1984] Goodman, D. W. (1984). *Statistical Properties of Laser Speckle Patterns*. In [Dainty, 1984], 2nd edition.

[Granlund and Knutsson, 1995] Granlund, G. H. and Knutsson, H. (1995). *Signal Processing for Computer Vision*. Kluwer Academic Publishers, Dordrecht, Boston, London.

[Grimmett and Welch, 1990] Grimmett, G. R. and Welch, D. J. A., editors (1990). *Disorder in physical systems. A volume in honour of John M. Hammersley on the occasion of his 70th birthday*. Oxford Science Publication. Clarendon Press, Oxford.

[Grossmann and Morlet, 1984] Grossmann, A. and Morlet, J. (1984). Decomposition of Hardy functions into square integrable wavelets of constant shape. *J. Math. Anal. (SIAM)*, 15(4):723–736.

[Hammersley and Clifford, 1971] Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. Unpublished.

[Hampel, 1985] Hampel, F. R. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics*, 27(2):95–107.

[Hampel et al., 1986] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

[Harasaki et al., 2000] Harasaki, A., Schmit, J., and Wyant, J. C. (2000). Improved vertical-scanning interferometry. *Appl. Opt.*, 39(13):2107–2115.

[Harasaki and Wyant, 2000] Harasaki, A. and Wyant, J. C. (2000). Fringe modulation skewing effect in white-light vertical scanning interferometry. *Appl. Opt.*, 39(13):2101–2106.

[Hariharan, 2003] Hariharan, P. (2003). *Optical Interferometry*. Academic Press, New York, London, 2nd edition.

[Hartvig and Jensen, 2000] Hartvig, N. V. and Jensen, J. L. (2000). Spatial mixture modeling of fMRI data. *Hum. Brain Mapping*, 11:233–248.

[Hastings, 1970] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

[Hecht, 2001] Hecht, E. (2001). *Optics*. Addison–Wesley, Reading, Mass., 4th edition.

[Hissmann and Hamprecht, 2003] Hissmann, M. and Hamprecht, F. A. (2003). Bayes'sche Schätzung von Höhenkarten aus der Weißlicht-Interferometrie. In [Rose, 2003].

[Hissmann and Hamprecht, 2005] Hissmann, M. and Hamprecht, F. A. (2005). Bayesian surface estimation for white light interferometry. *Opt. Eng.*, 44(1):015601.

[Hjort and Mohn, 1984] Hjort, N. L. and Mohn, E. (1984). A comparison of some contextual methods in remote sensing classification. In *Proceedings of the 18th Int. Symp. on Remote Sensing of Enviroment*, volume 3, pages 1693–1702, Paris.

[Huber, 1981] Huber, P. J. (1981). *Robust Statistics*. Wiley Series in Probability & Mathematical Statistics. John Wiley & Sons.

[Hunt, 1977] Hunt, B. R. (1977). Bayesian methods in nonlinear digital image restoration. *IEEE Trans. Comput.*, 26(3):219–229.

[Jähne, 2002] Jähne, B. (2002). *Digital Image Processing*. Springer, Berlin, Heidelberg, New York, 5th edition.

[Katsaggelos, 1989] Katsaggelos, A. K. (1989). Iterative image restoration algorithms. *Opt. Eng.*, 28(7):735–748.

[Kendall et al., 1987] Kendall, M., Stuart, A., and Ord, J. K. (1987). *The Advanced Theory of Statistics*. Oxford Univ. Press, 4th edition.

[Kino and Chim, 1990] Kino, G. S. and Chim, S. S. C. (1990). Mirau correlation microscope. *Appl. Opt.*, 29(26):3775–3783.

[Kittler and Föglein, 1984] Kittler, J. and Föglein, J. (1984). Contextual classification of multispectral pixel data. *Image Vis. Comput.*, 2(1):13–29.

[Koch and Ulrich, 1991] Koch, A. and Ulrich, R. (1991). Fiber-optic displacement sensor with 0.02 μm resolution by white-light interferometry. *Sens. Actuators A. Phys.*, 25(1–3):201–207.

[Körber, 2004] Körber, A. (2004). Konzeption und Realisierung einer Shutter-Steuerung für die High-Speed Bildaufnahme bei einem Weißlichtinterferometer. Diploma thesis, FH Esslingen.

[Larkin, 1996] Larkin, K. G. (1996). Efficient nonlinear algorithm for envelope detection in white light interferometry. *J. Opt. Soc. Am. A*, 13(4):832–843.

[Lauritzen, 1996] Lauritzen, S. L. (1996). *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, Oxford.

[Leclerc, 1989] Leclerc, Y. G. (1989). Constructing simple stable descriptions for image partitioning. *Int. J. Comput. Vis.*, 3:73–102.

[Lee and Strand, 1990] Lee, B. S. and Strand, T. C. (1990). Profilometry with a coherence scanning microscope. *Appl. Opt.*, 29(26):3784–3788.

[Li, 2001a] Li, S. Z. (2001a). *Markov Random Field Modeling in Image Analysis*. Springer, Berlin, Heidelberg, New York.

[Li, 2001b] Li, S. Z. (2001b). *Modeling Image Analysis Problems Using Markov Random Fields*, pages 1–50. Volume 20 of [Rao, 2001].

[Mallat, 1999] Mallat, S. (1999). *A Wavelet Tour Of Signal Processing*. Academic Press, London, 2nd edition.

[McKechnie, 1976] McKechnie, T. S. (1976). Image-plane speckle in partially coherent illumination. *Opt. Quantum Electron.*, 8:61–67.

[Medeiros et al., 1998] Medeiros, F. N. S., Mascarenhas, N. D. A., and da Costa, L. F. (1998). Combined use of MAP estimation and k-means classifier for speckle noise filtering in SAR images. In *Proceedings of the IEEE Symp. on Image Analysis and Interpretation, Apr. 1998*, pages 250–255.

[Mehlhorn, 1984] Mehlhorn, K. (1984). *Data Structures And Algorithms 2: Graph Algorithms And NP-Completeness*. EATC Monographs on Theoretical Computer Science. Springer, Berlin, Heidelberg, New York.

[Meixner et al., 2003] Meixner, A., Zeh, T., Riemenschneider, M., Purde, A., and Koch, A. W. (2003). Formvermessung an bewegten technischen Oberflächen mittels der Speckle-Interferometrie. *Techn. Messen*, 70(2):93–98.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and E., T. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1091.

[Mohn et al., 1987] Mohn, E., Hjort, N. L., and Storvik, G. O. (1987). A simulation study of some contextual classification methods for remotely sensed data. *IEEE Trans. Geosci. Remote Sensing*, GE-25(6):796–804.

[Moon and Stirling, 2000] Moon, T. K. and Stirling, W. C. (2000). *Mathematical Methods and Algorithms for Signal Processing.* Prentice Hall, Upper Saddle River, NJ.

[Natter, 2003] Natter, R. (2003). Untersuchung von Wavelet-basierten Algorithmen in der Auswertung von Weißlichtinterferogrammen. Diploma thesis, FH Karlsruhe.

[Nordberg et al., 1994] Nordberg, K., Gösta, G., and Knutsson, H. (1994). Representation and learning of invariance. In *Image Processing 1994, Proceedings, IEEE International Conference, ICIP-94*, volume 2, pages 585–589, Austin, Texas.

[Olszak and Schmit, 2003] Olszak, A. and Schmit, J. (2003). High-stability white-light interferometry with reference signal for real-time correction of scanning errors. *Opt. Eng.*, 42(1):54–59.

[Oppenheim and Schafer, 1999] Oppenheim, A. V. and Schafer, R. W. (1999). *Discrete-Time Signal Processing.* Prentice Hall Signal Processing Series. Prentice Hall, New Jersey, 2nd edition.

[Perona and Malik, 1990] Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Machine Intell.*, 12(7):629–639.

[Pförtner and Schwider, 2001] Pförtner, A. and Schwider, J. (2001). Dispersion error in white-light Linnek interferometers and its implications for evaluation procedures. *Appl. Opt.*, 40(34):6223–6228.

[Prüssmann et al., 1999] Prüssmann, K. P., Weiger, M., Scheidegger, M. B., and Bösiger, P. (1999). SENSE: sensitivity encoding for fast MRI. *Magn. Reson. Med.*, 42:952–962.

[Rao, 2001] Rao, C. R., editor (2001). *Handbook of Statistics.* Elsevier Science.

[Recknagel et al., 1998] Recknagel, R., Feigl, T., Duparré, A., and Notni, G. (1998). Wide-scale surface measurement using white light interferometry and atomic force microscopy. In *SPIE Proceedings Vol. 3479: Laser Interferometry IX: Applications*, pages 36–42.

[Recknagel et al., 2000] Recknagel, R.-J., Kowarschik, R., and Notni, G. (2000). High-resolution defect detection and noise reduction using wavelet methods for surface measurement. *J. Opt. A: Pure Appl. Opt.*, 2:538–545.

[Recknagel and Notni, 1998] Recknagel, R.-J. and Notni, G. (1998). Analysis of white light interferograms using wavelet methods. *Opt. Commun.*, 148:122–128.

[Reif, 1965] Reif, F. (1965). *Fundamentals Of Statistical And Thermal Physics*. McGraw-Hill, Singapore.

[Restle, 2003] Restle, J. (2003). *Optimierung der Weißlichtinterferometrie für Applikationen der industriellen Qualitätskontrolle*. PhD thesis, Univ. Heidelberg.

[Restle et al., 2004] Restle, J., Hissmann, M., and Hamprecht, F. A. (2004). Nonparametric smoothing of interferometric height maps using "confidence" values. *Opt. Eng.*, 43(4):866–871.

[Rose, 2003] Rose, T., editor (2003). *Oberflächenmesstechnik*, VDI-Berichte Nr. 1806, Düsseldorf. VDI Verlag.

[Saleh and Teich, 1991] Saleh, B. E. A. and Teich, M. C. (1991). *Fundamentals of Photonics*. John Wiley & Sons.

[Sandoz, 1997] Sandoz, P. (1997). Wavelet transform as a processing tool in white-light interferometry. *Opt. Lett.*, 22(14):1065–1067.

[Sandoz et al., 1997] Sandoz, P., Devillers, R., and Plata, A. (1997). Unambiguous profilometry by fringe-order identification in white-light phase-shifting interferometry. *J. Mod. Opt.*

[Sandoz and Jacquot, 1997] Sandoz, P. and Jacquot, M. (1997). Processing of white light correlograms: simultaneous phase and envelope measurements by wavelet transformation. In *SPIE Proceedings Vol. 3098: Optical Inspection and Micromeasurements II*, pages 73–82.

[Scharr et al., 2003] Scharr, H., Felsberg, M., and Forssén, P.-E. (2003). Noise adaptive channel smoothing of low-dose images. In *Computer Vision for the Nano-Scale (Workshop accompanying CVPR 2003)*.

[Schlesinger, 2003] Schlesinger, D. (2003). Gibbs probability distributions for stereo reconstruction. In Michaelis, B. and Gerhard, K., editors, *Proceedings of the 25rd DAGM-Symposium on Pattern Recognition*, Lecture Notes in Computer Science vol. 2781, pages 394–401, Berlin, Heidelberg, New York.

[Schraud, 2000] Schraud, J. (2000). Optimierung der Signalaufnahme und Signalverarbeitung am optischen 3D-Sensor Kohärenzradar. Diploma thesis, Univ. Erlangen-Nürnberg.

[Schwarzer et al., 1996] Schwarzer, H., Teiwes, S., and Wyrowski, F. (1996). Why is it sensible to use wavelets in matched filtering? In *SPIE Proceedings Vol. 2969: Proceedings of the 2nd Int. Conf. on Optical Information Processing*, pages 604–609.

[Seiffert, 2005] Seiffert, T. (2005). *Verfahren zur schnellen Signalaufnahme in der Weißlichtinterferometrie*. PhD thesis, Univ. Erlangen-Nürnberg.

[Shekhar et al., 2002] Shekhar, S., Schrater, P. R., Vatsavai, R. R., Wu, W., and Chawla, S. (2002). Spatial contextual classification and prediction models for mining geospatial data. *IEEE Trans. Multimedia*, 4(2):174–188.

[Sodnik et al., 1991] Sodnik, Z., Fischer, E., Ittner, T., and Tiziani, H. J. (1991). Two-wavelength double heterodyne interferometry using a matched grating technique. *Appl. Opt.*, 30:3139–3144.

[Stahel, 2002] Stahel, W. (2002). *Statistische Datenanalyse. Eine Einführung für Naturwissenschaftler*. Vieweg, Braunschweig, 4th edition.

[The Mathworks, Inc., 2002] The Mathworks, Inc. (2002). MATLAB Version 6.5, Release 13.

[Tikhonov and Arsenin, 1977] Tikhonov, A. and Arsenin, V. (1977). *Solutions of Ill-Posed Problems*. Winston, Washington, DC.

[Wang, 2003] Wang, Y. (2003). Three-dimensional profilometer for super-smooth surface. *Opt. Eng.*, 42(10):3013–3016.

[Weickert, 1998] Weickert, J. (1998). *Anisotropic Diffusion in Image Processing*. B. G. Teubner, Stuttgart.

[Windecker et al., 1999] Windecker, R., Fleischer, M., and Tiziani, H. J. (1999). White-light interferometry with an extended zoom range. *J. Mod. Opt.*, 46(7):1123–1135.

[Windecker and Tiziani, 1999] Windecker, R. and Tiziani, H. J. (1999). Optical roughness measurements using extended white-light interferometry. *Opt. Eng.*, 38(6):1081–1087.

[Winkler, 2003] Winkler, G. (2003). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer, Berlin, Heidelberg, New York, 2nd edition.

[Zhang et al., 1990] Zhang, M. C., Haralick, R. M., and Campbell, J. B. (1990). Multispectral image context classification using stochastic relaxation. *IEEE Trans. Syst., Man, Cybern.*, 20(1):128–140.