Technical Report

Pattern Recognition and Image Processing Group
Institute of Computer Aided Automation
Vienna University of Technology
Favoritenstr. 9/1832
A-1040 Vienna AUSTRIA
Phone: +43 (1) 58801-18351
Fax: +43 (1) 58801-18392
E-mail: nob@prip.tuwien.ac.at
URL: http://www.prip.tuwien.ac.at/

PRIP-TR-70 July 1, 2002

Robust Analysis of Spot Array Images

Norbert Brändle

Abstract

Computer-aided image analysis deals with the automatic recovery of visual information from digital images. Recent image analysis research is trying to find more general and more efficient algorithms. Given a digital image with a certain contents or certain problem domain, it is still mandatory to manually evaluate different approaches and processing sequences to extract useful and plausible information from the image. Once a generic image analysis approach for the problem domain has been found, the method can often be applied independent of the operator's intuition and previous experience. Optical character recognition (OCR) is an example for a successful development from initial research to off-the-shelf products. This work provides a generic image analysis system for a class of images having a characteristic contents denoted as *spot arrays*. Spots are defined as simply connected, irregularly shaped regions lighter (or darker) than their background. Representatives of this image class include images of Braille paper for blind persons and DNA arrays - a tool of modern biotechnology. Analysis of spot array image has three main tasks. The first task is to detect the spots present in the image and therefore deals with the spatial localization process. The spots in the image are located on a grid which may be distorted in the course of the image production process. The second task therefore deals with the fitting of a grid to the detected spots, such that they can be correctly addressed. Once the spots are detected and addressed, they are characterized by their shape, intensity and local background. The automatic image analysis presented in this work is composed of a set of tools arranged in a general framework. This general framework enables to analyze spot arrays of high spot density with possibly multiple overlapping spots. Furthermore, the concept is robust in order to cope with outliers in the spot array and artifacts like image contaminations. These requirements can be fulfilled by robust statistical models: A key principle of grid fitting is to fit straight line models to the rows and columns of the grid. The input for the straight line models is given by a maximum search in matched filter response. Spot characterization is performed by fitting a parametric spot model to the corresponding pixels with the help of a robust M-estimator. In a consecutive step, a semi-parametric fit is possible in order to cope with deviations from the spot model assumptions. Analysis of DNA array images serves as a demonstration of the presented general framework. Here, the intensity of a spot represents the amount of genetic material bound to the corresponding array element. The ultimate image analysis goal of this application is to quantify as exactly as possible the intensity of tens of thousands of possibly overlapping spots. The output of DNA array image analysis yields the raw data for the discovery of specific genes and the genetic control system of organisms. The results of DNA array images demonstrate the successful application of the framework presented in this work on thousands of images resulting from various biological experiments.

Acknowledgments

There are a number of individuals who have made contributions to this work. I am grateful to my thesis supervisor, Prof. Horst Bischof, for his outstanding support for my work. He has always shared his ideas with me and demonstrated to me what it means to work in a scientific manner. His valuable suggestions and criticism have influenced this thesis to a large extent and made the completion of this work possible. A considerable part of this work has been performed in collaboration with the Inflammatory Diseases unit of the Novartis Research Institute Vienna. I am grateful to Hilmar Lapp for complementing my work with a substantial number of time-consuming experiments and some beers. His critical reasoning helped guaranteeing a good quality of this work. I would also thank Horng-Yang Chen for his valuable contributions from the statistical point of view. I want to thank Prof. Walter G. Kropatsch for his kind support and for his vibrating ideas. He has made it possible for me to perform this work in the stimulating environment of the Pattern Recognition and Image Processing Group. Finally, I would like to express my gratitude to my parents, Horst and Isolde Brändle, who have made my studies possible and who always encouraged and supported me in my goals.

Contents

List of Symbols

$\text{argmax}_i(\text{expr}(i))$	the value of i that causes $\text{expr}(i)$ to be maximal
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$	general sets
$\text{card}\,(\mathcal{A})$	cardinality (number of elements) of set $\mathcal{A}$
$\mathcal{A} \subset \mathcal{B}$, $\mathcal{B} \supset \mathcal{A}$	set $\mathcal{A}$ is included in set $\mathcal{B}$
$\mathcal{A} \cap \mathcal{B}$	intersection between sets $\mathcal{A}$ and $\mathcal{B}$
$\emptyset$	empty set
$\mathbb{N}$	set of cardinal numbers (inclusive zero)
$\mathbb{Z}$	set of integer numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}^2$	two-dimensional plane.
$\mathbb{R}^{M \times N}$	set of $M \times N$ matrices over $\mathbb{R}$ $M, N \geq 1$
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots \in \mathbb{R}^{M \times N}$	(uppercase bold) $M \times N$ matrices over $\mathbb{R}$
$\mathbf{A}^T, \mathbf{B}^T, \ldots \in \mathbb{R}^{N \times M}$	transpose of $M \times N$ matrices over $\mathbb{R}$
$\mathbf{a}, \mathbf{b}, \mathbf{c}, \cdots \in \mathbb{R}^d$	(lowercase bold) vectors (points)
$\mathbf{a} \cdot \mathbf{b}$	scalar product between vectors $\mathbf{a}$ and $\mathbf{b}$
F, P, ...	functions
$\partial \text{F} / \partial x$	partial derivative of the function F with respect to x
$\text{sgn}(x)$	Signum function
$\lvert x \rvert$	abolute value of a scalar x
$\widehat{x}$	estimate of the value x
$\circ(x)$	largest integer which is not bigger than $x + 0.5$
$\lceil x \rceil$	smallest integer not smaller than x
$x \propto y$	x is proportional to y

Chapter 1

Introduction and Overview

Image analysis deals with the automatic recovery of visual information from digital images. This information extraction is usually divided into *high-level computer vision methods* and *low-level image processing* [Sonka et al., 1999]. High-level computer vision tries to imitate human cognition and the ability to make decisions according to the information contained in the image [Ullman, 1996]. Low-level methods, on the other hand, usually use very little knowledge about the content of images. Knowledge about image content is normally provided by high-level algorithms or directly by a human who knows the problem domain. Recent research on low-level methods is trying to find more efficient and more general algorithms. Sonka et al. [Sonka et al., 1999] point out a central image analysis problem:

> A complicated and so far unsolved problem is how to order the low-level steps to solve a specific task, and the aim of automating this problem has not yet been achieved. It is usually still a human operator who finds a sequence of relevant operations, and domain-specific knowledge and uncertainty cause much to depend on this operator's intuition and previous experience.

Given a digital image with a certain contents or a certain problem domain, it is therefore still mandatory even for low-level methods to manually evaluate different approaches and processing sequences to extract useful and plausible information from the image. Once a generic image analysis approach for the problem domain has been found, the method can often be applied independent of the operator's intuition and previous experience. Optical character recognition (OCR), for example, has been a very active field for research and development [Mori et al., 1992]. Today, reasonably good OCR packages are off-the-shelf products. The current research in OCR is now addressing documents that are not well handled by the available systems, including severely degraded, omni-font machine-printed text and (unconstrained) handwritten text [Trier et al., 1996]. A generic image analysis system for invoices is presented in [Bayer and Mogg-Schneider, 1997]: The system automatically extracts the requested items such as invoice code, sender information, date, etc. from invoices with arbitrary form layout in arbitrary domains. It consists of two components, an OCR tool which need not be adapted to the current domain, and a component which contains the knowledge about the domain.

In a similar manner, this work provides a generic image analysis system for a class of images having a characteristic contents we denote as a *spot arrays*. In the following, we first describe

spot images and spot array images, their characteristics and the image analysis tasks one might want to solve.

1.1 Spot Images

The term "spot" is not uniquely defined in the literature. The simplest definition of a spot is that of a pixel in a digital image whose intensity is either higher or lower than all the intensities of its eight neighbors [Sher and Rosenfeld, 1989], [Shneier, 1983]. Blanford and Tanimoto [Blanford and Tanimoto, 1988] do not constrain the spot extension to a single pixel and describe bright spots as localities noticeably brighter than their immediate surroundings. Van der Heijden et al. [van der Heijden et al., 1997] define spots in images as phenomena which correspond to certain objects in the scene, where the projections of these objects are so small (relative to the image resolution) that the internal structure of the objects cannot be observed. An encyclopedic definition characterizes a spot as a small area which is different in color, material, or texture from the surrounding area, especially ones that are more or less circular [Microsoft, 2001]. For this work we use the characteristic features described in [Minor and Sklansky, 1981] and [Danker and Rosenfeld, 1981] and define spots as *irregularly shaped regions lighter (or darker) than their background*, which share the following characteristics: they are usually *simply connected* and they are usually *surrounded by a smoothly curved edge*. Note that this definition originally describes the very similar concept of a *blob*, where a blob is often distinguished from a spot by its size [Shneier, 1983]. In many application domains, however, the defined phenomena are denoted as spots.

The galactic nebula image in Fig. 1.1a illustrates characteristic features of a spot image: First, the stars are brighter than the surrounding background area. Second, the stars are convex and have an approximately circular or elliptic shape. If an ellipse in Fig. 1.1a is very elongated, i.e. has a high ratio between the two principal axes, it is likely that it belongs to two or more stars that *overlap* in the digital image. Overlap between adjacent spots can occur due occlusions

(a) Galactic Nebula (b) Microcalcifications (c) Gel Electrophoresis

Figure 1.1. Spot Images. A spot image contains simply connected objects which deviate moderately but not severely from convexity.

caused by different distances of the stars. Another reason for star overlap might be the point spread function of the optical system, describing the intensity variation of the image of an isolated point object located on a uniformly black background [Nalwa, 1993]. If bright spots are blurred by the point spread function, they will interfere with adjacent spots in the digital image. The second example in Fig. 1.1b shows an X-ray image of the breast (mammogram) that helps physicians detect and evaluate breast abnormalities. The abnormalities include small targets called microcalcifications which are tiny calcium deposits that have accumulated in the breast tissue, and appear as a small bright spot in the mammogram [Boccignone et al., 2000]. The third spot image example in Fig. 1.1c is from the domain of biotechnology [Takahashi et al., 1997], [Takahashi and Watanabe, 1998]. It shows a scanned autoradiograph (X-ray film to visualize radioactively labeled molecules) [Johnston et al., 1990] of a 2D gel electrophoresis of DNA. Electrophoresis in this context is the phenomenon of the movement of DNA fragments through a liquid as a result of an electric field formed between electrodes immersed in the medium. Dark spots in Fig. 1.1c correspond to (radioactively labeled) DNA fragments.

Spot image analysis is often partitioned into the following two main tasks (Fig. 1.2): **Spot detection** deals with the spatial localization process: image analysis requires that spots be reliably detected and distinguished from *image noise*. For example, biologists are interested in the position of a DNA fragment in an electrophoresis autoradiograph, because it reveals information about the length of the DNA strand (the longer the DNA fragments, the slower they move within the gel [Lewin, 1997]). Further applications of spot localization include X-ray imagery for ma-

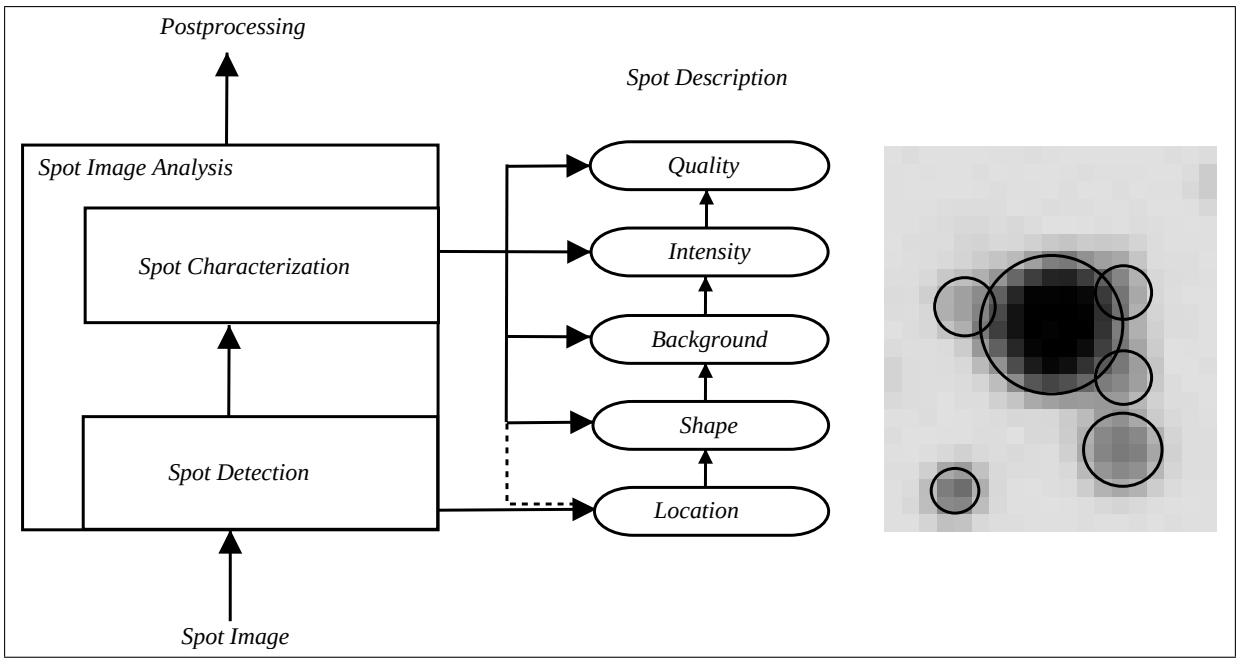


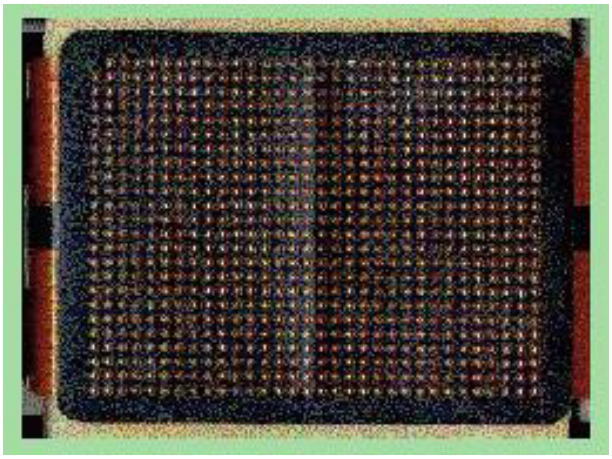
Figure 1.2. Spot Image Analysis. Spot image analysis is often partitioned into spot detection yielding spot locations in the image and subsequent spot characterization. A spot is fully characterized by its location, shape (spatial extension), background and spot signal intensity. Both spot detection and spot characterization can be influenced by interfering image structures like artifacts or overlapping spots. The example spot image in this figure is a zoomed and inverted subimage of Fig. 1.1a. The manually drawn circles show the partly overlapping stars detected by a human.

terial analysis and marker detection in navigation systems. Applications in which the number of spots is primarily more relevant than their exact position include the detection of microcalcifications (Fig. 1.1b) and the counting of trees in aerial photographs of forest [Rewo, 1984]. The second image analysis task, **spot characterization** deals with the exact shape reconstruction of the spots, estimation of spot intensity and spot background (some existing spot characterization methods include detection, see Sect. 1.3). For example, astronomers are not only interested in the position of celestial objects. In order to determine their distance, they also want to extract the red shift of the stars, i.e. the spot intensity in the red light band (Sloan Digital Sky Survey Project [York et al., 2000, Chen et al., 2001]). Spot intensities are often considered as additive signals to a background signal. In the gel electrophoresis example (Fig. 1.1c), one must account for the so-called unspecific radioactivity, producing intensities in the autoradiograph not resulting from DNA-fragments. The distribution of the background signal is usually non-uniform and one must find a reliable way to separate signal and background. The grey level patch in Fig. 1.2 shows a "negative" subimage of Fig. 1.1a and with superimposed stars (circles) detected by a human observer in the original image. The largest star interferes with three adjacent small stars. Spot detection and characterization algorithms should deal with overlapping spots and separate them. If the application requires the spot intensity and background estimation, it is important to assess the reliability of the data obtained. Without assessing the reliability of the data, the conclusions drawn from analyzing such data may be misleading.

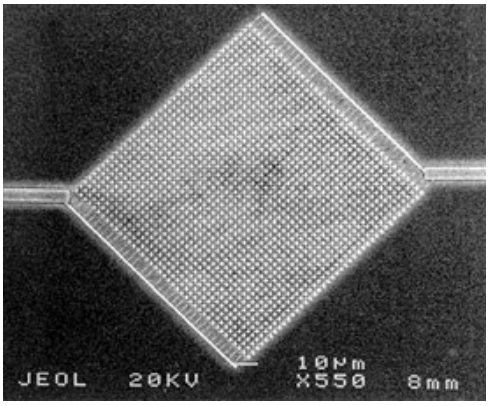
1.2 Spot Array Images

A spot array image consists of an array of spots *arranged in a grid*. Consider, for example, embossed printing with Braille letters for blind persons [Collins and Schneider, 1998]. Fig. 1.3a shows the spot array of a scanned Braille paper with three-dimensional bright spots formed towards a scanner window and dark spots formed backwards a scanner window. With an "Optical Braille Recognition" (OBR) image analysis system blind readers would dispose of a Braille-to-Braille copy machine. Sighted people would have a tool for easy communication with the Braille world. Furthermore, producers of Braille prints would receive a tool for automatic proofreading, for cheap storage and for easy re-printing of old, worn-out unique Braille originals [Halousek, 1999]. Another facility for non-sighted people are *tactile systems* enhancing access to computer graphics and the learning of visual concepts. A tactile array (Fig. 1.3) takes visual information from a computer monitor or camera and transforms it into patterns of small electrical stimuli that can be felt on the skin [Eye Institute, 1996]. Image analysis of tactile arrays would support computer-aided quality control of producers. A similar image analysis application would be the inspection of super-conducting arrays of Fig. 1.3 [Rimberg et al., 1997]. Figure 1.3d shows an image of a high-density DNA array , a tool becoming gradually standard equipment in biotech labs [Chee et al., 1996] [Schena, 2000]. Chapter 3 of this work is devoted to the production and properties of DNA array images. The intensity of a spot in an array element is proportional to the amount of DNA. A noisy example of a high-density DNA array is shown in Fig. 1.3e. In this example, the small bright spots are contaminations, i.e. signals that do not result from genetic material. Furthermore, it is clearly visible that some adjacent spots do overlap. Fig. 1.3f shows a DNA array in which the spots correspond to proteins [MacBeath and Schreiber, 2000].

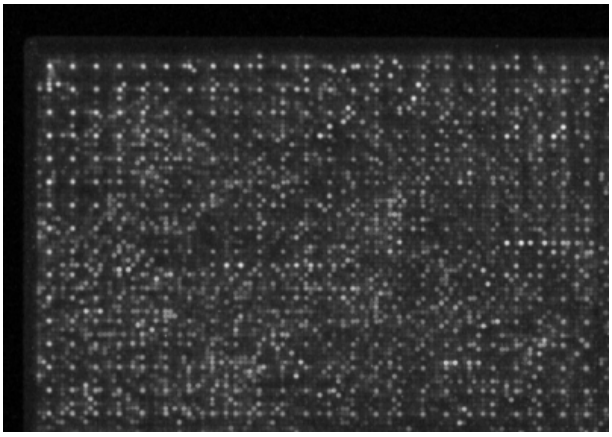
(a) Braille Paper



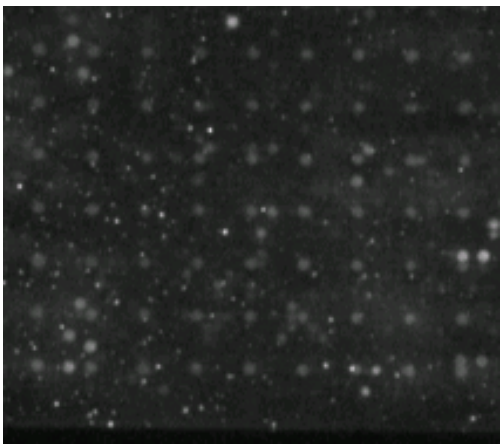
(b) Tactile Array



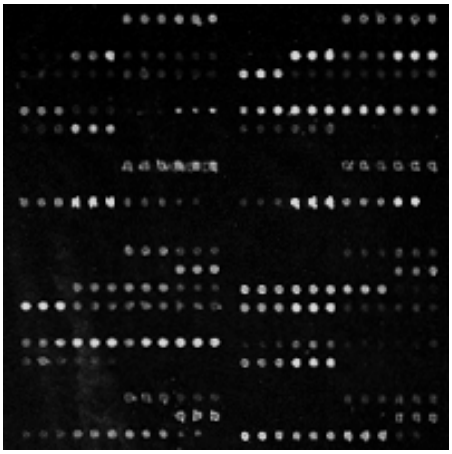
(c) Superconducting Array



(d) High-Density DNA Array



(e) DNA Array with Contaminations



(f) Protein Array

Figure 1.3. Spot Array Images. The spots are arranged in a grid, but the image of the spot array may be distorted.

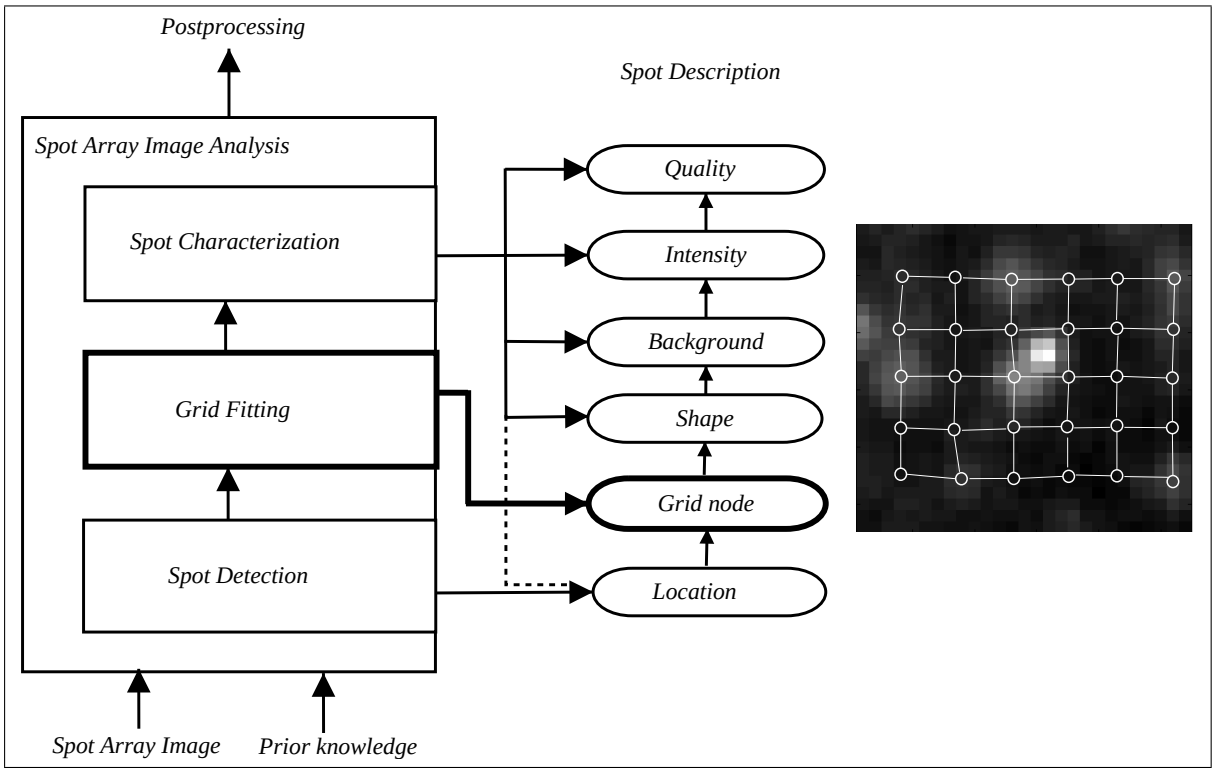


Figure 1.4. Spot Array Image Analysis. In addition to the spot detection and characterization as outlined in Fig. 1.2, spot array image analysis must solve a grid fitting problem. Grid fitting deals with the correct assignment of a detected spot location to a grid node. Furthermore, a fully restored grid reveals location information of undetected spots. The example spot array image is from a DNA array. The grid was manually drawn by a human and illustrates the major image analysis problems. The grid fitting algorithm must tolerate a certain degree of irregularity in spot spacing. At the same time, the algorithm must not be "distracted" by the artifact lying between the grid node.

The image analysis tasks for spot array images are identical to the tasks of spot image analysis described in Sect. 1.1: The spot locations must be detected and its shape must be characterized. There is, however, one task that distinguishes spot array image analysis from general spot image analysis: The spots locations need not merely be detected but rather be assigned to the correct address of the spot array (Fig. 1.4). For example, a detected bright spot in the Braille paper must be assigned to the correct row and column of the corresponding Braille letter in order to yield correct reading results. This addressing task is denoted as **grid fitting** and should cope with the following tasks:

- **Index Detected Spots:** While the spots in a spot array are inherently arranged in rows and columns, the spots in the digital image need not be present in a regular manner, meaning that distortions from a *regular grid* a very likely to occur. Such distortion are often a result of non-linear deformations of the medium carrying the spots (Fig. 1.3a,b). Sometimes the distortions are a result of the (linear) rotation of the spot array in the image (Fig. 1.3c).
- **Cope with Lacking Spots:** Another key issue for grid fitting algorithms is to index the spots correctly even at very low levels of intensity. The spot can be absent (as can occur

in a production error) or have such a low signal that it is not detected. Lacking spots make simple horizontal and vertical ordering algorithms for indexing prohibitive. If one wants to restore the full grid, the locations of the lacking spots must be inferred from the set of the detected spot locations.

- **Ignore Artifacts** A grid fitting algorithm must deal effectively with two opposing criteria. First, due to variation in spot position, as described above, the algorithm must tolerate a certain degree of irregularity in the spot spacing. At the same time, the algorithm must not be "distracted" by artifacts (Fig. 1.3e) that could be adjacent to a true arrayed spot.

1.3 Related Work

To get an idea about the demands a spot array analyzing method must meet, we discuss some existing methods that detect spots and methods that characterize spots. Some methods accomplish spot detection and characterization together. We also show approaches and applications that perform grid fitting on arrays.

1.3.1 Spot Detection

The most straightforward method to detect spots in an image is *matched filtering* [Pratt, 1991] [Schmidt, 1990]. First, the spot image is convolved with a kernel that resembles the shape of the spot. Then, all positions corresponding to a (local) maximum of the convolved image are marked as candidate spots (i.e. non-local maximum suppression). Finally, all candidate spots for which the convolved image exceed a certain threshold are accepted as detected spots. The matched filter approach is sensitive to interfering image structures (edges and lines resulting from other objects in the scene) and overlapping spots. Another approach is to explicitly use the known intensity value of a spot [Schmidt, 1990]. By increasing the error for spots with different intensities, the sensitivity of detecting the spot increases dramatically. However, the variation in their detection value is so large that it is more difficult to distinguish a spot from noise than in ordinary matched filtering.

The second class of spot detectors are approaches based on a *statistical parametric model* of the image data. Here, the spot shape is assumed to correspond to a statistical model with adjustable parameters which are optimized for the observed gray level data. Rewo [Rewo, 1984] models convex objects (trees) as two-dimensional second order polynomials. He derives a spot detection mask which enhances convex objects by minimization of the squared error between the model and the observed data. A pixel is a spot element if the detector output exceeds a positive threshold t. Noordmans and Smeulders [Noordmans and Smeulders, 1998] detect spots with the help of Gaussian (isotropic) spot models of different sizes. For every image position and every selected spot size (standard deviation), they fit the remaining parameters to the data and compute a match error between the model and the data. If the match error is a local minimum and the spot intensity above a threshold, the corresponding position is a spot with the fitted parameters. The parameters are refined in a characterization step, where it is possible to deal with overlapping spots (see next section). Van der Heijden [van der Heijden et al., 1997] also models spots as two-dimensional Gaussians and provides numerical algorithms to find the optimal parameters of the spot detector. The optimizations are performed for the covariance

model (cvm) operator [van der Heijden, 1995], which is a parallel bank of K filters with different kernels, the squared outputs of which are summed with weights. The first kernel has a positive weight and resembles the spot shape. The other kernels have negative weights and have zero output at the position of spot. Correspondingly, their squared (and thus positive) outputs are subtracted from the squared output of the first filter. These filters therefore serve to sharpen the peaks in the detector output without affecting the maximum of these peaks.

An efficient way of detecting differently sized spots is given by the class of *hierarchical operators*. Blanford and Tanimoto [Blanford and Tanimoto, 1988] provide hierarchical algorithms to detect bright spots in image pyramids. An image pyramid is a collection of images of a single scene at exponentially decreasing resolutions. The bottom level of the pyramid is the original image. In the simplest case, each successive level of the pyramid is obtained from the previous level by a filtering operation followed by a sampling operator [Haralick et al., 1991]. The basic bright-spot detection algorithm works with this pyramid as follows: starting at the top level (consisting of one pixel), a path is traced down to the original image, at each step moving from the current cell to is child with the highest value. They provide extensions to detect several spots. The algorithms only work well for images where the spot and background intensities are relatively constant [Blanford and Tanimoto, 1988]. Another hierarchical way for the detection of spots is given by the wavelet transform [Antoine et al., 1993], [Mallat, 1999]. Wavelets are scaled waveforms that measure signal variations and can therefore be used to detect sharp signal transitions. Strickland and Hahn [Strickland and Hahn, 1996] employ a wavelet transform which acts as a bank of multiscale matched filters for detecting microcalcifications in mammograms (Fig. 1.1b). Since the matched filter technique requires specific knowledge of the target signal, they introduce the assumptions that the spots to be detected are truly Gaussian objects.

By choosing the Laplacian (second derivative) of Gaussian wavelets, the wavelet transform performs as a multiscale matched filter: A filter bank convolves a signal with a low-pass filter and a high-pass filter and subsamples by 2 the output (high image frequencies correspond to sharp signal transitions). At each scale, the output of the different subbands can be thresholded to produce a detect/no detect result. The detection threshold is experimentally chosen as a fixed percentile of the histogram of each channel. A drawback of this approach is that a certain number of pixels are always detected at each threshold. An approach which automatically determines the threshold function by relying on an information-theoretic tool is presented in [Boccignone et al., 2000].

1.3.2 Spot Characterization

Spot characterization deals with the exact shape reconstruction of the spots. If the spot detection is performed with a statistical model like a Gaussian, it is possible to reduce errors of the model in the characterization steps as follows: Based on the dimensions of spot, a local image is extracted around the spot. From the local image, the provisional models of the neighboring spots are subtracted to reduce the disturbing influence of artifacts and overlapping spots [Noordmans and Smeulders, 1998]. When no statistical model is used, the shape reconstruction is accomplished as an *image segmentation* task. Image segmentation results in a set of disjoint regions corresponding uniquely with spots in the input image. Image segmentation is a wide field of research in the image analysis literature [Sonka et al., 1999]. We will only review methods specifically dedicated to the segmentation of spot-like objects. If a spot detection is

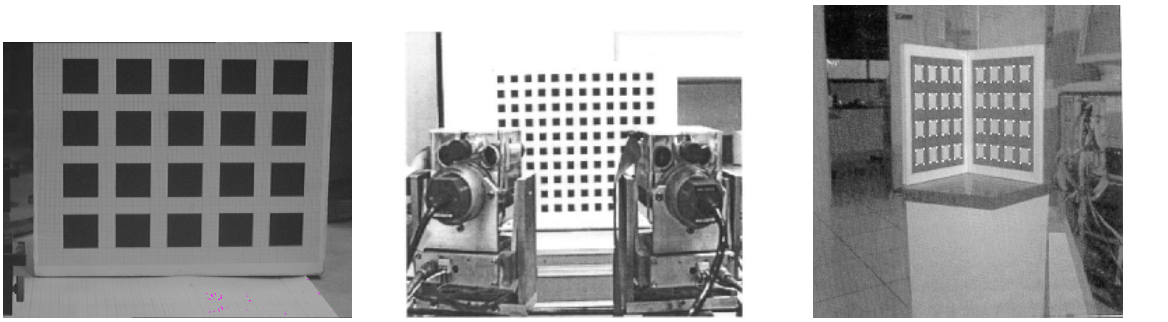
performed prior to spot characterization, segmentation can be constrained to local images extracted around detected spot centers. On the other hand, segmentation is often performed on the entire image and therefore serves as a tool for both spot detection and characterization.

Thresholding

The simplest segmentation approach is to examine all image elements which exceed a given *threshold*. Many different techniques have been used to select good thresholds for this purpose [Sahoo et al., 1988]. Threshold selection involves choosing a gray level t such that all gray levels greater than t are mapped into the "spot" label, while all other gray levels are mapped into the "background" label (segmentation by pixel classification). However, if the image is noisy, thresholding will produce noisy results which may not be cleaned up in postprocessing. Thresholding may extract regions that are not bounded by edge but are smooth continuations of the background if the gray level fluctuations in the background cross the threshold level. In order to overcome these problems, several local thresholds can be extracted from various parts of the picture, and can be applied only in those regions. Shneier [Shneier, 1983] describes a method of identifying parts of a picture on which to apply a threshold, and a means of calculating a local threshold for each of these parts. The method involves constructing an image pyramid, each of lower resolution than its predecessor (Sect. 1.3.1). At some level of the pyramid, it is to be expected that any spot-like object should be contained in a single point which has a higher value than its neighbors. Thus, by detecting such points in low-resolution images, the interesting regions in the picture can be discovered, and only those regions need to be thresholded. Danke and Rosenfeld describe a relaxation approach [Danker and Rosenfeld, 1981] that can be used to defer the classification decisions (spot pixel or background pixel?) until more information is available. In their approach, a degree of membership for each pixel in each class is computed (a probability that it belongs to each class). These membership values are adjusted, based on the values at neighboring pixels and the compatibilities of the various possible combinations of class memberships of pairs of neighbors. After a few iterations, the membership values stabilize, with some values becoming or remaining relatively high and others becoming very low, so that it becomes easier to make the final classification decision.

Edge-based segmentation

Edge-based segmentations rely on edges found in an image by edge detecting operators, which are pre-processing methods used to locate changes in the intensity function [Sonka et al., 1999]. Since the image resulting from edge detection cannot be used as a segmentation result, supplementary processing steps must follow. For example, an early work using an edge detector is described in [Minor and Sklansky, 1981], where the spots are found by an edge-based segmentation, subsequent labelling and classification of labels into spot and non-spot. However, the edge detector may respond in the interior of the spot or background as a result of noise or may fail to respond strongly on the spot/background border because of blur. A very effective method to detect spots with a known boundary, e.g. circular, is the Hough transform [Illingworth and Kittler, 1988], which can also be used successfully in segmentation of overlapping spots. A theoretical statistical approach with maximum likelihood estimation to detect spot boundaries in noisy images was described in [Cooper, 1979]. However, the spots are re-



(a) Calibration pattern (b) Stereo cameras (c) Detected corners [Faugeras, 1993]

Figure 1.5. Grid Fitting for explicit camera calibration. In this examples, the corner points of the squares on the calibration object have to be determined as accurately as possible and identified on the calibration grid.

stricted to a constant foreground and background.

Region-based segmentation

The aim of thresholding and and edge-based segmentation described in the previous sections was to find borders between regions. Region-based segmentation methods construct regions directly. Seeded Region growing [Adams and Bischof, 1994] is well suited for spot images. After specification of seeds, the algorithm proceeds by growing all the foreground and background regions simultaneously until all pixels in the image have been allocated to one of the regions. At each stage, all pixels which are as yet unallocated, but which have at least one neighbor which has already been allocated, are considered for allocation. Out of all these region-neighboring pixels, the algorithm selects the one whose pixel value is nearest (in terms of absolute gray level difference) to the average of the pixel values in the neighboring region. The process repeats until all pixels have been allocated. For spot images, the foreground and background seeds are chosen using the spot detector output. The spots can also be grown from the detected locations using the watershed transformation, which is motivated by the topographic view of images [Beucher and Meyer, 1993]. Binary morphological segmentations are able to cope with overlapping spots, where morphological gray-level segmentation produces similar outputs than edge-based segmentation [Sonka et al., 1999].

1.3.3 Grid Fitting

Camera calibration [Abdel-Aziz and Karara, 1971], [Slama, 1980] was perhaps the first application where a grid fitting problem had to be solved. Camera calibration is a necessary step in 3D computer vision in order to extract metric information from 2D images. It relates the locations of pixels in the digital image to points in the three-dimensional scene [Jain et al., 1995]. Two calibrated cameras are used to make three-dimensional measurements with the help of stereo vision [Gennery, 1979, Sonka et al., 1999]. Explicit camera calibration is performed by

observing a calibration object [1] whose geometry in the three-dimensional space is known with very good precision [Trucco and Verri, 1998]. The calibration pattern is often a planar grid of known-sized boxes on a contrasting background [Tsai, 1987], [Zhang, 1999]. Figure 1.5a shows an example for a calibration pattern, Fig. 1.5b shows a camera setup for stereo vision [Batista et al., 1999]. Before the parameters of the camera model are computed, two preprocessing steps are necessary

1. **Detect Feature Points:** The feature points of the calibration objects have to be reliably extracted. For box grids the feature points are the four corner points of each box (Fig. 1.5c). If the feature points are selected manually as in [Broadhurst and Cipolla, 1999], the image is often preprocessed with a corner detector [Haralick and Shapiro, 1992]. Automatic feature point detectors usually first apply an edge detector [Canny, 1986], find box edges and fit lines to each box edge [Thacker and Lacey, 2000] and intersect lines to find corner points [Bryant et al., 2000]. For calibration patterns with circular features, the center points are usually located with subpixel precision ([Heikkilä and Silvén, 1997] describe how the distortions of perspective projections of circles can be corrected).

2. **Identify Points:** The grid fitting problem for camera calibration is to reliably assign the detected feature points in the image to the three-dimensional coordinates of the calibration pattern. In typical calibration environments usually all the feature points are detected. Simple point sorting algorithms are then used to identify the points on the basis of their vertical and horizontal ordering. A camera calibration application for an autonomous underwater vehicle in which the detection of all points cannot be relied upon is described in [Bryant et al., 2000]. They identify a box in the calibration pattern by computing a planar projective invariant index [Rothwell, 1995] with all other boxes. 5 predetermined corner points of each box pair are chosen such that certain collinearity constraints hold [Bryant et al., 2000]. This method is generally not applicable to spot array images since no organization in simple sub-units like “boxes” can be assumed.

Some grid fitting methods have been published in the field of DNA arrays [Hartelius, 1996], [Zhou et al., 2001] [Bergemann et al., 2001], [Steinfath et al., 2001]. They will be reviewed in chapter 4.

1.4 The Goals of this Work

The ultimate goal of spot array image analysis is an automatic system that utilizes algorithms to find, index and characterize the spots without the need for any human intervention. This work provides a set of tools and a framework which only requires the user to specify the spot array configuration (i.e. the number of rows and columns, image resolution). Automatic image analysis of an undistorted and fully-occupied array with displaced spots is not a difficult task and could be immediately solved with the methods described in Sect. 1.3. The original contribution of this work is a **framework of spot array image analysis methods** which is able to deal with

1. **high spot density** with possibly **multiple overlapping spots**

[1] Implicit camera calibration techniques do not use any calibration object, see for example [Maybank and Faugeras, 1992]

2. **outliers in the spot array**
3. **artifacts in the spot array image**
4. **non-regular spot distances**.

Conditions 1 to 4 require that the spot array image analysis method be robust. In order to accomplish this task, a novel robust grid fitting procedure and a novel robust spot characterization method based on statistical models is presented. It is assumed that the spot array is planar and approximately parallel to the image plane of the image acquisition system (camera, scanner). Large perspective distortions of an array like the calibration pattern in Fig. 1.5 are therefore not allowed. This restriction does not limit the usability of the framework, since its main application lies in the field of inspection, where the image acquisition systems are mainly oriented in the above-mentioned manner for practical reasons. It is also due to the industrial setup that prior knowledge about spot array configuration and imaging parameters can be assumed to be available.

1.5 Overview

Each chapter starts with a brief introduction and is concluded with a short summary. In Chapter 2 the core methods of the spot array image analysis framework for robust grid fitting and spot fitting are presented. A major application of spot array image analysis is introduced in Chapter 3, where key principles and techniques for DNA array production are described. Readers familiar with the field DNA array technology may skip this chapter and move on to Chapter 4, where the spot array image analysis is applied to various DNA arrays. Furthermore, related work is presented and compared to an existing method. Chapter 5 concludes this work and gives a brief outlook towards future work.

Chapter 2

General Framework

This chapter provides a general framework of tools and processing sequences to solve the image analysis task for an image containing a spot array. Fig. 2.1 illustrates the task decomposition proposed in this work. Spot detection and grid fitting finds potential spot locations in the image and fits a consistent grid to the locations of the possibly distorted grid nodes. Spot detection and grid fitting is divided into three substeps, where the first substep is an amplification of the spot locations (Sect. 2.2). The subsequent rotation estimation step (Sect. 2.3) determines the global rotation of the spot array in the image. Finally, a consistent grid is spanned with the help of the rotation estimation and the amplified spot locations (Sect. 2.4). The spot characterization part is divided into the background estimation (Sect. 2.5) and the spot fitting step (Sect. 2.6). The general framework is based upon the following principles:

1. **Keep as much as information as possible:** Processes involving loss of information like geometric image transformations and thresholding should be avoided.
2. **Bring in as much prior knowledge as possible:** By considering different constraints inherent in the imaging process a fast image analysis should be obtained without using unnecessarily complicated processes.

2.1 Formal representation

This section introduces the notation and definitions of the main modeling concepts of the spot array image and the spot grid.

2.1.1 Spot Array Image Representation

A $M \times N$ *spot array image* is represented as a matrix $\mathbf{S} \in \mathbb{Z}^{M \times N}$ with pixel coordinates (m, n) and pixel intensities $\mathbf{S}[m, n] \in 2^B$ with $B \in \mathbb{N}$ as the bit-depth or radiometric resolution:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}[1,1] & \mathbf{S}[1,2] & \dots & \mathbf{S}[1,N] \\ \mathbf{S}[2,1] & \mathbf{S}[2,2] & \dots & \mathbf{S}[2,N] \\ & \dots & & \\ \mathbf{S}[M,1] & \mathbf{S}[M,2] & \dots & \mathbf{S}[M,N] \end{bmatrix} \tag{2.1}$$

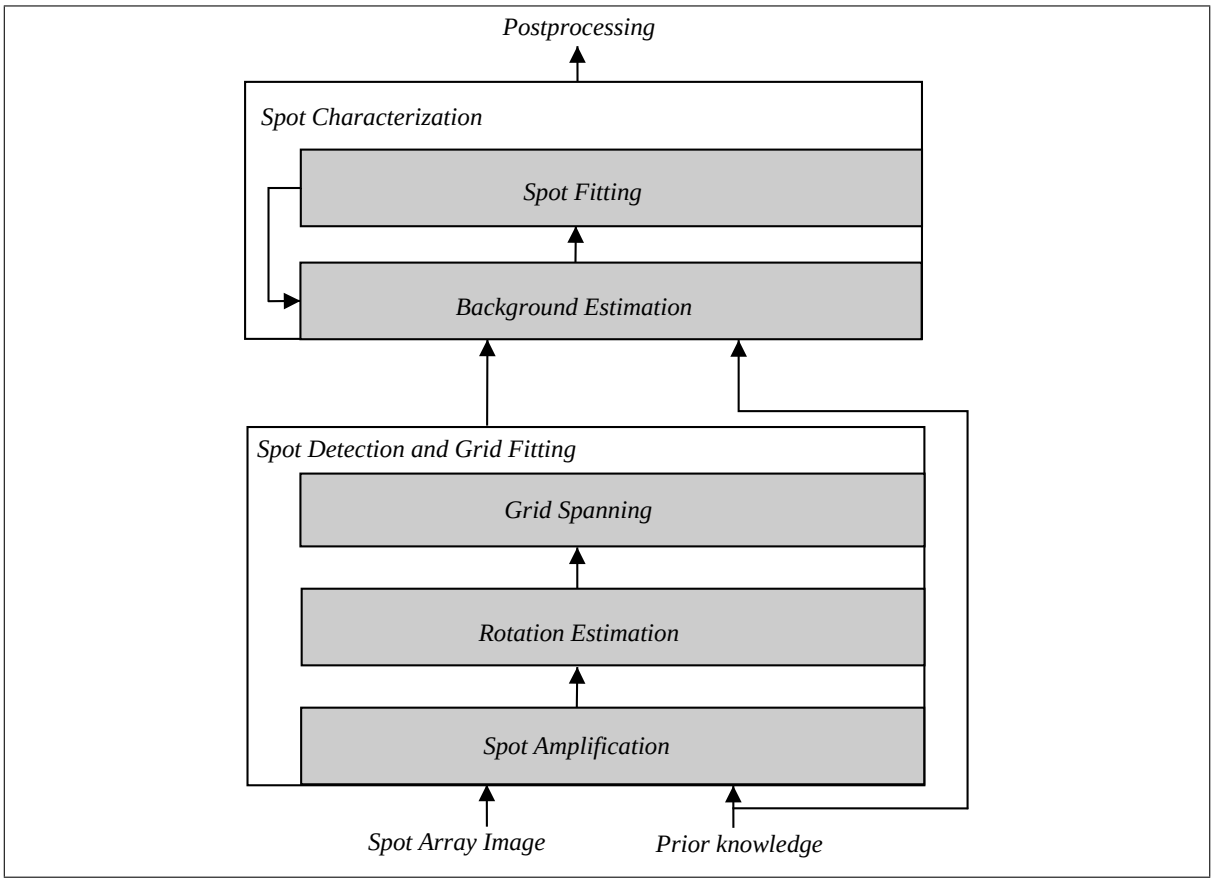


Figure 2.1. General Spot Array Image Analysis Framework. Overview of the general approach for the image analysis of spot array images.

Some algorithms are more conveniently formulated in Cartesian coordinates rather than the above pixel coordinates: The *spatial coordinate system* (x, y) has its origin at the upper left corner with x increasing to the right and y increasing downward. The *central coordinate system* (x', y') has its the origin at the image center width x increasing to right and y increasing upward. Cartesian coordinates are also represented as vectors $\mathbf{p} = [x\ y]^T \in \mathbb{R}^2$.

2.1.2 Grid Representation

A *grid* $\mathcal{G}$ is a set of nodes in $\{1 \dots I_{\mathrm{G}}\} \times \{1 \dots J_{\mathrm{G}}\}$, with I_{G} as the number of grid rows and J_{G} as the number of grid columns. The grid row extraction relation $\mathrm{R}_{\mathcal{G}}$ extracts J_{G} grid nodes belonging to one row of a grid $\mathcal{G}$:

$$\mathrm{R}_{\mathcal{G}}(\{1 \dots I_{\mathrm{G}}\}) = \{(i,j) \mid (i,j) \in \mathcal{G} \wedge i = k\} \qquad \forall k \in \{1 \dots I_{\mathrm{G}}\} \tag{2.2}$$

Similarly, the column extraction relation $\mathrm{C}_{\mathcal{G}}$ extracts I_{G} grid nodes belonging to one column of a grid $\mathcal{G}$:

$$\mathrm{C}_{\mathcal{G}}(\{1 \dots J_{\mathrm{G}}\}) = \{(i,j) \mid (i,j) \in \mathcal{G} \wedge j = l\} \qquad \forall l \in \{1 \dots J_{\mathrm{G}}\}. \tag{2.3}$$

Grid fitting is the location assignment function $\mathrm{L} : \mathcal{G} \times \mathbf{S} \mapsto \mathcal{L}$, where $\mathcal{L} = \{\mathbf{p} \mid \mathbf{p} \in \mathbb{R}^2\}$. The location of a grid node $g_{ij} = (i, j) \in \mathcal{G}$ in the spot array image $\mathbf{S}$ will sometimes be simply denoted as $\mathrm{L}(g_{ij})$.

2.1.3 Prior Knowledge

The prior knowledge apart from the grid dimensions I_{G} and J_{G} consists of the theoretical distance between two spot locations in the spot array image. The *theoretical horizontal spot distance* $S_x \in \mathbb{R}$ and the *theoretical vertical spot distance* $S_y \in \mathbb{R}$ are the distances in sub-pixels between two adjacent spots in the horizontal and vertical direction computed as

$$S_x = \frac{N_x}{\Delta x} \qquad \text{and} \qquad S_y = \frac{N_y}{\Delta y} \tag{2.4}$$

with N_x and N_y as the spot distances on the medium in millimeters and Δx and Δy as the scanner or camera resolution in millimeters per pixel. Note that S_x and S_y are not rounded to integer values.

2.2 Spot Amplification

The first step towards a detection of spots is an amplification of their locations. We want to find signals in the spot array image which resemble the shape of the spots. We chose matched filters (MF), which are filters whose shape match the shape of the signals one is trying to find. The MF is optimal with respect to Gaussian noise. In order to find a random signal with non-zero mean in white noise, the filter should be matched to the mean of the signal The MF for spots is constructed by forming an average template from a number of representative spots in the following way: The image patches $\mathbf{S}_k$, $k \in \{1 \dots S\}$ with the dimension $M_{\mathrm{MF}} \times N_{\mathrm{MF}}$ contain the intensity values of the S spots which are manually selected by the user. The matched filter dimensions M_{MF} and N_{MF} should cover the spot extension for a given image resolution and spotting geometry. We formally define the matched filter dimensions as a function of the theoretical spot distances S_y and S_x defined in (2.4):

$$M_{\mathrm{MF}} = \begin{cases} \circ(S_y) + 2 & \text{for } \circ\,(S_y) = 2k + 1 \\ \circ(S_y) + 1 & \text{otherwise} \end{cases} \tag{2.5}$$

and

$$N_{\mathrm{MF}} = \begin{cases} \circ(S_x) + 2 & \text{for } \circ\,(S_x) = 2k + 1 \\ \circ(S_x) + 1 & \text{otherwise} \end{cases} \tag{2.6}$$

with $k \in \mathbb{N}$ and $\circ(.)$ as the rounding operator. Irrespective of its parity, the theoretical spot size rounded to the next integer is increased to the next higher odd number in order to guarantee to include some background information in the filter.

The S image patches $\mathbf{S}_k$ containing the spots are formally rearranged as D-dimensional vectors $\mathbf{s}_k$ by lexicographical ordering, with $D = M_{\mathrm{MF}} \cdot N_{\mathrm{MF}}$. The vectors $\mathbf{s}_k$ are normalized as

$$\tilde{\mathbf{s}}_k = \frac{\mathbf{s}_k - \mu_{\mathbf{s}_k} \cdot \mathbf{1}}{\sigma_{\mathbf{s}_k}}, \tag{2.7}$$

where $\mathbf{1}$ is a $D \times 1$ vector of ones and $\mu_{\mathbf{s}_k}$ is the mean intensity value of the the image patch defined as

$$\mu_{\mathbf{s}_k} = \frac{1}{D} \sum_{i=1}^{D} \mathbf{s}_k[i], \tag{2.8}$$

and $\sigma_{\mathbf{s}_k}$ is the intensity standard deviation

$$\sigma_{\mathbf{s}_k} = \sqrt{\frac{1}{D-1} \sum_{i=1}^{D} (\mathbf{s}_k[i] - \mu_{\mathbf{s}_k})^2}. \tag{2.9}$$

The matched filter is constructed by averaging the G normalized examples $\tilde{\mathbf{s}}_k$:

$$\tilde{\mathbf{m}} = \frac{1}{G} \sum_{k=1}^{G} \tilde{\mathbf{s}}_k. \tag{2.10}$$

Although the filter $\tilde{\mathbf{m}}$ is guaranteed to have zero mean, it must be renormalized to a contrast of one [Nayar and Poggio, 1996]:

$$\mathbf{m} = \tilde{\mathbf{m}} / \sigma_{\tilde{\mathbf{m}}} \tag{2.11}$$

with the standard deviation $\sigma_{\tilde{\mathbf{m}}}$ computed similar as in (2.9). Filtering of the spot array image $\mathbf{S}$ with the matched filter $\mathbf{m}$ results in a response image $\mathbf{R}^{\text{M}}$ which is constructed as follows: If $\mathbf{p}_{[m,n]}$ denotes an $M_{\text{MF}} \times N_{\text{MF}}$ image patch around a pixel $\mathbf{S}[m,n]$ rearranged as a D-dimensional vector, the image patch $\mathbf{p}_{[m,n]}$ is first normalized to zero mean and contrast 1 similarly to (2.7) as

$$\tilde{\mathbf{p}}_{[m,n]} = \frac{\mathbf{p}_{[m,n]} - \mu_{\mathbf{p}_{[m,n]}}}{\sigma_{\mathbf{p}_{[\mathbf{m},\mathbf{n}]}}} \cdot \mathbf{1}. \tag{2.12}$$

$$\mathbf{R}^{\text{M}}[m,n] = \tilde{\mathbf{p}}_{[m,n]} \cdot \mathbf{m} \tag{2.13}$$

corresponding to the statistical cross-correlation between the image patch and the matched filter. Thus, response values close to one indicate that the image patch is strongly correlated with the filter. Figure 2.2 shows a matched filter example for a simple spot array image.

Note that when the matched filter and the image patch are not normalized by the standard deviation, the response image will correspond to the statistical covariance between the the image patch and the matched filter. In this case, the spot intensity will also be taken into account and yield high response values for bright spots compared to low response values for dark spots.

2.3 Rotation Estimation

The rotation estimation step provides the basis data for the grid spanning step with the help of *intensity projections*. A one-dimensional projection of a two-dimensional continuous function $\mathrm{f}(x,y)$ is a line integral in a certain direction θ, also known as the Radon transform [Jain, 1986]. In general, the Radon transform of a function $\mathrm{f}(x,y)$ is the line integral of f parallel to the y'-axis:

$$\mathrm{R}_\theta(x') = \int_{-\infty}^{\infty} \mathrm{f}(x' \cos\theta - y' \sin\theta, x' \sin\theta - y' \cos\theta) dy' \tag{2.14}$$

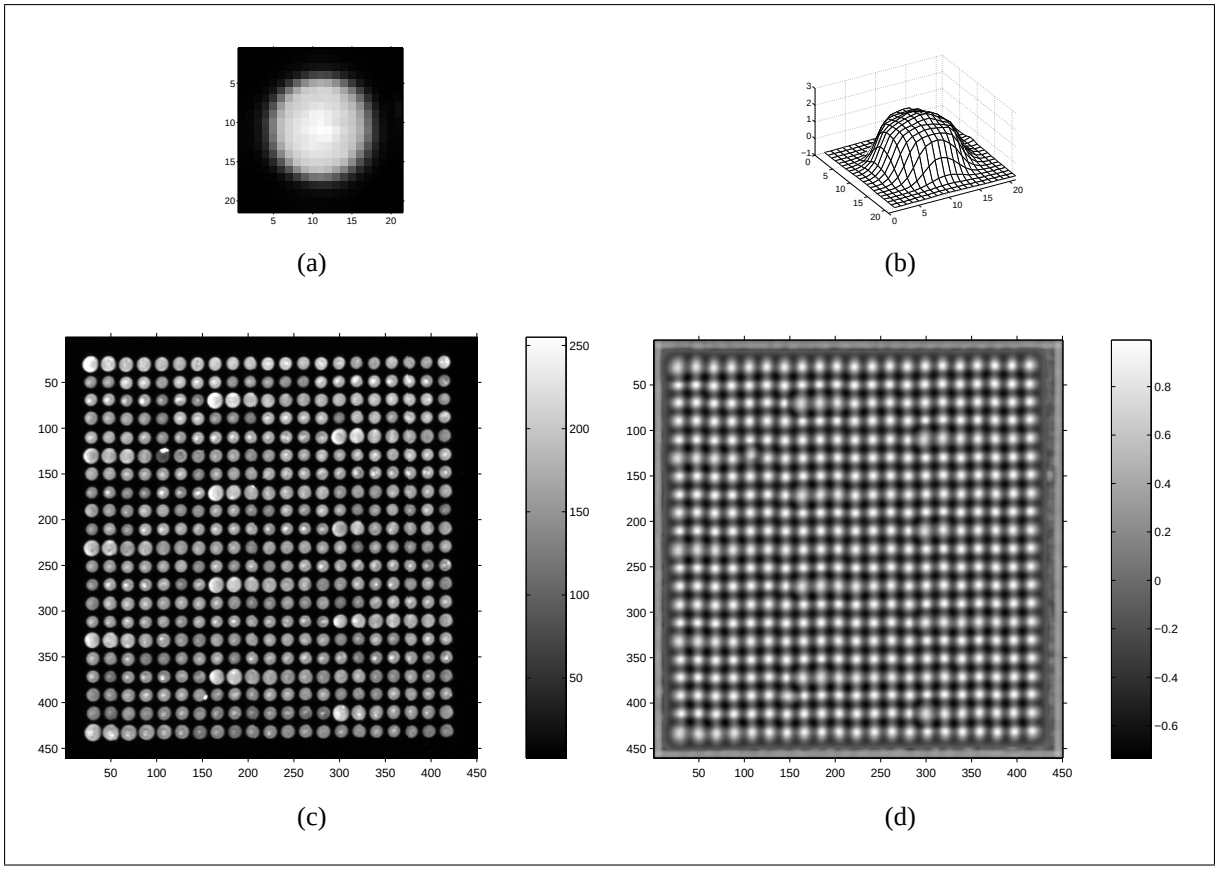


Figure 2.2. Matched Filtering. (a) Matched filter for spots in images of the type shown in (c). It is a normalized spot template built from manually selected spot examples. The dimensions of the filter are determined by the prior knowledge about the image resolution and the spot distance. (b) Three-dimensional plot of matched filter. (c) Spot Array Image Example (d) Response image for spot array image (c) filtered with matched filter (a). Response values close to 1 indicate strong correlation between the image patch and the matched filter.

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{2.15}$$

Figure 2.3 illustrates the geometry of the Radon transform. The *discrete* Radon transform $\mathbf{R}^{\mathrm{T}}$ is defined as an $R \times C$ matrix. The number of rows R corresponds to the number of directions for which the projection is computed. R depends on the angle resolution $\Delta\theta$ of the projection angles and the maximum rotation angle θ_{M}:

$$R = \frac{2\theta_{\mathrm{M}} + 1}{\Delta\theta} \tag{2.16}$$

Note that also negative rotation angles must be considered, hence the factor 2 in (2.16). The rotation angle $\theta(r)$ belonging to a row index $r \in \{1 \ldots R\}$ is given by

$$\theta(r) = r\Delta\theta - \theta_{\mathrm{M}} - 1. \tag{2.17}$$

The number of columns C is defined as the size of the spot array image diagonal:

$$C = \left\lceil \sqrt{M^2 + N^2} \right\rceil . \tag{2.18}$$

The simplest projections $\mathrm{P}(\mathbf{S}, \theta)$ are those for $\theta = 0$ and $\theta = \pi/2$, i.e. the projections to the image coordinates x and y. Figure 2.4a shows the projection $\mathrm{P}(\mathbf{S}, \pi/2)$ of the intensities of a spot array image to the y-axis. One would expect that the spots will be most distinguishable in the projection for the correct global rotation angle $\theta_{\mathcal{G}}$: For a non-distorted grid, the line integral for the correct angle will intersect all spot centers in the spot array image, giving rise to a high projection value (low projection value for dark spots). Similarly, background regions will give rise to low projection values (high projection values in the case of bright spots). Note that while the horizontal projection and vertical projection of the spot image $\mathbf{S}$ might provide useful information for a part of the subsequent grid spanning step, the global rotation angle $\theta_{\mathcal{G}}$ can be best estimated with projections of the spot amplification response image $\mathbf{R}^{\mathrm{M}}$. Figure 2.4b shows a projection $\mathrm{P}(\mathbf{R}^{\mathrm{M}}, \theta_{\mathcal{G}_{\mathrm{C}}})$ through the grid columns at the correct grid rotation angle. The highest projection values correspond to the spot array columns. Fig. 2.4c shows the different projections (discrete Radon transform) $\mathrm{P}(\mathbf{R}^{\mathrm{M}}, \theta)$ with $\theta = (-6 : 0.125 : 6) * \pi/180$. The global grid rotation estimate $\theta_{\mathcal{G}} = \theta(r)$ is in the row r of the Radon transform $\mathbf{R}^{\mathrm{T}}$ having the highest projection values. Let $\mathcal{M}_r$ be the set of the J_{G} (number of spots in a row) highest projection values in the row r of $\mathbf{R}^{\mathrm{T}}$. The estimated rotation angle is then found in the row with the highest median of $\mathcal{M}_r$:

$$r = \mathrm{argmax}_r\{\mathrm{median}(\mathcal{M}_r)\}. \tag{2.19}$$

The median is used because of the following reason: The row with the true rotation angle has the highest projection values but large intervals of low projection values. The other rows have

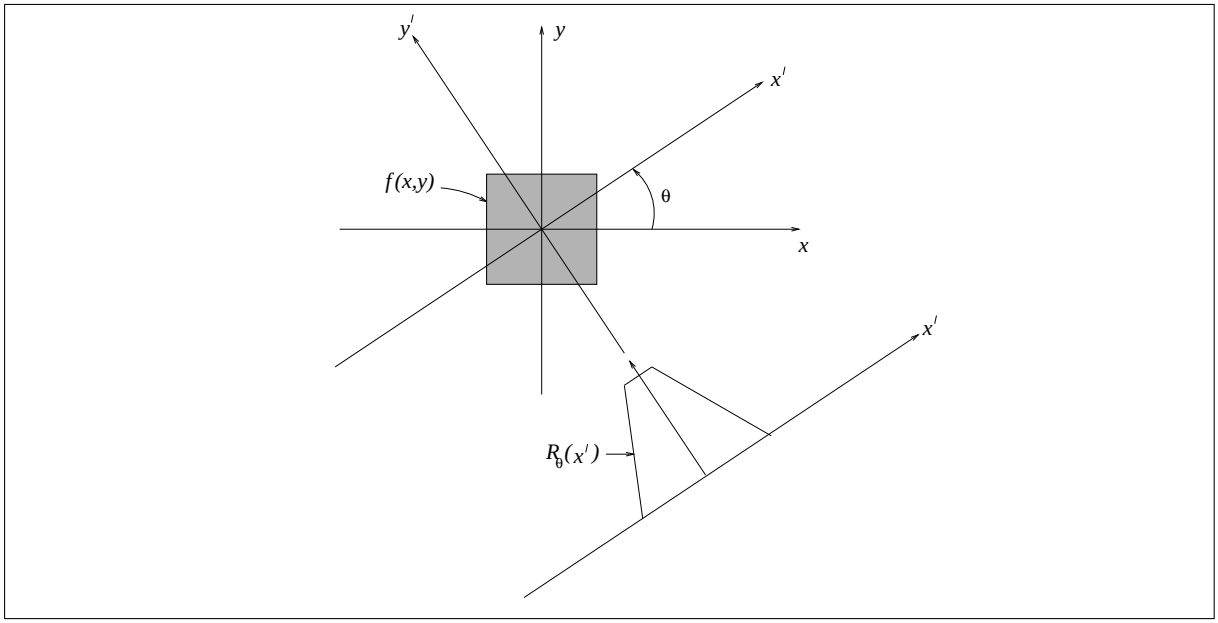


Figure 2.3. Geometry of the Radon transform. A function $\mathrm{f}(x, y)$ is projected along the y'-axis of a coordinate system rotated by the angle θ.

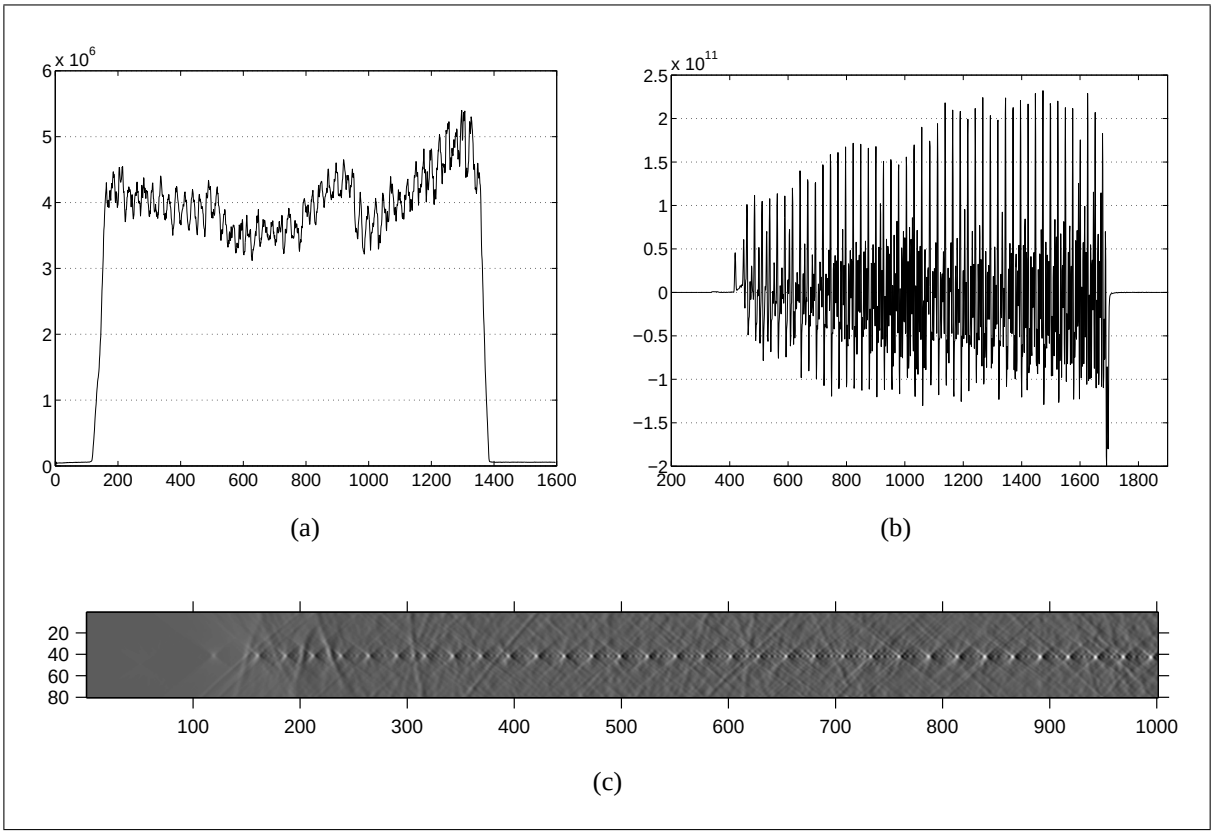


Figure 2.4. Rotation estimation by projections. The rotation estimation is determined by intensity projections. (a) Horizontal projection of the spot image $\mathbf{S}$ to the y-axis. (b) Projection of $\mathbf{R}^{\mathrm{M}}$ at the correct grid angle. (c) Discrete Radon transform between the angles -6° and 6° in steps of 0.125°.

lower projection values, but they are distributed over the columns. A simple horizontal addition of the projection values would therefore not lead to reliable rotation estimations.

The computation of a row of the Radon transform $\mathbf{R}^{\mathrm{T}}$ involves a transformation of all the pixels of the spot array image $\mathbf{S}$. In order to increase the efficiency, one should compute the Radon transform in a hierarchical manner, determine a reasonable maximum rotation angle θ_{M} and an angle resolution $\Delta\theta$. These issues are covered in the Appendix.

2.4 Grid Spanning

Grid spanning is the core step of grid fitting and tries to reconstruct the spot grid of the spot array image with the help of the matched filter response and the rotation estimation. The information gained with the MF response image and the rotation estimation is used to define an initial spot grid. Two ways to define such an initial spot grid are presented in Sections 2.4.1 and 2.4.2. The initial spot grid is subsequently parameterized in order to remove false negatives and false positive spot locations. Grid parameterization is described in Sect. 2.4.3.

2.4.1 Algorithm I: Initial Grid after Maximum Search

One way to achieve an initial spot grid is to extract a set of potential spot locations $\mathcal{L}$ by a local maximum search in $\mathbf{R}^{\text{M}}$ as a first step. The detected locations $\mathcal{L}$ are subsequently aligned on a transformed prior spot grid defined by the theoretical spot distances.

Maximum Search Algorithm

Pass 1 The initial spot location set $\mathcal{L}$ consists of the locations of the maximum value in every *non-overlapping* $M_{\text{L}} \times N_{\text{L}}$ window. The window dimensions M_{L} and N_{L} are the next smaller odd number of the theoretical spot distance S_y and S_x:

$$M_{\text{L}} = \begin{cases} \circ(S_y) - 2 & \text{if } \circ\,(S_y) = 2k+1 \\ \circ(S_y) - 1 & \text{otherwise} \end{cases} \tag{2.20}$$

and

$$N_{\text{L}} = \begin{cases} \circ(S_x) - 2 & \text{if } \circ\,(S_x) = 2k+1 \\ \circ(S_x) - 1 & \text{otherwise} \end{cases} \tag{2.21}$$

for $k \in \mathbb{N}$. This window size avoids that two spot locations fall into one search window and therefore avoids the cancellation of potential spot locations. Figure 2.5 illustrates this step. The black pixels indicate the maxima in the $M_{\text{L}} \times N_{\text{L}}$ windows. Note the two close maxima in the second spot row.

Pass 2 For every detected maximum location $[x\ y]^T \in \mathcal{L}$, select the location $[x'\ y']^T$ with the maximum response value in an $M_{\text{L}} \times N_{\text{L}}$ window around $[x\ y]^T$. If $[x'\ y']^T \neq [x\ y]^T$, remove

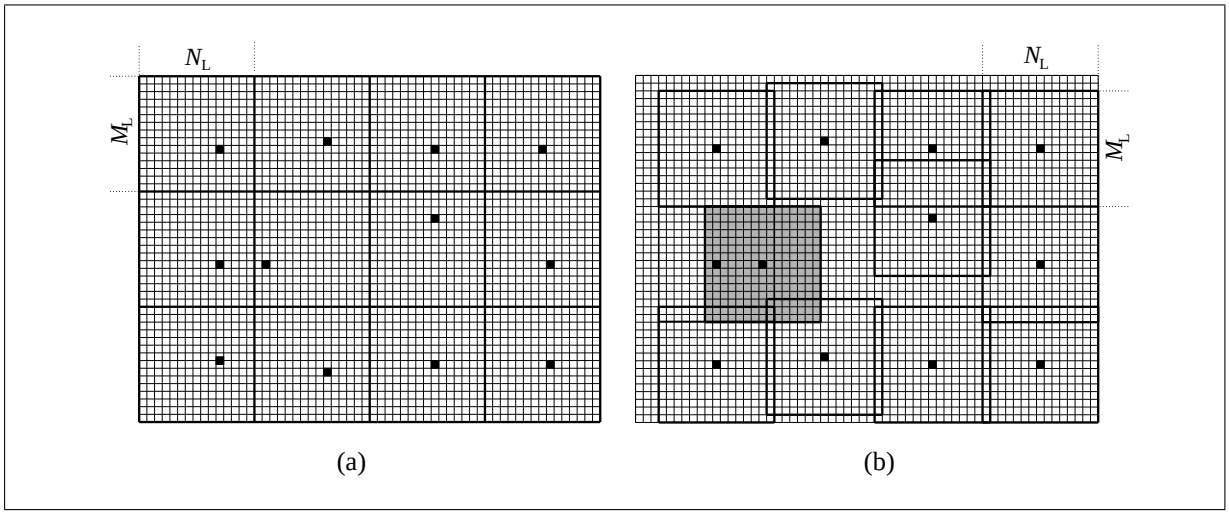
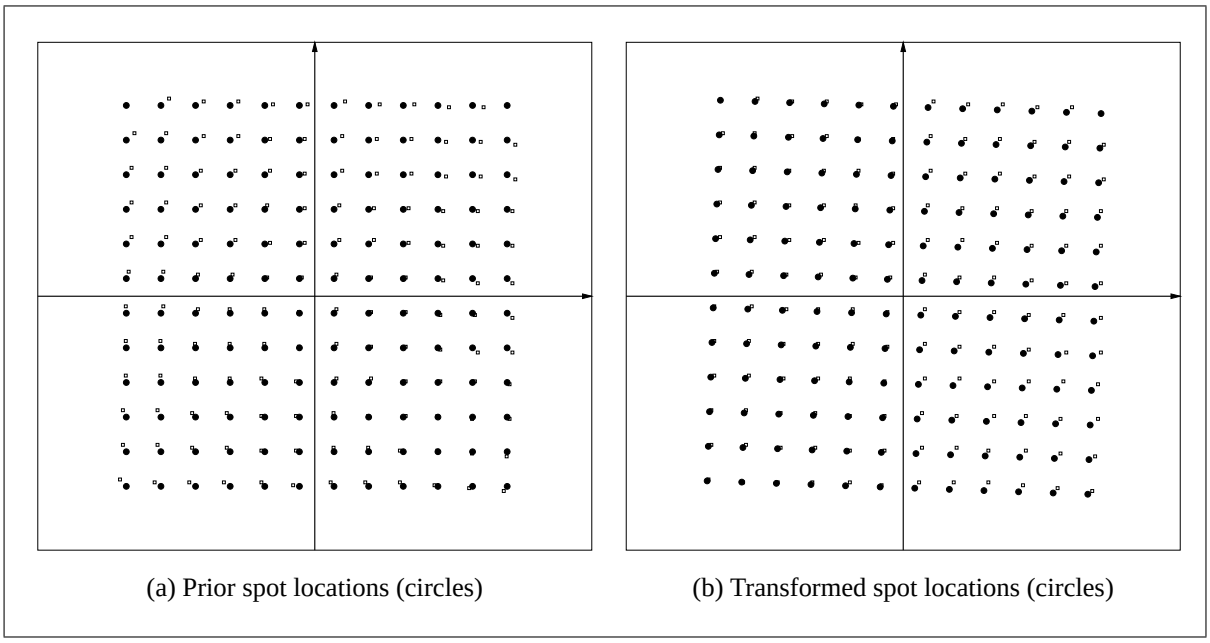


Figure 2.5. Maximum Search. Maxima are first searched in non-overlapping $M_{\text{L}} \times N_{\text{L}}$ windows which are smaller than the theoretical spot distance. Black pixels in (a) indicate local maxima. The maximum search is repeated in a second pass in order to get plausible spot distances. In (b) the lower maximum from the first pass will be canceled in the shaded area.



(a) Prior spot locations (circles) (b) Transformed spot locations (circles)

Figure 2.6. Alignment of detected spot locations: The black circles in (a) indicate the prior spot locations with theoretical spot distances, the squares indicate the detected guide spot locations. The prior spot locations are rigidly transformed such that they define a reference grid for the detected spot locations. A local search in the neighborhood of the reference locations then tries to map the detected spot locations to the correct grid nodes.

$[x\ y]^T$ from $\mathcal{L}$ and add $[x'\ y']^T$ to $\mathcal{L}$. After this pass it is guaranteed that two spot locations have plausible distances. In Figure 2.5b the grey-shaded area illustrates the plausible distance criterion. Assuming that in the second row the first local maximum value is higher than the second maximum value, the second maximum will be canceled.

Pass 3 Since prior knowledge about the expected number of spots is available, the size of the location set can be constrained to

$$\text{card}\,(\mathcal{L}) = I_{\text{G}} \cdot J_{\text{G}}. \tag{2.22}$$

As background regions will normally result in low MF response values and therefore low local maxima, it is natural to select the $I_{\text{G}} \cdot J_{\text{G}}$-highest local maxima to remain in $\mathcal{L}$. Clearly, noise will produce false positives and must be dealt with accordingly in the subsequent steps.

Transforming Prior Spot Locations

The potential spot locations which were found with the maximum search in the MF response image are only an unordered set $\mathcal{L}$ with no information about the node position on the spot grid. It is therefore necessary to map the detected locations to the correct nodes of the spot grid. This mapping is also denoted as the *alignment* of the detected spot locations. The idea to align the detected locations is to rigidly transform *prior spot locations* $\text{L}_{\text{P}}(\mathcal{G}, \mathbf{S})$ in a way that they

can serve as reference locations for the detected locations. Figure 2.6 outlines the idea behind this concept. The white squares indicate locations found by the maximum search. The black circles in Fig. 2.6a are prior spot locations with theoretical horizontal and vertical inter-spot distances S_x and S_y defined in (2.4). The absolute location of the prior guide spot grid within the spot array image does not matter; nevertheless, a position near the spot array can define a region of interest in order avoid unnecessary false positives. The black circles in Fig. 2.2b show rigidly transformed prior locations denoted as *reference locations* $\mathrm{L}_{\mathrm{R}}(\mathcal{G}, \mathbf{S})$. A search for a detected location $\mathrm{L}_{\mathrm{D}}((i,j))$ in a small neighborhood of $\mathrm{L}_{\mathrm{R}}((i,j))$ should then provide the correct location mapping.

A reference location $[x_{\mathrm{R}}\ y_{\mathrm{R}}]^T$ in the central coordinate system is computed from a prior spot location $[x_{\mathrm{P}}\ y_{\mathrm{P}}]^T$ as follows:

$$\begin{bmatrix} x_{\mathrm{R}} \\ y_{\mathrm{R}} \end{bmatrix} = \begin{bmatrix} \cos\theta_{\mathcal{G}} & \sin\theta_{\mathcal{G}} \\ -\sin\theta_{\mathcal{G}} & \cos\theta_{\mathcal{G}} \end{bmatrix} \begin{bmatrix} x_{\mathrm{P}} \\ y_{\mathrm{P}} \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}, \tag{2.23}$$

where $\theta_{\mathcal{G}}$ is an estimate of the global grid rotation angle according to Sect. 2.3 and $\mathbf{t} = [t_x\ t_y]^T$ is an estimate for the translation vector of the rotated prior spot locations. It is sufficient to determine the location of one grid node of the reference grid $\mathrm{L}_{\mathrm{R}}(\mathcal{G})$: If, for example, the translation $\mathbf{t}_{\mathrm{UL}}$ to the reference location $\mathrm{L}_{\mathrm{R}}((1,1))$ of the upper left node of $\mathcal{G}$ is known, the shift for all the other grid nodes is the same. Formally, the translation vector $\mathbf{t}_{\mathrm{UL}}$ is given by

$$\mathbf{t}_{\mathrm{UL}} = \widehat{\mathrm{L}}_{\mathrm{D}}((1,1)) - \mathrm{L}_{\theta}((1,1)). \tag{2.24}$$

The rotated prior spot location $\mathrm{L}_{\theta}((1,1))$ of the upper left node of $\mathcal{G}$ is subtracted from an estimate for the detected spot location of the upper left corner node of $\mathcal{G}$. Note that the upper left location $\mathrm{L}_{\mathrm{D}}((1,1))$ is not yet available, since the determination of the grid of detected spot locations $\mathrm{L}_{\mathrm{D}}(\mathcal{G})$ is the task of the grid fitting itself. We must therefore determine an estimate $\widehat{\mathrm{L}}_{\mathrm{D}}((1,1))$ for the location $\mathrm{L}_{\mathrm{D}}((1,1))$ of the upper left corner node from the set $\mathcal{L}$ of detected spot locations. We also estimate the location of the lower right corner node $(I_{\mathrm{G}}, J_{\mathrm{G}})$ of $\mathcal{G}$ and define the translation vector $\mathbf{t}_{\mathrm{LR}}$ as

$$\mathbf{t}_{\mathrm{LR}} = \widehat{\mathrm{L}}_{\mathrm{D}}((I_{\mathrm{G}}, J_{\mathrm{G}})) - \mathrm{L}_{\theta}((I_{\mathrm{G}}, J_{\mathrm{G}})). \tag{2.25}$$

The translation vector $\mathbf{t}$ is then determined as

$$\mathbf{t} = \mathrm{mean}(\mathbf{t}_{\mathrm{UL}}, \mathbf{t}_{\mathrm{LR}}). \tag{2.26}$$

The location estimate $\widehat{\mathrm{L}}_{\mathrm{D}}((1,1))$ can be computed in the following way: try to extract from the set of detected spot locations the locations $\mathrm{L}_{\mathrm{D}}(\mathrm{R}_{\mathcal{G}}(1))$ belonging to the first row and the locations $\mathrm{L}_{\mathrm{D}}(\mathrm{C}_{\mathcal{G}}(1))$ belonging to the first column of $\mathcal{G}$. $\widehat{\mathrm{L}}_{\mathrm{D}}((1,1))$ is then the intersection point of the straight lines fitted to the locations of the first row and first column, respectively. Similarly, extract the locations $\mathrm{L}_{\mathrm{D}}(\mathrm{R}_{\mathcal{G}}(I_{\mathrm{G}}))$ belonging to the last row and the locations $\mathrm{L}_{\mathrm{D}}(\mathrm{C}_{\mathcal{G}}(J_{\mathrm{G}}))$ belonging to the last column of $\mathcal{G}$ and intersect the corresponding fitted straight lines in order to estimate the location of the lower right grid node of $\mathcal{G}$.

Extracting the First Spot Row. An algorithm to extract the locations $\mathrm{L}_{\mathrm{D}}(\mathrm{R}_{\mathcal{G}}(1))$ belonging to the first row of $\mathcal{G}$ works as follows (see also Fig. 2.7):

1. Select as initial location $(x_0, y_0) \in \mathcal{L}$ the detected spot location with the highest y-coordinate in the central coordinate system.

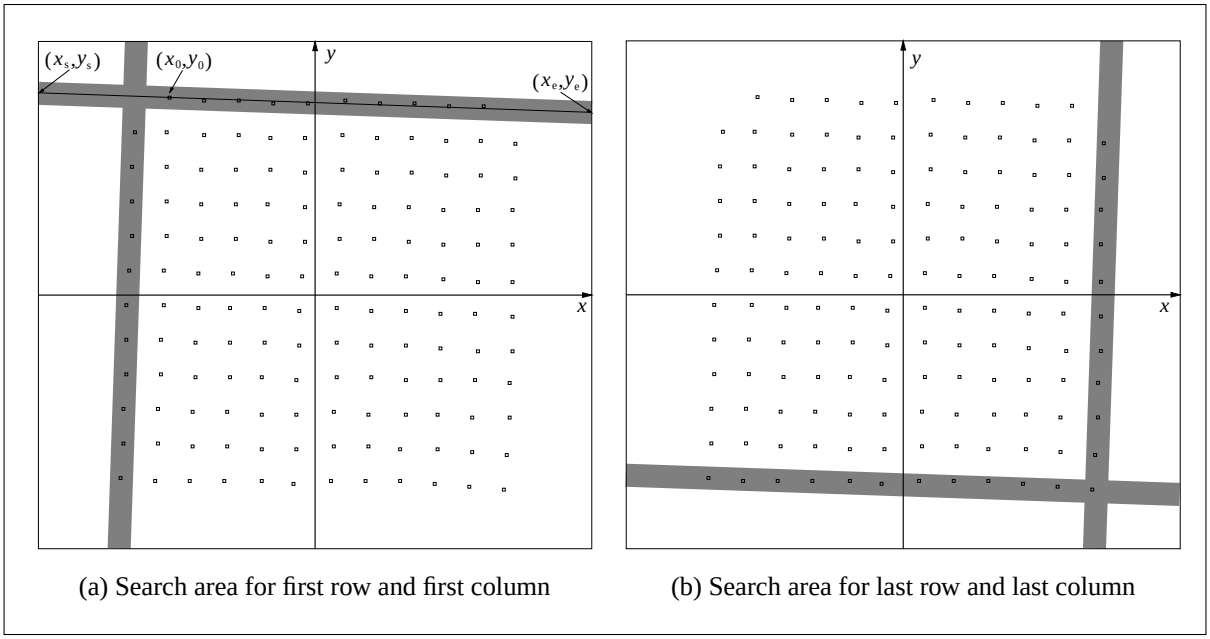
2. Determine the end points (x_s, y_s) and (x_e, y_e) of a digital straight line crossing the initial location (x_0, y_0). The straight line has a slope corresponding to the estimated global grid rotation angle $\theta_\mathcal{G}$ (Sect. 2.3). The straight line reaches from the first pixel column of the $M \times N$ spot image $\mathbf{S}$ to the last pixel column of the spot image.

3. Define an area $\mathcal{A}$ in which to look for the detected locations belonging to the first row. The search area $\mathcal{A}$ includes locations of the digital straight line between the end points (x_s, y_s) and (x_e, y_e). These locations belonging to the digital straight line are determined by a modified Bresenham algorithm [Bresenham, 1965], [Watt, 1994]. Add also all locations to the search area $\mathcal{A}$ belonging to straight lines between $(x_s, y_s \pm \tau)$ and $(x_e, y_e \pm \tau)$, where τ is a pixel tolerance parameter with $\tau \in \{1 \ldots \circ (S_y/2)\}$. The height of the search area therefore corresponds to the theoretical spot distance.

4. The locations belonging to the first grid row of $\mathcal{G}$ correspond to the intersection of the set of detected spot locations $\mathrm{L}_\mathrm{D}(\mathcal{G})$ with the locations of the search area $\mathcal{A}$:

$$\mathrm{L}_\mathrm{D}(\mathrm{R}_\mathcal{G}(1)) = \mathcal{A} \cap \mathcal{L}. \tag{2.27}$$

   The location set is only accepted if the number of locations in the intersection (2.27) is at



(a) Search area for first row and first column

(b) Search area for last row and last column

Figure 2.7. Determination of the points belonging to the first/last row/column. A search area $\mathcal{A}$ (one of the four shaded bars) is spanned by a straight line with a slope corresponding to the estimated grid rotation angle going through the point with the smallest/biggest y/x-coordinate. The width of the search area corresponds to the theoretical spot distance.

least 50 % of the spot grid width J_{G}. We therefore have the plausibility condition

$$\text{card}\,(\mathrm{L_D}(\mathrm{R}_{\mathcal{G}}(1))) \geq J_{\mathrm{G}}/2. \tag{2.28}$$

Condition (2.28) is necessary to cope with outliers. If inequality (2.28) does not hold, the extraction restarts with step 1. Before, all the locations previously assigned to $\mathrm{L_D}(\mathrm{R}_{\mathcal{G}}(1))$ are removed from the set of detected spot locations $\mathrm{L_D}(\mathcal{G})$.

The locations $\mathrm{L_D}(\mathrm{R}_{\mathcal{G}}(I_{\mathrm{G}}))$ belonging to the nodes of the last row are determined in the same way except that the initial point (x_0, y_0) is chosen as the point with the smallest y-coordinate in the central coordinate system. Likewise, the locations belonging to the nodes of the first and last column $\mathrm{L_D}(\mathrm{C}_{\mathcal{G}}(1))$ and $\mathrm{L_D}(\mathrm{C}_{\mathcal{G}}(J_{\mathrm{G}}))$ are determined in a similar way, except that the digital straight lines reach from the first row to the last row of the spot image.

Robust Straight Line Fitting. Once the R locations $(x_k, y_k) \in \mathrm{L_D}(\mathrm{R}_{\mathcal{G}}(1))$, $k \in \{1 \dots R\}$ belonging to the first row are determined, it is possible to fit to the location data the parameters a_r and b_r of a straight line model

$$y(x) = y(x; a_r, b_r) = a_r + b_r x. \tag{2.29}$$

Robust fitting is performed with the following algorithm:

1. **Initialization.** Mark all the locations of the first row as valid. Fit the parameters a_r and b_r in the standard least squares sense , i.e. minimize the sum of the squares of the residuals e_k between the location data points (x_k, y_k) and the model:

$$\sum_{k=1}^{R} {e_k}^2 = \sum_{k=1}^{R} (y_k - a_r - b_r x_k)^2 \rightarrow \text{min}. \tag{2.30}$$

The optimal solution is given by [Press et al., 1992, Cherkassky and Mulier, 1998]

$$a_r = \mu_y - b_r \mu_x \qquad \text{and} \qquad b_r = \frac{\sigma_{xy}}{\sigma_x^2}. \tag{2.31}$$

The mean values μ_x and μ_y are computed as

$$\mu_x = \frac{1}{R} \sum_{k=0}^{R} x_k \qquad \text{and} \qquad \mu_y = \frac{1}{R} \sum_{i=k}^{R} y_k, \tag{2.32}$$

with $(x_i, y_i) \in \mathrm{L_D}(\mathrm{R}_{\mathcal{G}}(I_{\mathrm{G}}))$. The variance σ_x^2 of the x-coordinates is given by

$$\sigma_x^2 = \frac{1}{R-1} \sum_{k=0}^{R} (x_k - \mu_x)^2, \tag{2.33}$$

and the covariance σ_{xy} between the x-coordinates and the y-coordinates is given by

$$\sigma_{xy} = \frac{1}{R} \sum_{k=0}^{R} (x_k - \mu_x)(y_k - \mu_y). \tag{2.34}$$

2. **Large residual removal.** Sort the residuals e_k of the initial fit and mark the nodes with the ρ highest residuals as invalid, where ρ is a percentage P_1 of the valid spots.

3. **Refitting.** Re-fit the parameters a_r and b_r if still at least 2 nodes in the row are valid; if fewer than 2 nodes are valid, go to 5.

4. **Outlier removal.** Sort the residuals e_k. If the largest residual is below a threshold t [pixel], the fitting is finished and the parameters are valid; otherwise mark the corresponding node as invalid. If more than 50 % of the initial number of valid nodes K_0 are valid, go to 3. Else go to 5.

5. **Algorithm abortion.** Mark the *parameters* as invalid and return.

A reasonable choice is to set $P_1 = 10$, meaning that the straight lines are re-fitted at least once with 10 % fewer nodes than at the first fit (the 10 % with the highest residuals). After the re-fit, the absolute values of the residuals are regarded. The parameters are re-fitted

- if the distance between a model point and a data point exceeds a threshold t [pixels] and
- if not more than 50 % of the initial valid nodes have already been marked as invalid.

The threshold t is a percentage P_2 of the theoretical spot distance S_x and S_y, respectively. It is reasonable to also set $P_2 = 10$. The 50 % limit is the *breakdown point* of this fitting procedure indicating the limit to which the percentage of outliers can increase which the estimator still can tolerate.

The parameters for the straight line of the last row are estimated in the same way. As for the parameter computation of the straight line of the first and last column, the x- and y-coordinates are swapped. This is necessary because an unrotated grid would result in an infinite slope. A straight line of the form $y' = a_c + b_c x'$ in the swapped x', y'-coordinate system has the form $y = -a_c/b_c - 1/b_c x$ in the original coordinate system (provided that $b_c \neq 0$).

Intersecting straight lines. After having estimated the straight line parameters a_r and b_r for the first row and a_c and b_c for the first column, the intersection point can be determined with the following equation:

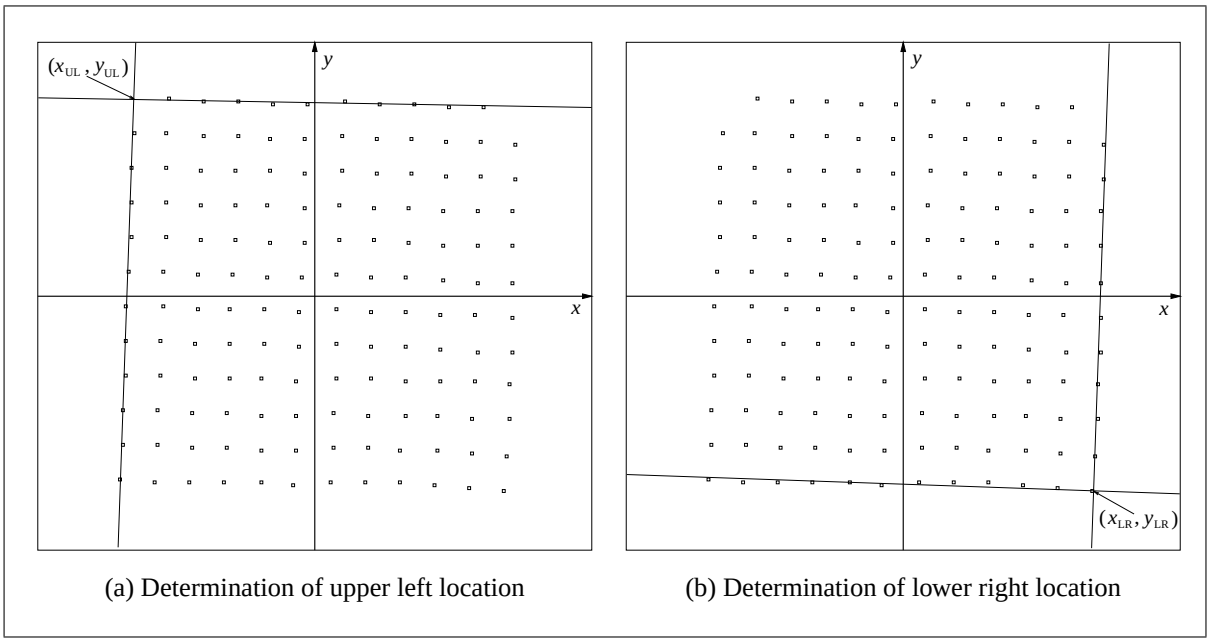
$$a_r + b_r x = -\frac{a_c}{b_c} - \frac{1}{b_c} x \qquad \text{for } b_c \neq 0. \tag{2.35}$$

Note that the right hand side of the equation is the straight line of the first column transformed to the coordinate system of the straight line of the first row. After some manipulation we have the following coordinates for the intersection point $(x_{\text{UL}}, y_{\text{UL}}) = \widehat{\text{L}}_{\text{D}}((1,1))$ of the upper left corner:

$$x_{\text{UL}} = \begin{cases} \left(-a_r - \dfrac{a_c}{b_c}\right) \Big/ \left(b_r - \dfrac{1}{b_c}\right) & \text{for} \quad b_c \neq 0 \\ a_c & \text{for} \quad b_c = 0 \end{cases} \tag{2.36}$$

$$y_{\text{UL}} = a_r + b_r x_{\text{UL}}. \tag{2.37}$$

The intersection point $(x_{\text{LR}}, y_{\text{LR}}) = \widehat{\text{L}}_{\text{D}}((I_{\text{G}}, J_{\text{G}}))$ for the lower right corner is computed in a similar manner to (2.36) and (2.37). The intersection points $(x_{\text{UL}}, y_{\text{UL}})$ and $(x_{\text{LR}}, y_{\text{LR}})$ are illustrated in Fig. 2.8. The final translation vector $\mathbf{t}$ is the mean vector according to (2.26).



(a) Determination of upper left location

(b) Determination of lower right location

Figure 2.8. Determination of the location of the upper left grid node and the location of the lower right grid node of the spot grid. The locations are determined by the intersection of straight line models whose parameters are fitted to the locations belonging to the first/last row/column.

Alignment of locations to grid nodes

Having an estimate of the rotation angle $\theta_{\mathcal{G}}$ and the translation vector $\mathbf{t} = [t_x\ t_y]^T$, it is possible to rigidly transform the prior spot locations $[x_{\text{P}}\ y_{\text{P}}]^T \in \text{L}_{\text{P}}(\mathcal{G})$ to the reference spot locations $[x_{\text{R}}\ y_{\text{R}}]^T \in \text{L}_{\text{R}}(\mathcal{G})$ according to (2.23). The reference spot locations $\text{L}_{\text{R}}(i,j)$ are expected to be near the detected spot locations $\mathcal{L}$. The detected spot locations can therefore be assigned to spot grid nodes (i,j) by investigating a rectangular $M_{\text{L}} \times N_{\text{L}}$ location window $\text{W}\left(\text{L}_{\text{R}}(i,j)\right)$ with $\text{L}_{\text{R}}(i,j)$ in the window center:

$$\text{L}_{\text{D}}(i,j) = \text{W}\left(\text{L}_{\text{R}}(i,j)\right) \cap \mathcal{L}. \tag{2.38}$$

The dimensions M_{L} and N_{L} of W correspond to the size of maximum search window as defined in (2.20) and (2.21). The grid spanning with prior spot locations works reliable if the spot array in the image is not too distorted. If, however, the spot array is stretched in one or two directions, the transformed prior locations will not cover the whole array and some spots will never be assigned to grid nodes. The following section describes an alternative grid spanning which is able to cope with large distortions in the spot array.

2.4.2 Algorithm II: Initial Grid before Maximum Search

This section describes an alternative way which first constructs an initial spot grid and then performs a local maximum search in the MF response image. The main idea of this approach is based on the concept of the inverse Radon transform: Every point of a projection $\text{P}(\mathbf{R}^{\text{M}}, \theta)$

corresponds to a straight line with orientation θ in the image $\mathbf{R}^{\mathrm{M}}$. Approximate spot locations are expected at *intersections* of straight lines going through the spot grid rows and spot grid columns. Straight lines going through spot grid columns correspond to maximum values in the projection $\mathrm{P_C} = \mathrm{P}(\mathbf{R}^{\mathrm{M}}, \theta_{\mathcal{G}_{\mathrm{C}}})$ (see Fig. 2.4b). In order to search straight lines going through the spot grid *rows* it is necessary to compute a second projection $\mathrm{P_R} = \mathrm{P}(\mathbf{R}^{\mathrm{M}}, \theta_{\mathcal{G}_{\mathrm{R}}})$, with $\theta_{\mathcal{G}_{\mathrm{R}}} \approx \theta_{\mathcal{G}_{\mathrm{C}}} + \pi/2$. It is a good idea to determine $\theta_{\mathcal{G}_{\mathrm{R}}}$ with a second rotation estimation based on projections through the grid rows. The Radon transform for this second rotation estimation can be efficiently computed in a small neighborhood of $\theta_{\mathcal{G}_{\mathrm{C}}} + \pi/2$. The straight lines for the back-projection are represented in polar coordinates

$$x \cos\theta + y \cos\theta = p, \tag{2.39}$$

with θ as the angle between the x-axis and the straight line normal and p as the position on the projection. Two straight lines with parameters $(\theta_{\mathcal{G}_{\mathrm{C}}}, p_i \in \mathrm{P_C})$ and $(\theta_{\mathcal{G}_{\mathrm{R}}}, p_j \in \mathrm{P_R})$ intersect at the point $\chi(\theta_{\mathcal{G}_{\mathrm{C}}}, p_i, \theta_{\mathcal{G}_{\mathrm{R}}}, p_j) = (x_\chi\ y_\chi)$, where

$$\begin{pmatrix} \cos\theta_{\mathcal{G}_{\mathrm{C}}} & \sin\theta_{\mathcal{G}_{\mathrm{C}}} \\ \cos\theta_{\mathcal{G}_{\mathrm{R}}} & \sin\theta_{\mathcal{G}_{\mathrm{R}}} \end{pmatrix} \begin{pmatrix} x_\chi \\ y_\chi \end{pmatrix} = \begin{pmatrix} p_i \\ p_j \end{pmatrix}. \tag{2.40}$$

Straight lines going through spot grid rows and columns will give rise to local maxima in the corresponding projections. On the other hand, not all local maxima in the projections will correspond to spot array rows or columns. Local maxima in the projections are therefore good initial *hypotheses* for the locations of the spot array rows and columns. False hypotheses are removed by considering small rectangular areas $\mathrm{W}((x_\chi, y_\chi))$ around the intersection points of the back-projected straight lines. The small area W is used to be more robust in order to cope with non-linearities of the spot grid. The dimensions of W depend on the image resolution and the spot size and density. We formulate the removal of false straight line hypotheses as the following optimization problem:

$$\sum_{p_i \in \mathrm{P_C}, p_j \in \mathrm{P_R}} \mathrm{M}_1\left[\mathrm{W}(\chi(p_i, p_j)\right] \rightarrow \max, \tag{2.41}$$

where $\mathrm{M}_1[.]$ is a brightness measure (typically the intensity mean) of the intersection window W. The idea behind the maximization problem (2.41) is that the intersections of a false column hypothesis $p_i \in \mathrm{P_C}$ with the row hypotheses $p_j \in \mathrm{P_R}$ will be no spot locations and therefore will have low response values in the spot amplified image. We therefore want to enforce hypotheses which have high brightness measures M_1 at the intersection points. Fig. 2.9 shows a visualization of the intersection brightness measures M_1 for 72 row hypotheses and 72 column hypotheses of a 48×48 spot grid. The optimization problem (2.41) is constrained in the sense that two neighboring straight lines must have a minimum distance. Formally, we define for all $p_m \in \mathrm{P_C}, \mathrm{P_R}$ the neighborhood operator $\mathrm{N}_2(p_m) = p_n$ and have the constraint

$$|p_m - p_n| - S/2 \geq 0, \tag{2.42}$$

where S is the theoretical spot size [pixels] which is expected from the imaging parameters.

The constrained optimization problem (2.41) could be solved with the help of Lagrange-Multipliers [Bertsekas, 1999]. In practice, however, the search space for the general optimization problem will be too large. The false hypotheses removal can be implemented as a greedy

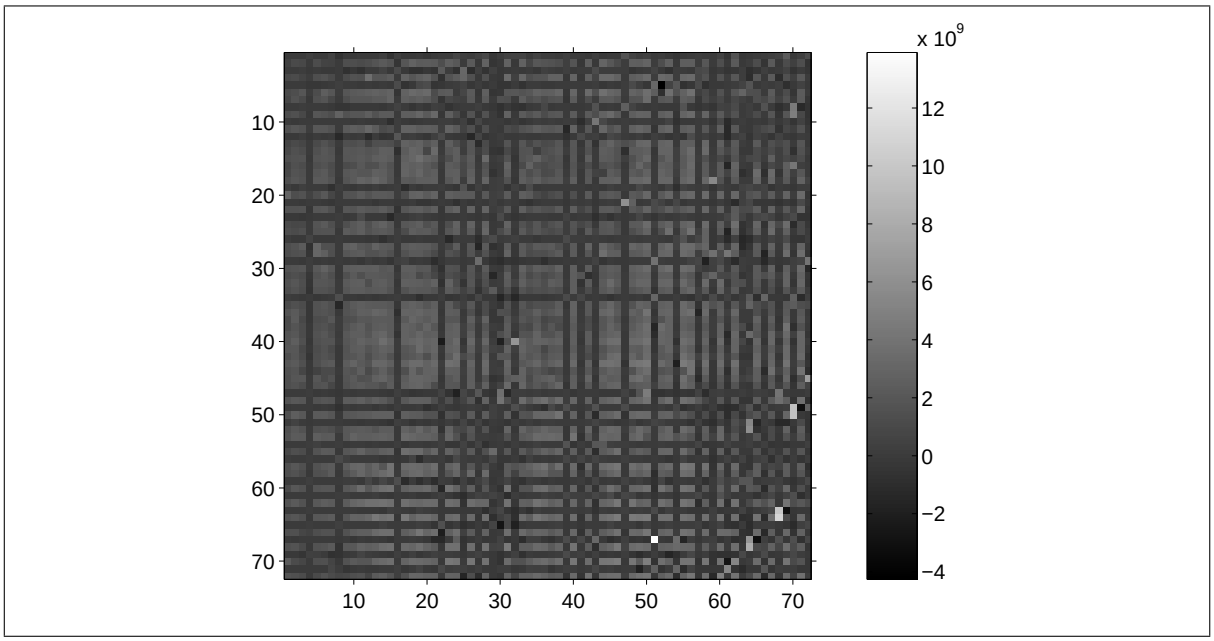


Figure 2.9. Spot location hypotheses. Intersection brightness of 72 column hypotheses with 72 row hypotheses of a 48×48 spot grid. Wrong hypotheses a characterized as dark rows or columns.

search algorithm, where the J_{G} correct column hypotheses of $\mathrm{P_C}$ are selected in a single step and the I_{G} correct row hypotheses of $\mathrm{P_R}$ are selected in a second step. The column hypotheses $p_i \in \mathrm{P_C}$ are indexed with regard to a weight M_2, which is a measure of the intersection intensities of p_i with all row hypotheses $p_j \in \mathrm{P_R}$. The rows and columns of the false hypotheses in Fig. 2.9 mostly have significantly lower brightness measures than rows and columns of true hypotheses. Nevertheless, some intersection points of false hypotheses have a very bright intersection measure M_1. The weight measure M_2 must therefore be determined in a robust manner. The weight M_2 for a hypothesis $p_i \in \mathrm{P_C}$ is chosen as the $P\%$ percentile of the intersection intensities $\mathrm{M}_1(\mathrm{W}(\chi(p_i, p_j))$ with the hypotheses $p_j \in \mathrm{P_R}$. The percentage P is given as $P = 50 + b((H - J_{\mathrm{G}})/J_{\mathrm{G}}) * 100$, where H is the total number of hypotheses and b is fraction of the false hypotheses that is considered. If, for example $b = 0.8$, we do not consider the 20% brightest intersections.

After the false hypotheses have been removed, the spot locations are refined by looking for maximum values in small windows of the MF response image $\mathbf{R}^{\mathrm{M}}$.

2.4.3 Spot Grid Parameterization

The algorithms for the determination of the initial spot grid in Sect. 2.4.1 and 2.4.2 try to minimize the number of false positives and false negatives. Nevertheless, after the initialization it is not guaranteed that all grid nodes are associated with the correct spot locations. One possible approach to get a consistent spot grid is to robustly fit straight lines to every row and column of the spot grid $\mathcal{G}$ and estimate a lacking location by the intersection of the straight line of row i

and and the straight line of column j. We define a parameter set $\mathcal{P}$ for a grid $\mathcal{G}$ as

$$\mathcal{P} = \{((a_{r_i}, b_{r_i}), (a_{c_j}, b_{c_j})) \mid 1 \leq i \leq I_{\mathrm{G}}, 1 \leq j \leq J_{\mathrm{G}}\}, \tag{2.43}$$

with a_{r_i} and b_{r_i} as the parameters of a row straight line model

$$y(x) = y(x; a_{r_i}, b_{r_i}) = a_{r_i} + b_{r_i} x \tag{2.44}$$

and a_{c_i} and b_{c_i} as the parameters of a column straight line model (2.44).

2.4.4 Robust Straight Line Fitting

The algorithm for the fitting of a straight line belonging to row i corresponds to the straight line fitting for the first grid row already defined in Sect. 2.4.1. The main points are repeated here.

1. **Initialization.** Fit the parameters a_{r_i} and b_{r_i} according to (2.31) if $K_0 \geq 2$, where K_0 is the number of valid nodes in row i. If $K_0 < 2$ go to 5.
2. **Large residual removal.** Sort the residuals e_k of the initial fit and mark the nodes with the ρ highest residuals as invalid, where ρ is a percentage P_1 of the valid spots K_0.
3. **Refitting.** Re-fit the parameters a_{r_i} and b_{r_i} if still at least 2 nodes in the row i are valid; if fewer than 2 nodes are valid, go to 5.
4. **Outlier removal.** Sort the residuals e_k. If the largest residual is below a threshold t [pixel], the fitting is finished and the parameters are valid; otherwise mark the corresponding node as invalid. If more than 50 % of the initial number of valid nodes K_0 are valid, go to 3. Else go to 5.
5. **Algorithm abortion.** Mark the *parameters* as invalid and return.

2.4.5 Grid Parameter Correction

After fitting of the grid rows and columns, the parameters are checked:

1. a_{r_i} and b_{r_i} might be marked as invalid because of the failure of the straight line fitting algorithm
2. the absolute value of the slope b_{r_i} might differ too much ($0.5°$) from the absolute mean value of all the valid slopes of the grid rows. This situation may occur since it is not guaranteed that the node with the highest residual value must necessarily be the outlier. Sometimes it is therefore possible that - even after the repeated refitting - false positives do survive.

If 1) or 2) holds for the parameters a_{r_i} and b_{r_i} of a row, they are estimated with the help of the parameters of the nearest neighborhood row a_{r_n} and b_{r_n}. Since we do not expect large variation between neighbored slopes we can set $b_{r_i} = b_{r_n}$. The intercept a_{r_i} can be estimated by setting

$$a_{r_i} = a_{r_n} - (n - i)S_y, \tag{2.45}$$

i.e. subtracting the theoretical spot distances from the neighbor depending on how far away the neighbor on the grid is.

2.4.6 Abortion Criterion

After the robust determination of the field parameter sets (2.43), the entire spot grid $\mathcal{G}$ is tested for consistency. If – due to a very bad image quality – the final spot grid is not plausible, the distance between the locations of at least two nodes must be too large. Formally, there must exist at least one node $(i, j) \in \mathcal{G}$ for which the following condition holds:

$$||\mathrm{L}((i,j)) - \mathrm{L}((i+1, j+1))|| > t \tag{2.46}$$

where t is the residual threshold entity of Sect. 2.4.4. If (2.46) holds, the grid fitting is aborted and the user is notified.

2.5 Background Estimation

The general framework for spot array image analysis assumes nonuniform, smoothly varying background values. In order to obtain contiuously varying background values for the entire image, the background is estimated in a global manner. In principle, the ideal output of the background estimation would be a *background image* $\mathbf{B}$ comprising the spot array image $\mathbf{S}$ without the spots. However, initially after grid fitting the spots are not yet characterized. In order to tackle this chicken-egg problem, the background is estimated in several (at least two) passes. Firstly, the grid fitting procedure yields information about the spot locations. If one assumes additive noise, one could subtract the pixels belonging to the spots from the spot array image. In order to get smooth results, a hierarchical interpolation method basen on Gaussian image pyramids [Jolion and Rosenfeld, 1994] is used. An image pyramid combines the advantages

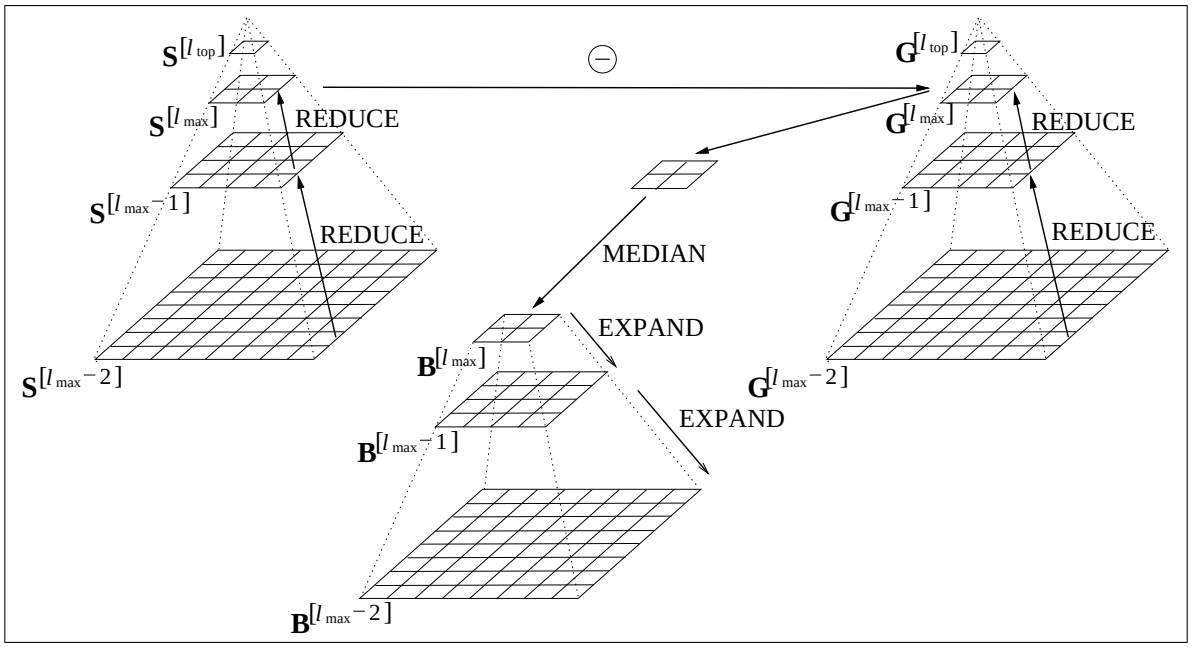


Figure 2.10: Background estimation.

of high and low resolution. It is a collection of images $\mathbf{S}^{[l]}$ of a single scene at exponentially decreasing resolutions $l \in \{0 \dots l_{\text{top}}\}$. The bottom level of the pyramid is the original image. In the simplest case, each successive level of the pyramid is obtained from the previous level by a filtering operation followed by a sampling operator [Haralick et al., 1991]. The image size of the pyramid levels and the grey values are determined by the REDUCE function: In our case the image size decreases by a factor of 4 with every pyramid level. We furthermore use a 5×5 *reduction window* indicating that for the intensity of a pixel in $\mathbf{S}^{[l+1]}$ the mean of a 5×5 window in $\mathbf{S}^{[l]}$ with Gaussian weighting is computed. Figure 2.10 plots the principle of the background estimation with image pyramids.

Pass 1 This pass is performed after grid fitting and before the spot characterization. A pyramid $\mathbf{S}^{[l]}$ of the spot image and a pyramid $\mathbf{G}^{[l]}$ of the synthetic spot image are built. The resolution of the spot image decreases with increasing pyramid levels l. At a certain level l_{max} the resolution of the image is so low such that the spot grid structure is no longer present in the image, meaning that the spots are merged. The merging level l_{max} is chosen as the pyramid level where the equation

$$\min({S_x}^l, {S_y}^l) < 0.5 \tag{2.47}$$

holds (${S_x}^l$ is the theoretical spot width and ${S_y}^l$ is the theoretical spot height for pyramid level l). At the merging level l_{max} we subtract $\mathbf{G}^{[l_{\text{max}}]}$ from $\mathbf{S}^{[l_{\text{max}}]}$ resulting in a low-resolution background image $\mathbf{B}^{[l_{\text{max}}]}$. In order to get a background image at the original resolution of the spot image, the levels of the background pyramid are computed by the EXPAND function, which consists of bicubic interpolation of the grey values [Press et al., 1992].

Pass 2 The background is estimated a second time after the spot fitting procedure has finished. After spot characterization it is possible for every spot to reconstruct a complete synthetic spot image. The pyramid $\mathbf{G}^{[l_{\text{max}}]}$ in Fig. 2.10 is now the reconstructed synthetic spot image. The subtraction and the expansion are performed the same way as in pass 1.
Supplementary passes will further increase accuracy: A more accurate background estimation will increase spot fitting accuracy, leading itself to an improved background computation etc. The trade-off between accurarcy and computation time will depend on the application.

2.6 Parametric Spot Fitting

A parametric fit on a set of intensities (pixels) belonging to a spot assumes a given analytic model the unknown parameters of which have to be determined. The approximate initial locations of the spots are given by the grid fitting procedure. The extension of the pixel set belonging to a spot is determined by the prior knowledge about the theoretical spot size and equals the $M_{\text{MF}} \times N_{\text{MF}}$ matched filter. Let $\mathcal{S}_{ij} = \{(\mathbf{p}_k, z_k),\ \mathbf{p}_k \in \mathbb{R}^2,\ z_k \in \mathbb{R}\}$ be a set of N points (pixels) corresponding to a spot grid node $g_{ij} \in \mathcal{G}$, where z_k denotes the intensity pixel intensity at location $\mathbf{p}_k$. A parametric model Z_{ij} for a spot g_{ij} with location $\mathbf{p}$ predicts a functional relationship between the measured independent and dependent variables,

$$\mathrm{Z}_{ij}(\mathbf{p}) = \mathrm{Z}_{ij}(\mathbf{p}, \mathbf{q}) \tag{2.48}$$

with the adjustable parameter vector $\mathbf{q} \in \mathbb{R}^D$. Given a particular data set $\mathcal{S}_{ij}$, the question in data modeling is: "Given a particular parameter vector $\mathbf{q}$, what is the probability that the data set $\mathcal{S}_{ij}$ could have occurred?" If the z_k take on continuous values, the probability will always be zero unless the phrase "...plus or minus some fixed Δy on each data point" [Press et al., 1992]. If the probability of obtaining the data set is infinitesimally small, then one can conclude that the parameters under consideration are unlikely to be right. On the other hand, the data set should not be improbable for the correct choice of parameters.

In order to deal with overlapping spots, one can recompute the model Z_{ij} by using the modified spot patch

$$\mathcal{S}^*_{ij} = \mathcal{S}_{ij} - \sum_{k,l\in\{-1,0,1\},(k,l)\neq(0,0)} \mathrm{Z}_{i+k\ j+l} \tag{2.49}$$

i.e. subtracting neighboring spot models. One can iterate this procedure for every spot g_{ij} over the whole image. The results are gradually better models for every spot, where the iteration is aborted when the parameters of the model for each spot stabilize.

2.6.1 Least Squares as a Maximum Likelihood (ML) Estimator

Maximum likelihood estimation is a form of parameter estimation which *maximizes* the likelihood defined in the above way. One can suppose that each data point z_k has a measurement error that is independently random and distributed as a normal (Gaussian) distribution around the "true" model $\mathrm{Z}_{ij}(\mathbf{p}_k, \mathbf{q})$. If it is furthermore supposed that the standard deviations σ of these normal distributions are the same for all points, then the probability of the data set is the product of the probabilities of each point:

$$P \propto \prod_{k=1}^{N} \left\{ \exp\left[-\frac{1}{2} \left(\frac{z_k - \mathrm{Z}_{ij}(\mathbf{p}_k, \mathbf{q})}{\sigma} \right)^2 \right] \Delta y \right\}. \tag{2.50}$$

Maximizing (2.50) is equivalent to maximizing its logarithm, or minimizing the negative of its logarithm:

$$\left[\sum_{k=1}^{N} \frac{(z_k - \mathrm{Z}_{ij}(\mathbf{p}_k, \mathbf{q}))^2}{2\sigma^2} \right] - N \log \Delta y \rightarrow \ \min. \tag{2.51}$$

Since N, σ, and Δy are all constants, minimizing this equation is equivalent to

$$\sum_{k=1}^{n} (z_k - \mathrm{Z}_{ij}(\mathbf{p}_k, \mathbf{q}))^2 \ \rightarrow \ \min \tag{2.52}$$

Minimizing the squared error between the model and the data is a maximum likelihood estimation of the fitted parameters *if* the measurement errors are independent and normally distributed with constant standard deviation. For real data, the normal distribution of the measurement errors is often rather poorly realized. For example, contaminations in the spot array image are regarded as outliers. Their probability of occurrence in the assumed Gaussian model is so small that the maximum likelihood estimator is willing to distort the whole model to try to bring them, mistakenly, into line.

Least Squares Spot Fitting of Noordmans/Smeulders

Noordmans and Smeulders [Noordmans and Smeulders, 1998] introduce a spot model Z with the parameters spot center $\boldsymbol{\mu}$, spot intensity a, size σ, orientation ϕ and local background level b. The parameters are summarized in the parameter vector $\mathbf{q} = [\boldsymbol{\mu}\ a\ \sigma\ \phi\ b]^T$. In the spot *detection* phase, they already match spot models with a range of parameter values to each image position. For each image position the best matching model is retained. The detection phase ends by selecting positions where the spot match is locally optimal. In the characterization phase, the remaining spots are one by one considered on the basis of the same model used to refine the parameter vector. In the following, we describe the core least squares fitting approach in their paper, where the notation has been adapted to be consistent with the symbols used in this work.

The error E between the intensity data z_k and the spot model $\mathrm{Z}(\mathbf{p}_k, \mathbf{q})$ is computed as

$$\mathrm{E} = \frac{\sum_{k=1}^{n} \left\{ [z_k - \mathrm{Z}(\mathbf{p}_k, \mathbf{q})]^2 \, \mathrm{W}(\mathbf{p}_k, \mathbf{w}) \right\}}{\sum_{k=1}^{n} \mathrm{W}(\mathbf{p}_k, \mathbf{w})} \ \rightarrow\ \min, \tag{2.53}$$

where the squared error (2.52) is weighted by a function $\mathrm{W}(\mathbf{p}_k, \mathbf{w})$ with a parameter vector $\mathbf{w}$: The weight function is maximal at the center of the spot and drops to zero at greater distances to reduce distortions caused by other proximate image structures.

The definition of the error E in (2.53) has the advantage that the minimizing for spot intensity a and local background b can be calculated analytically. To prove this, they explicitly write down the dependence of the model on a and b:

$$\mathrm{Z}(\mathbf{p}, \mathbf{q}) = a\mathrm{G}(\mathbf{p}, \mathbf{r}) + b, \qquad \mathbf{r} = [\sigma\ \phi]^T, \tag{2.54}$$

where $\mathrm{G}(\mathbf{p}, \mathbf{r})$ depends only on the remaining elements of the parameter vector. Recall that in the detection phase of the algorithm, a spot model is fitted at *every* pixel position. Omitting the need for subpixel accuracy, the spot center $\boldsymbol{\mu}$ is therefore not in parameter vector $\mathbf{r}$. Substitution into (2.53) gives

$$\mathrm{E} \ = \ \frac{1}{c_0}\{a^2 c_2 + b^2 c_0 + d_1 + 2abc_1 - 2ad_2 - 2bd_0\}$$

where c_i and d_i have been defined as

$$\begin{aligned} c_0 &= \sum_{k=1}^{n} \mathrm{W}(\mathbf{p}_k, \mathbf{w}) & d_0 &= \sum_{k=1}^{n} z_k \mathrm{W}(\mathbf{p}_k, \mathbf{w}) \\ c_1 &= \sum_{k=1}^{n} \mathrm{G}(\mathbf{p}, \mathbf{r}) \mathrm{W}(\mathbf{p}_k, \mathbf{w}) & d_1 &= \sum_{k=1}^{n} z_k^2 \mathrm{W}(\mathbf{p}_k, \mathbf{w}) \\ c_2 &= \sum_{k=1}^{n} \mathrm{G}(\mathbf{p}, \mathbf{r})^2 \mathrm{W}(\mathbf{p}_k, \mathbf{w}) & d_2 &= \sum_{k=1}^{n} z_k \mathrm{G}(\mathbf{p}, \mathbf{r}) \mathrm{W}(\mathbf{p}_k, \mathbf{w}) \end{aligned} \tag{2.55}$$

The optimal amplitude and background values are found by minimizing E with respect to the amplitude a and the background b

$$\frac{\partial \mathrm{E}}{\partial a} = 0 \Rightarrow ac_2 + bc_1 - d_2 = 0, \tag{2.56}$$

$$\frac{\partial \mathrm{E}}{\partial b} = 0 \Rightarrow bc_0 + ac_1 - d_0 = 0. \tag{2.57}$$

Solving (2.56) and (2.57) gives

$$a = \frac{d_0 c_1 - d_2 c_0}{{c_1}^2 - c_0 c_2} \tag{2.58}$$

$$b = \frac{d_2 c_1 - d_0 c_2}{{c_1}^2 - c_0 c_2} \tag{2.59}$$

From (2.56) and (2.57) it is also possible to derive the expressions for optimal a and b when one of these is constant

$$a = a_0 \Rightarrow b = \frac{d_0 - a_0 c_1}{c_0} \tag{2.60}$$

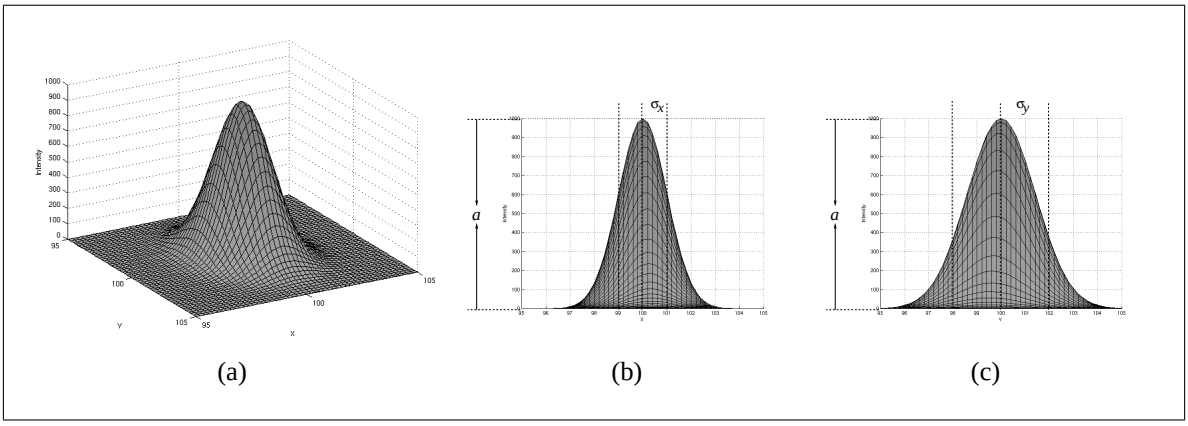
$$b = b_0 \Rightarrow a = \frac{d_2 - b_0 c_1}{c_2} \tag{2.61}$$

As for the remaining elements of the parameter vector $\mathbf{r}$, the minimizing will generally not be analytical.

ML-Estimators for Gaussian Spots

In this work, Gaussian Spots are used as a model for a broad range of spots. For Gaussian spot models, the term $\mathrm{G}(\mathbf{p}, \mathbf{r})$ in (2.54) is defined as

$$\mathrm{G}(\mathbf{p}, \mathbf{r}) = \mathrm{G}(\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\left[-\frac{1}{2}(\mathbf{p} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{p} - \boldsymbol{\mu})\right], \tag{2.62}$$



(a) (b) (c)

Figure 2.11: Gaussian Spot Model.

with $\boldsymbol{\mu}$ as the mean of the Gaussian model corresponding to the center of the spot and $\boldsymbol{\Sigma}$ as the 2×2 covariance (dispersion) matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}. \tag{2.63}$$

Figure 2.11 illustrates the shape of the Gaussian model. The definition of the general covariance matrix (2.63) allows elliptical spot expansion with any orientation. In the following, we show how the parameters of a Gaussian spot model can be fitted with ML-Estimators. For the general framework of spot array image analysis, the following differences to the Noordmans/Smeulders ML fitting described above are vital:

1. There is no need for a (further) spot detection phase – it has already been performed via the grid fitting described in the sections above. There is already information about locations of spot centers. However, the information might be only approximate. Furthermore, subpixel accuracy is required. It is therefore necessary to explicitly estimate the spot center $\boldsymbol{\mu}$.

2. Equations (2.59) and (2.61) show how to compute the background in an analytical matter. The additive background b competes with the amplitude a of the Gaussian, which is also computed analytically. This is because there are infinite possibilities to combine the two parameters to yield the observed intensity pixel. This may lead to heavily discontinuous background values. If a smooth background is a required feature, the preferred way is to estimate the background iteratively in a global manner as described in Sect. 2.5. With this global approach, the intensity data is first corrected with the estimated background value $\widehat{b}$ as follows:

   $$z_k := \max(\mathrm{I}(\mathbf{p}_k) - \widehat{b}, 0). \tag{2.64}$$

   Since the background can be overestimated, – especially in the first background estimation – negative values are corrected to zero.

The ML estimate $\widehat{\boldsymbol{\mu}}$ of the center $\boldsymbol{\mu}$ is computed as the *sample average* of the coordinates weighted by the corrected pixel intensities, which in turn can be weighted by a function W as in (2.53)

$$\widehat{\boldsymbol{\mu}} = \sum_{k=1}^{n} z_k \mathbf{p}_k \mathrm{W}(\mathbf{p}_k, \mathbf{w}) \Bigg/ \sum_{k=1}^{n} z_k \mathrm{W}(\mathbf{p}_k, \mathbf{w}) \,. \tag{2.65}$$

Similarly, the ML estimate $\widehat{\boldsymbol{\Sigma}}$ of the covariance matrix $\boldsymbol{\Sigma}$ is given by the sample average of the outer product $(\mathbf{p}_k - \widehat{\boldsymbol{\mu}})(\mathbf{p}_k - \widehat{\boldsymbol{\mu}})^T$ [Bishop, 1995] weighted by the pixel intensities and the weight function:

$$\widehat{\boldsymbol{\Sigma}} = \sum_{k=1}^{n} z_k (\mathbf{p}_k - \widehat{\boldsymbol{\mu}})(\mathbf{p}_k - \widehat{\boldsymbol{\mu}})^T \mathrm{W}(\mathbf{p}_k, \mathbf{w})^2 \Bigg/ \sum_{k=1}^{n} z_k \mathrm{W}(\mathbf{p}_k, \mathbf{w}) \,. \tag{2.66}$$

The estimate $\widehat{a}$ for the amplitude a is given by (2.60): It is assumed that the background $b(\mathbf{p}_k)$ is slowly varying over the image – for the characterization of the background of a spot it is sufficient to take one background sample $b(\boldsymbol{\mu})$ at the spot center $\boldsymbol{\mu}$. Hence one can assume

a constant b and therefore employ (2.61) rather than (2.59). Since the background has already been subtracted , one has $b_0 = 0$ and

$$\widehat{a} = \frac{d_2}{c_2} = \frac{\sum_{k=1}^{n} z_k \mathrm{G}(\mathbf{p}_k, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \mathrm{W}(\mathbf{p}_k, \mathbf{w})}{\sum_{k=1}^{n} \mathrm{G}(\mathbf{p}_k, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})^2 \mathrm{W}(\mathbf{p}_k, \mathbf{w})}. \tag{2.67}$$

As mentioned above, the weight function W could be maximal at the center of the spot and drop to zero at greater distances (for example a Gaussian). While this may reduce the influence of distortions at the spot border, it does not cope with distortions in regions near the spot center. The weight function of *robust estimators* is data-driven rather than topology-driven. In the next sections, robust estimators are introduced and adapted to our needs of Gaussian models.

2.6.2 Robust M-Estimators

Parametric spot fitting is expected to yield consistent estimates of the unknown parameters at the idealized model. If the fitting is robust, the parameters will not drift too far away if the model is only approximately true. The theory of robust statistics is well researched in literature. One quality measure of a robust estimator is the *breakdown point*. The breakdown point ϵ^* gives the limit to which the percentage of outliers can increase which the estimator still can tolerate. We give a short introduction to robust estimation with M-estimators. Other robust estimators include L-, R-, and MM-estimators and are discussed in [Jureckova, 1984] , [Jureckova and Sen, 1996] or [Huber, 1981]. In general, M-, L- and R-estimators are asymptotically equivalent under certain smoothness conditions [Jureckova, 1984]. However, M-estimators are chosen in this work because they are the most flexible ones and they generalize straightforwardly to multiparameter problems [Huber, 1981] needed to be solved in this work.

Let the difference $z_k - \mathrm{Z}(\mathbf{p}_k, \mathbf{q})$ in (2.53) between the k^{th} observation and its fitted value be denoted as the residual r_k. The ML estimator tries to minimize $\sum_k r_k^2$, which is unstable if there are outliers present in the data. Outlying data give an effect so strong in the minimizing that the parameters thus estimated are distorted. M-estimators try to reduce the effect of outliers by replacing the squared residuals r_k^2 by a function of the residuals, yielding

$$\sum_{k=1}^{n} \rho(r_k) \;\rightarrow\; \min, \tag{2.68}$$

where ρ is a symmetric, positive-definite function with a unique minimum at zero, and is chosen to be less increasing than square. Instead of solving directly this problem, it can be implemented as an iterated re-weighted least-squares one.

The M-estimator for the parameter vector $\mathbf{q}$ based on the function $\rho(r_k)$ is the vector $\mathbf{q}$ which is the solution of the following D equations:

$$\sum_{k=1}^{n} \psi(r_k) \frac{\partial r_k}{\partial q_j} = 0, \qquad \text{for } j = 1, \ldots D, \tag{2.69}$$

where the derivative $\psi(x) = d\rho(x)/dx$ is called the *influence function*. If the weight function W is defined as

$$\mathrm{W}(x) = \frac{\psi(x)}{x}, \tag{2.70}$$

then (2.69) becomes

$$\sum_{k=1}^{n} \mathrm{W}(r_k) r_k \frac{\partial r_k}{\partial q_j} = 0, \qquad \text{for } j = 1, \ldots D. \tag{2.71}$$

This is exactly the system of equations that are obtained is the following iterated re-weighted least-squares problem is solved

$$\sum_{k=1}^{n} \mathrm{W}(r_k^{i-1}) r_k^2 \;\rightarrow\; \min, \tag{2.72}$$

where the superscript ^{i} indicates the iteration number. The weight $\mathrm{W}(r_k^{i-1})$ should be recomputed after each iteration in order to be used in the next iteration.

The influence function $\psi(x)$ measures the influence of a datum on the value of the parameter estimate. For example, for the least-squares with $\rho(x) = x^2/2$, the influence function is $\psi(x) = x$. This means that the influence of a datum on the estimate increases linearly with the size of its error, confirming the non-robustness of the ML estimate. When an estimator is robust, it may be inferred that the influence of any single observation (datum) is insufficient to yield any significant offset. A robust M-estimator must therefore have a bounded influence function. Furthermore, a robust estimator should be unique.

Table 2.1 lists a few commonly used influence functions. They are graphically depicted in Fig. 2.12:

- L_2 (least squares) estimators are not robust because their influence function is not bounded.
- L_1 (absolute value) estimators are not stable because the ρ-function $|x|$ is not strictly convex in x. The second derivative at $x = 0$ is unbounded, and an indeterminate solution may result.
- L_1 estimators reduce the influence of large errors, but they still have an influence because the influence function has no cut off point.
- $L_1 - -L_2$ estimators take both the advantage of the L_1 estimators to reduce the influence of large errors and that of L_2 estimators to be convex.
- The L_p (least powers) function represents a family of functions. It is L_2 with $\nu = 2$ and L_1 with $\nu = 1$. The smaller the ν, the smaller is the incidence of large errors in the estimate $\mathbf{q}$. The parameter ν must be fairly moderate in order to provide a relatively robust estimator. The selection of an optimal ν has been investigated, and for ν around 1.2, a good estimate may be expected [Zhang, 1995]. However, the computation results in many difficulties when parameter ν is in the range of interest $1 < \nu < 2$, because zero residuals are troublesome.

Type	$\rho(x)$	$\psi(x)$	W(x)
L_2	$x^2/2$	x	1
L_1	$\lvert x\rvert$	sgn(x)	$\frac{1}{\lvert x\rvert}$
$L_1 - -L_2$	$2(\sqrt{1+x^2/2}-1)$	$\frac{x}{\sqrt{1+x^2/2}}$	$\frac{1}{\sqrt{1+x^2/2}}$
L_p	$\frac{\lvert x\rvert^\nu}{\nu}$	sgn$(x)\lvert x\rvert^{\nu-1}$	$\lvert x\rvert^{\nu-2}$
"Fair"	$c^2[\frac{\lvert x\rvert}{c} - \log(1+\frac{\lvert x\rvert}{c})]$	$\frac{x}{1+\lvert x\rvert/c}$	$\frac{1}{1+\lvert x\rvert/c}$
Huber $\begin{cases} \text{if} \lvert x\rvert \le k \\ \text{if} \lvert x\rvert \ge k \end{cases}$	$\begin{cases} x^2/2 \\ k(\lvert x\rvert - k/2) \end{cases}$	$\begin{cases} x \\ k\text{sgn}(x) \end{cases}$	$\begin{cases} 1 \\ k/\lvert x\rvert \end{cases}$
Cauchy	$\frac{c^2}{2}\log(1+(x/c)^2)$	$\frac{x}{1+(x/c)^2}$	$\frac{1}{1+(x/c)^2}$
Geman-McClure	$\frac{x^2/2}{1+x^2}$	$\frac{x}{(1+x^2)^2}$	$\frac{1}{(1+x^2)^2}$
Welsch	$\frac{c^2}{2}[1-\exp(-(x/c)^2)]$	$x\exp(-(x/c)^2)$	$\exp(-(x/c)^2)$
Tukey $\begin{cases} \text{if} \lvert x\rvert \le c \\ \text{if} \lvert x\rvert > c \end{cases}$	$\begin{cases} \frac{c^2}{6}(1-[1-(x/c)^2]^3) \\ (c^2/6) \end{cases}$	$\begin{cases} x[1-(x/c)^2]^2 \\ 0 \end{cases}$	$\begin{cases} [1-(x/c)^2]^2 \\ 0 \end{cases}$

Table 2.1: A few commonly used M-estimators

- The function "Fair" has everywhere defined continuous derivatives of first three orders, and yields a unique solution. Asymptotic efficiency on the standard normal distribution of 95 % is obtained with the tuning constant $c = 1.3998$ [Rissanen, 1987].
- Huber's function [Huber, 1964] is a parabola in the vicinity of zero, and increases linearly at a given level $|x| > k$. The 95 % asymptotic efficiency on the standard normal distribution is obtained with the tuning constant $k = 1.345$.
- Cauchy'function does not guarantee a unique solution. With a descending first derivative, such a function has a tendency to yield erroneous solutions in a way which which cannot be observed. The 95 % asymptotic efficiency on the standard normal distribution is obtained with the tuning constant $c = 2.3849$.
- The other remaining functions have the same problem as the Cauchy function. As can be seen from the influence function, the influence of large errors only decreases linearly with their size. The Geman-McClure and Welsh functions try to further reduce the effect of large errors, an the Tukey's biweight function even suppresses the outliers. The 95 % asymptotic efficiency on the standard normal distribution of the Tukey's biweight function is obtained with the tuning constant $c = 4.6851$; That of the Welsch function with $c = 2.9846$.

It seems difficult to select a ρ-function for general use without being rather arbitrary. For location (or regression) problems, the best choice is the L_p in spite of its theoretical non-robustness: They are quasi-robust. However, it suffers from computational difficulties. The second best

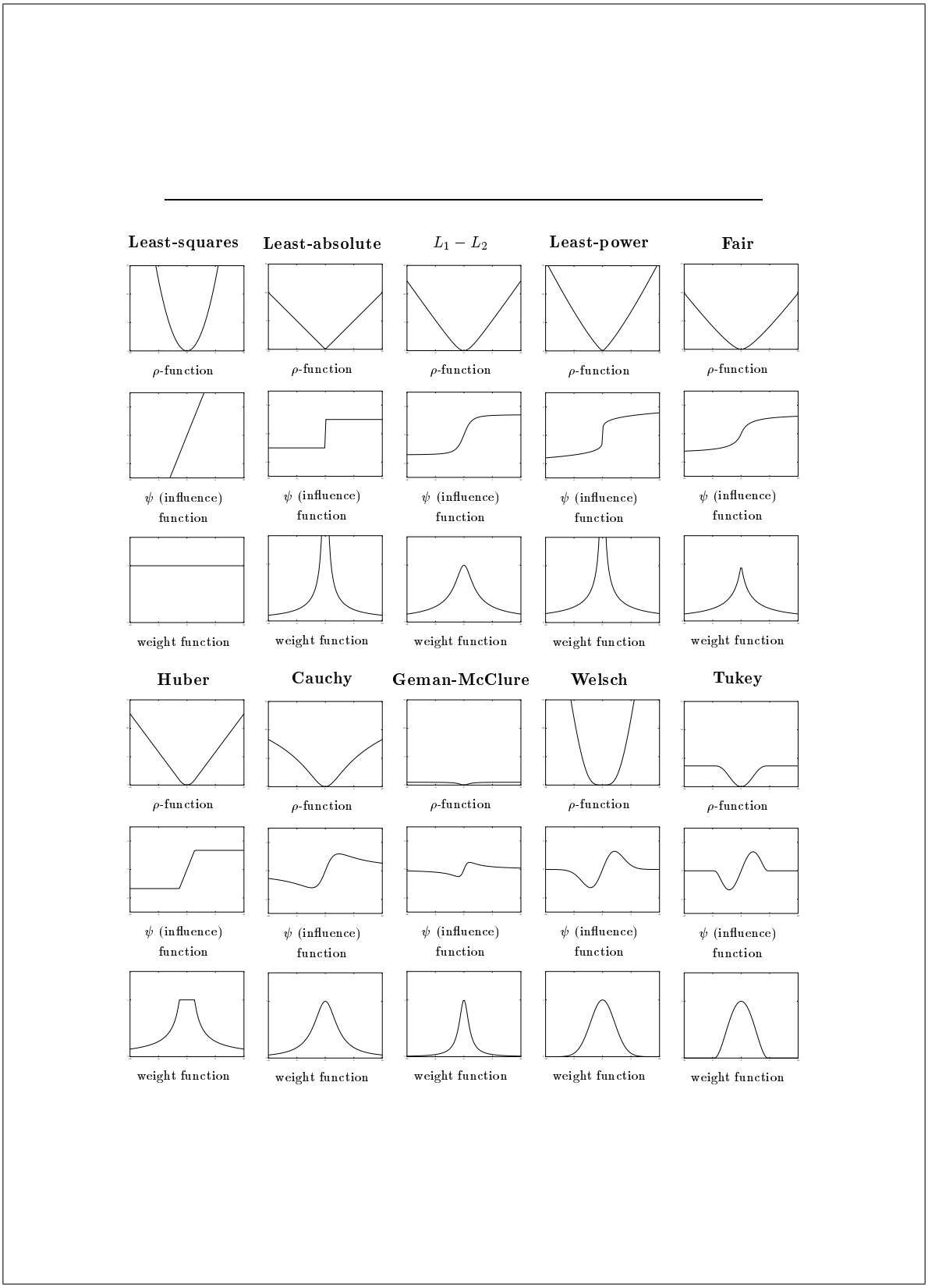


Figure 2.12: Graphic representation of a few common M-estimators [Zhang, 1995].

function is “Fair”, which can yield nicely converging computational procedures. All these functions and Huber’s function do not eliminate completely the influence of large gross errors. The four last functions do not guarantee unicity, but reduce considerably, or even eliminate completely, the influence of large gross errors.

Scale Invariance

The solution to (2.69) is not scale-invariant, since in general $\psi(cx) \neq c\psi(x)$. In practice it means that an M-estimator should be supplemented by an *estimator of scale* σ. The scale invariant version of the M-estimator is defined by the solution to the equation:

$$\sum_{k=1}^{n} \mathrm{W}(r_k/\sigma) r_k \frac{\partial(r_k/\sigma)}{\partial q_j} = 0, \qquad \text{for } j = 1, \ldots D. \tag{2.73}$$

This procedure is also called studentizing. Since σ is usually unknown it is replaced by a robust estimator of the scale. For univariate estimators of location, one usually takes the MAD (median absolute deviation) divided by 0.6745:

$$\tilde{\sigma} = \frac{\text{median}|x_k - \text{median}(x_k)|}{0.6745}, \tag{2.74}$$

where $x_1, x_2, \ldots, x_k$ is a sequence of identically independently distributed (i.i.d.) observations. The MAD is a robust estimator for $u_{0.75} \cdot \sigma = 0.6745\sigma$ ($u_{0.75}$ is the 0.75 quantile of the standard normal distribution). So MAD/0.6745 is a robust estimator for σ.) The breakdown point of this estimator is $\epsilon^* = 0.5$.

M-Estimators for Gaussian Spots

The theory of robust M-estimators for multivariate distributions with elliptically symmetric density function (a Gaussian, for example) has been studied by Maronna [Maronna, 1976]. The elliptically symmetric density function can be written as

$$f(\mathbf{p}) = |\boldsymbol{\Sigma}|^{-1/2} h\{[(\mathbf{p} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{p} - \boldsymbol{\mu})]^{-1/2}\} \tag{2.75}$$

where h is a density function in $\mathbb{R}^m$, for $\mathbf{p} \in \mathbb{R}^m$. Let $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k$ be a sequence of i.i.d observations with density stated in (2.75). Maronna’s estimators for location $\widehat{\boldsymbol{\mu}}$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}$ are computed as solutions of the equations:

$$\widehat{\boldsymbol{\mu}} = \sum_{k=1}^{n} \mathbf{p}_k \mathrm{W}(e_k) \Big/ \sum_{k=1}^{n} \mathrm{W}(e_k) \tag{2.76}$$

and

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^{n} (\mathbf{p}_k - \widehat{\boldsymbol{\mu}})(\mathbf{p}_k - \widehat{\boldsymbol{\mu}})^T (\mathrm{W}({e_k}^2))^2, \tag{2.77}$$

where

$${e_k}^2 = (\mathbf{p}_k - \widehat{\boldsymbol{\mu}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{p}_k - \widehat{\boldsymbol{\mu}}) \tag{2.78}$$

Sufficient conditions on the weight functions for which solutions to (2.76) and (2.77) is given in [Maronna, 1976]. If $\mathbf{p}_k$ has the distribution (2.75), then $\widehat{\boldsymbol{\mu}}$ estimates $\boldsymbol{\mu}$ consistently and $\widehat{\boldsymbol{\Sigma}}$ estimates $c \cdot \boldsymbol{\Sigma}$, where the constant c depends on on $\mathbf{W}(\cdot)$ and the probability distribution function.

We use Maronna's M-estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in order to compute the corresponding parameters of the Gaussian spot model (2.62). We use a weighting scheme based on the deviation from Gaussian model which is more suitable for spot overlap and outlier handling. In order to estimate the parameters altogether efficiently, we introduce a joint estimation for the mean $\boldsymbol{\mu}$, the covariance matrix $\boldsymbol{\Sigma}$ and amplitude a:

$$\widehat{\boldsymbol{\mu}} = \sum_{k=1}^{n} z_k \mathbf{p}_k \mathrm{W}(e_k) \Bigg/ \sum_{k=1}^{n} z_k \mathrm{W}(e_k) \tag{2.79}$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T} \sum_{k=1}^{n} z_k (\mathbf{p}_k - \widehat{\boldsymbol{\mu}})(\mathbf{p}_k - \widehat{\boldsymbol{\mu}})^T (\mathrm{W}(e_k))^2. \tag{2.80}$$

$$\widehat{a} = \sum_{k=1}^{n} z_k \mathrm{G}(\mathbf{p}_k, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \mathrm{W}(e_k) \Bigg/ \sum_{k=1}^{n} \mathrm{G}(\mathbf{p}_k, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})^2 \mathrm{W}(e_k) \ , \tag{2.81}$$

where T is the total sum of the intensities of the patch:

$$T = \sum_{k=1}^{n} z_k \tag{2.82}$$

and the studentized error e_k between the data and the model for each point is

$$e_k := e_k(\widehat{a}, \widehat{b}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) := \frac{(z_k - \widehat{a}\, \mathrm{G}(\mathbf{p}_k, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}))}{\sigma} \tag{2.83}$$

with unknown spread σ.

The equations for $\widehat{\boldsymbol{\mu}}$ (2.79), $\widehat{\boldsymbol{\Sigma}}$ (2.80) and $\widehat{a}$ (2.81) can be solved by the weighted least squares iteration:

$$\widehat{\boldsymbol{\mu}}_{i+1} = \sum_{k=1}^{n} z_k \mathbf{p}_k \mathrm{W}(e_{ki}) \Bigg/ \sum_{k=1}^{n} z_k \mathrm{W}(e_{ki}) \tag{2.84}$$

$$\widehat{\boldsymbol{\Sigma}}_{i+1} = \frac{1}{T} \sum_{k=1}^{n} z_k (\mathbf{p}_k - \widehat{\boldsymbol{\mu}})(\mathbf{p}_k - \widehat{\boldsymbol{\mu}})^T (\mathrm{W}(e_{ki}))^2. \tag{2.85}$$

$$\widehat{a}_{i+1} = \sum_{k=1}^{n} z_k \mathrm{G}(\mathbf{p}_k, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \mathrm{W}(e_{ki}) \Bigg/ \sum_{k=1}^{n} \mathrm{G}(\mathbf{p}_k, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})^2 \mathrm{W}(e_{ki}) \tag{2.86}$$

with

$$e_{ki} := e_k(\widehat{a}, \widehat{b}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) := \frac{(z_k - \widehat{a}\, \mathrm{G}(\mathbf{p}_k, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i))}{\sigma_i} \tag{2.87}$$

and

$$\sigma_j = \mathrm{median}_{k \in \mathcal{I}} \frac{|z_k - a_i\, \mathrm{G}(\mathbf{p}_k, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)|}{0.6745}, \tag{2.88}$$

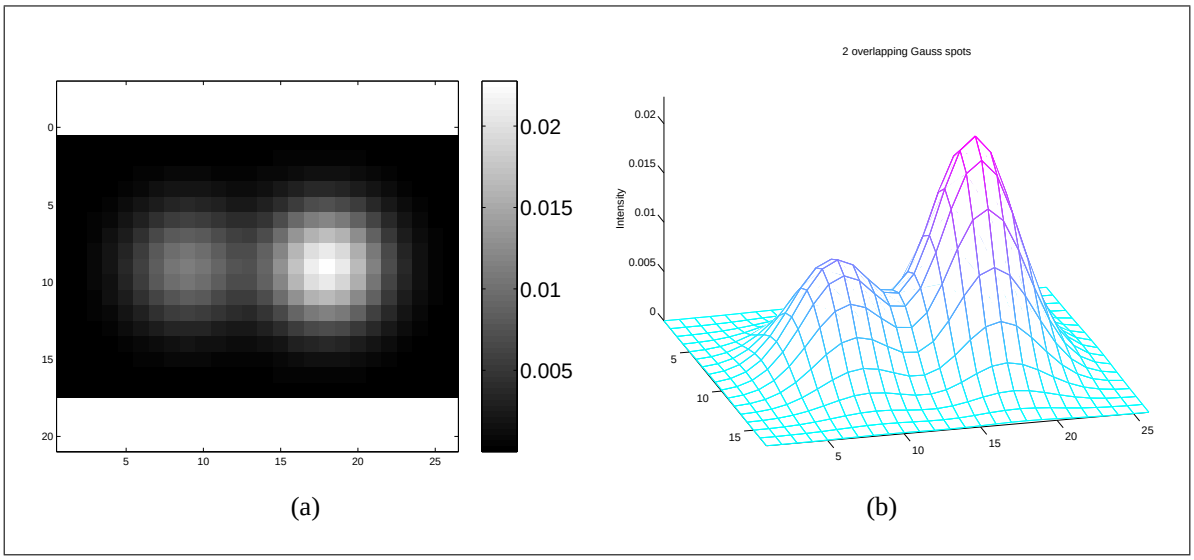


Figure 2.13. Overlapping Syntehtic Gaussian Spots. The right spot has the double intensity of the left spot.

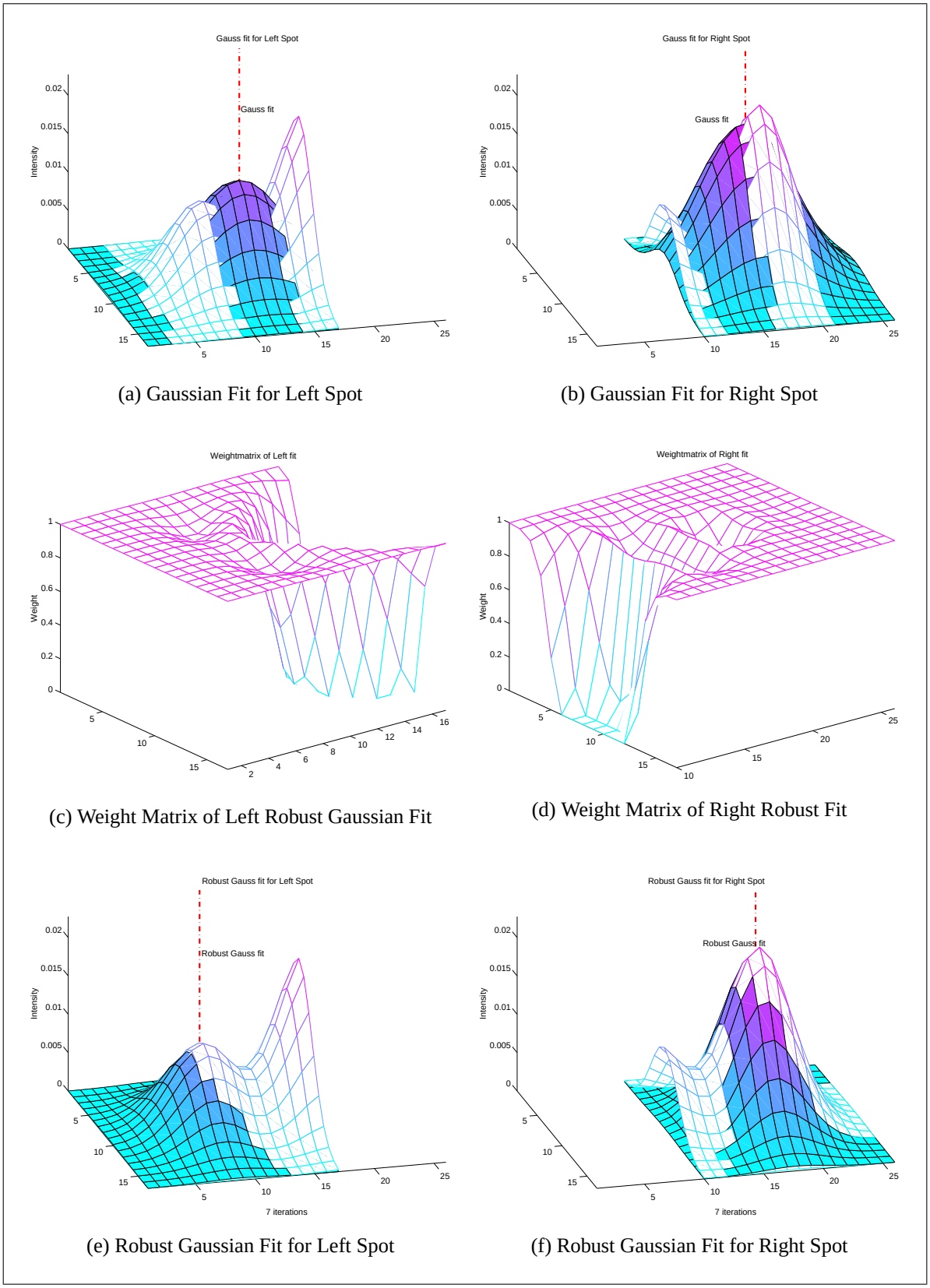
where the scale is estimated in analogy to (2.74). Not all data points are used for the scale estimation, because the estimator should not fit the points far from the spot center and classifying the center to be an outlier. This may happen when a bigger portion of background is visible. Only points 'inside' the ellipse defined by the covariance matrix $\widehat{\boldsymbol{\Sigma}}_i$ are considered. The region $\mathcal{I}$ in (2.88) is defined as

$$\mathcal{I} = \{k \mid (\mathbf{p}_k - \widehat{\boldsymbol{\mu}}_i)\widehat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{p}_k - \widehat{\boldsymbol{\mu}}_i)^T) < c\} \tag{2.89}$$

.

The threshold c for the Mahalanobis distance can be chosen such that 'inside' means the 0.95 percentile of the elliptical regions. An univariate standard normal distribution has its 0.95 percentile at 1.6449 (=Eucledian distance from 0). For the bivariate Gaussian distribution one choses $c = (1.6449)^2)$. For the initial values $\widehat{\boldsymbol{\mu}}_0$, $\widehat{\boldsymbol{\Sigma}}_0$ and $\widehat{a}_0$ one can use the ML-estimators (2.65) – (2.67). Alternatively one could take the median or the spot center for $\widehat{\boldsymbol{\mu}}_0$ and a matrix with the squared MAD in the diagonal and zero covariance for $\widehat{\boldsymbol{\Sigma}}_0$ and compute $\widehat{a}_0$ with this new $\widehat{\boldsymbol{\mu}}_0$ and $\widehat{\boldsymbol{\Sigma}}_0$.

Fig. 2.13 shows two synthetic Gaussian spots, where the right spot has twice the intensity of the left spot. It can be seen in Fig. 2.14a and b that the non-robust fit biases the the spot model towards the neighbor. Figures 2.14c and d shows the weights $\mathrm{W}(e_k)$ for every pixel $\mathbf{p}_k$ as used in Eqns (2.84)– (2.86). The weights are between 0 and 1 , where 'good' data points receive weight close to 1 while 'bad' data points which cause too high residuals receive weight close to 0. As expected, the data points corrupted by overlap are downweighted. Figures 2.14e and f show the robust Gaussian fit after six iterations of (2.84)– (2.86).



(a) Gaussian Fit for Left Spot

(b) Gaussian Fit for Right Spot

(c) Weight Matrix of Left Robust Gaussian Fit

(d) Weight Matrix of Right Robust Fit

(e) Robust Gaussian Fit for Left Spot

(f) Robust Gaussian Fit for Right Spot

Figure 2.14: Robust Spot Fitting for the Synthetic Data in Fig. 2.13.

2.6.3 Relative Error and Goodness-of-Fit

In order to quantitatively assess how well a model assumption holds for a given set of n intensities z_k belonging to a spot we introduce a measure for the error. We apply an approach also used in linear regression analysis as in [Hartung, 1989] by comparing the model to a "standard model Z_0":

$$T_1 := \frac{1}{n}\sum_{k=1}^{n}(z_k - \mathrm{Z}(\mathbf{p}_k))^2 \Bigg/ \frac{1}{n}\sum_{k=1}^{n}(z_k - z_0)^2 \tag{2.90}$$

with $z_0 := \frac{1}{n}\sum z_k$ as the mean of the given intensity values. The standard model Z_0 in this case is a plane parallel to the image plane at the height z_0, i.e. $Z_0 \equiv z_0$. T_1 relates the mean squared error between the standard model and the data to the mean squared error between a constant model and the data. T_1 can also be regarded as the mean squared error between the spot model and data normalized by the variance of the error. We will call T_1 the *relative squared error* or for short *relative error*. In the literature $1 - T_1$ is called the *goodness-of-fit*.

2.6.4 Quantification

The brightness V of the spot is quantified as the volume under the fitted Gaussian function:

$$\widehat{V} = \widehat{a}\,(2\pi)\,\sqrt{\det(\widehat{\boldsymbol{\Sigma}})}. \tag{2.91}$$

A derivation of the Gaussian integral (2.91) can be found in [Bishop, 1995].

2.7 Semi-parametric and non-parametric spot fitting

A semi-parametric approach can describe the spot shape more accurately in the case of deviations from the model assumptions. However, overlap handling will be difficult, because a semi parametric fit will lack an intrinsic declension of the tails of a parametric model.

The basic idea of this method is to reduce dimensionality of given data using prior knowledge. Assuming that the spot has elliptically symmetric shape the fit is computed in the following steps:

A. Find the spot center We first perform a Gaussian fit computing M-estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as described in (2.79) and (2.80). The estimate $\widehat{\boldsymbol{\mu}}$ is the spot center. Since the M-estimator of the location is robust it will also deal with spots with uncommon shapes. Passing a line perpendicular to the x, y-plane through $\widehat{\boldsymbol{\mu}}$ gives us an axis $\mathbf{a}$.

B. Transform the points The estimated dispersion matrix $\widehat{\boldsymbol{\Sigma}}$ gives us an ellipse in the x, y-plane. Let e_1 and e_2 be the two eigenvalues of $\widehat{\boldsymbol{\Sigma}}$, (without loss of generality $e_1 \geq e_2$), $\mathbf{v}_1$ and $\mathbf{v}_2$ the corresponding eigenvectors and ϵ be the half-plane spanned by $\lambda_1\mathbf{a} + \lambda_2\mathbf{v}_1;\ \lambda_1 \in \mathbb{R},\ \lambda_2 \in \mathbb{R}_0^+$. Consider the one parametric family of ellipses with the principle axis directions $\mathbf{v}_1$ and $\mathbf{v}_2$, diameters λe_1 and λe_2, $\lambda \in \mathbb{R}_0^+$ and center $\boldsymbol{\mu}$. The family covers the x, y-plane without intersection, each point in the x, y-plane lies exactly on one ellipse. We "rotate" the given

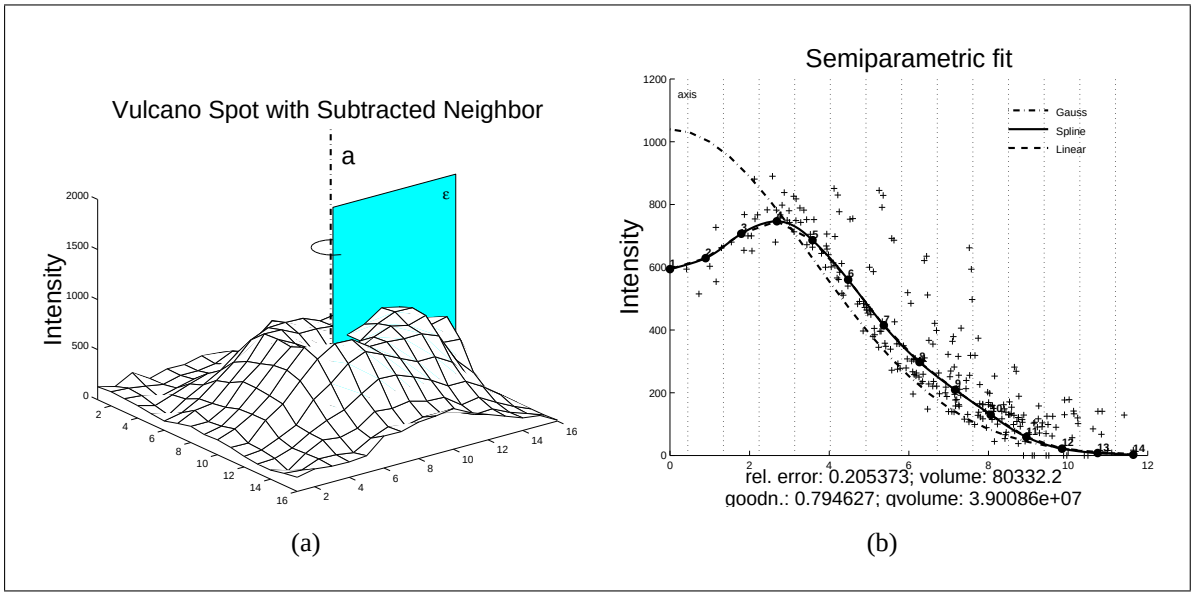


Figure 2.15: Semi-parametric Spot Fitting

intensity points $(\mathbf{p}_k, z_i)$ following the path corresponding to $\mathbf{p}_k$ into the half-plane ϵ yielding a point cloud $\mathbf{q}_i$ in 2-space (see Fig. 2.15a. The first coordinate can be easily computed by:

$$e_1 \cdot |\mathbf{p}|^2 / \sqrt{e_1^2(\mathbf{p} \cdot \mathbf{v}_1)^2 + e_2^2(1 - (\mathbf{p} \cdot \mathbf{v}_1)^2)} \tag{2.92}$$

and the second coordinate is the unchanged z-coordinate.

C. Compute a profile We introduce a simplified, efficient and robust version of curve approximation for scattered points suited to our purpose. First we compute m points $\mathbf{c}_i = (x_i, y_i),\ i = 1, \ldots, m$ well describing the shape of curve to be computed. Consider the vertical parallel strip with y-axis and $x \equiv \max x_i$ as borders. We then segment the strip into m commensurate parallel strips and compute $\mathbf{c}_i = \text{median}_k \mathbf{r}_{ki}$, where $\mathbf{r}_{ki}$ are those points $\mathbf{q}_k$ lying in the i^{th} strip, see Fig. 2.15b. We further cut off tails of the profile by gradually lowering the profile points down to zero in the last quarter, because 1) especially at the tail there may be some overlapping situation and 2) generally there are fewer points at the tail. For our purpose it is enough to interpolate the points $\mathbf{c}_i$ by a polygon and to perform a smoothing scheme on the profile points, e.g. by replacing each point with a weighted sum of its neighbors. Alternatively one can compute a spline interpolating the points $\mathbf{c}_i$ for the profile curve.

D. Compute Volume The profile curve is rotated following the elliptical paths as in step B. The brightness V of the spot is then estimated by taking

$$\widehat{V} = \frac{e_2}{e_1} \cdot \frac{1}{3} \sum_{i=2}^{m} (x_{i-1}^2 + x_{i-1}x_i + x_i^2)\pi(y_i - y_{i-1}). \tag{2.93}$$

Non-parametric quantification does not use any parameter in the desciption of the intensity distribution of the spot. An overview of non-parametric methods like the Parzen window approach can be found in [Bishop, 1995] or [Duda et al., 2000]. It will also have difficulties in

dealing with spot overlap. We also count the segmentation methods introducted in Sect. 1.3.2 to the class of non-parametric methods.

2.8 Chapter Summary

The proposed approach for spot array image analysis is composed of a set of robust tools arranged in a general framework (see Fig. 2.16). The image analysis starts with an amplification of the spot locations with the help of matched filters built by averaging a number of representative tranining spots. The matched filter response is expected to have maximum values at the spot locations. The grid rotation is estimated with the help of projections of matched filter responses along different directions. The projection at the correct rotation angle is expected to have maximum projection values. Two alternative grid spanning modules enable to span a consistent grid with the help of the matched filter response and the estimated rotation angle. The first grid spanning method transforms a prior spot grid according to the grid corner locations. The grid corner locations are estimated by robustly fitting straight lines to the first and last row and column of the grid. If the spot distances are too irregular, an alternative grid spanning projects intersects

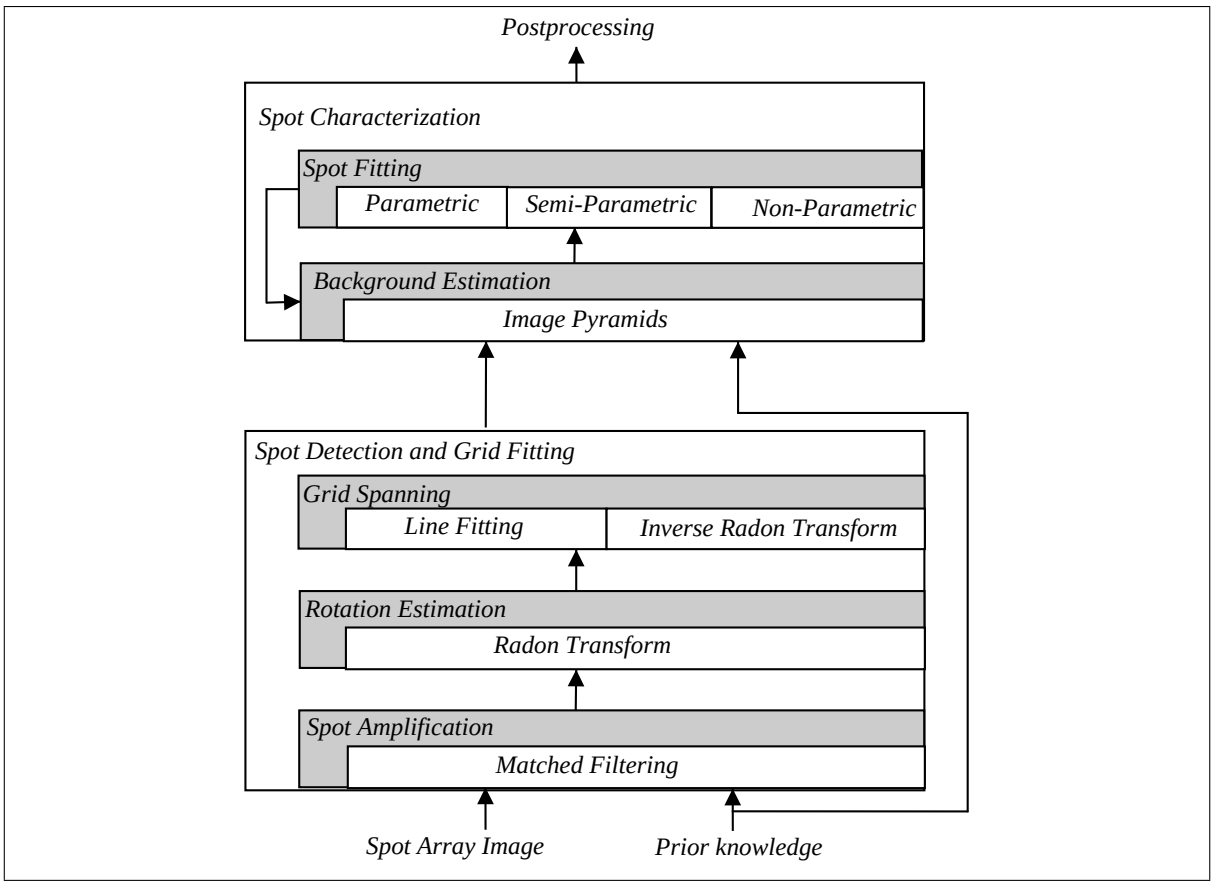


Figure 2.16. General Image Analysis Framework. Overview of the general approach for the image analysis of spot array images.

back-projected straight lines based on the inverse Radon transform. The spot characterization step consists of the interdependent background estimation and spot fitting. The background estimation uses a hierarchical pyramidal approach in order to yield smooth backgrounds. At a high pyramid level with low resolution, a synthetic image based on spot fitting is subtracted and the result is interpolated to the original spot image resolution.

The core spot fitting approach in our framework is based on statistical spot models. The parameters of the statistical model describe the observed distribution of pixel intensities and must be fitted to the observed data. Maximum Likelihood (ML) estimators find the parameters that best desribe the observed data based on the squared error between the data and the model. The parameters of a Gaussian model constist of an overall height (amplitude), two parameters for the spot center (mean) and three parameters for the dispersion of the spot (covariance matrix). The squared error of the ML estimator is non-robust against outliers. Robust M-estimators weight the contribution of a data point according to the residual error between the data and the model: A large residual will be scaled down or even be truncated in order not to bias the estimator. The semi-parametric approache is based on the robust parametric fit and then performs a dimension reduction by rotating a plane around the spot center. A robustly fitted curve in the plane then allows to model deviation the Gaussian form, e.g. a volcano spot. Non-parametric approaches include the classical segmentation methods to segregate signal and background.

The proposed approach has two main characteristics: 1. The tools do not need critical thresholds provided by users. The only user-information is the array configuration. 2. The spot characterization is based on the original image data. No information is lost by a geometric compensating rotation transform with its unavoidable interpolation.

Chapter 3

DNA Array Technology

In the domain of biotechnology, the past few years have witnessed an extraordinary surge of interest in the DNA array technology. This technology offers a great hope for providing a systematic way to explore the genome, which is defined as the full set of genetic information that an individual organism inherits from its parents. It permits very rapid analysis of thousands of genes for the purposes of gene discovery, sequencing, mapping and expression. The massive data generated by experiments based on DNA arrays need to be managed by powerful tools capable of rational analysis. This chapter outlines the basic principles of DNA array technology. It starts with a description of the biological knowledge indispensible for understanding DNA array technology (Sect. 3.1. This section can be skipped by readers familiar with fundamental knowledge about molecular biology. Section 3.2 introduces the different types of DNA arrays. A list of possible applications of DNA array technology can be found in Sect. 3.3. Section 3.4 describes the main step of a DNA array experiment.

3.1 Fundamentals of Molecular Biology

Inherited characteristics of an organism are defined by their ability to be passed from one generation to the next in a predictable manner. It is important to realize the distinction between the appearance of the organism (what we observe) and the underlying genetic constitution (which we must infer). Visible or otherwise measurable properties are called the **phenotype**, while the genetic factors responsible for creating the phenotype are called the **genotype**. The gene is the unit of inheritance. Each gene is a nucleic acid sequence that carries the information. A gene is a stable entity, but can suffer a change in sequence. Such a change is called a **mutation**. When a mutation occurs, the new form of the gene is inherited in a stable manner, just like the previous form. The basic paradigm in the science of genetics is that genes encode proteins, which in turn are responsible for the synthesis of other types of structures. The sequence of a gene specifies the sequence of a protein, and therefore its molecular structure. Each protein consists of a unique sequence of amino acids. Twenty amino acids are used to synthesize proteins. An extensive overview of Genetics an be found in [Lewin, 1997]

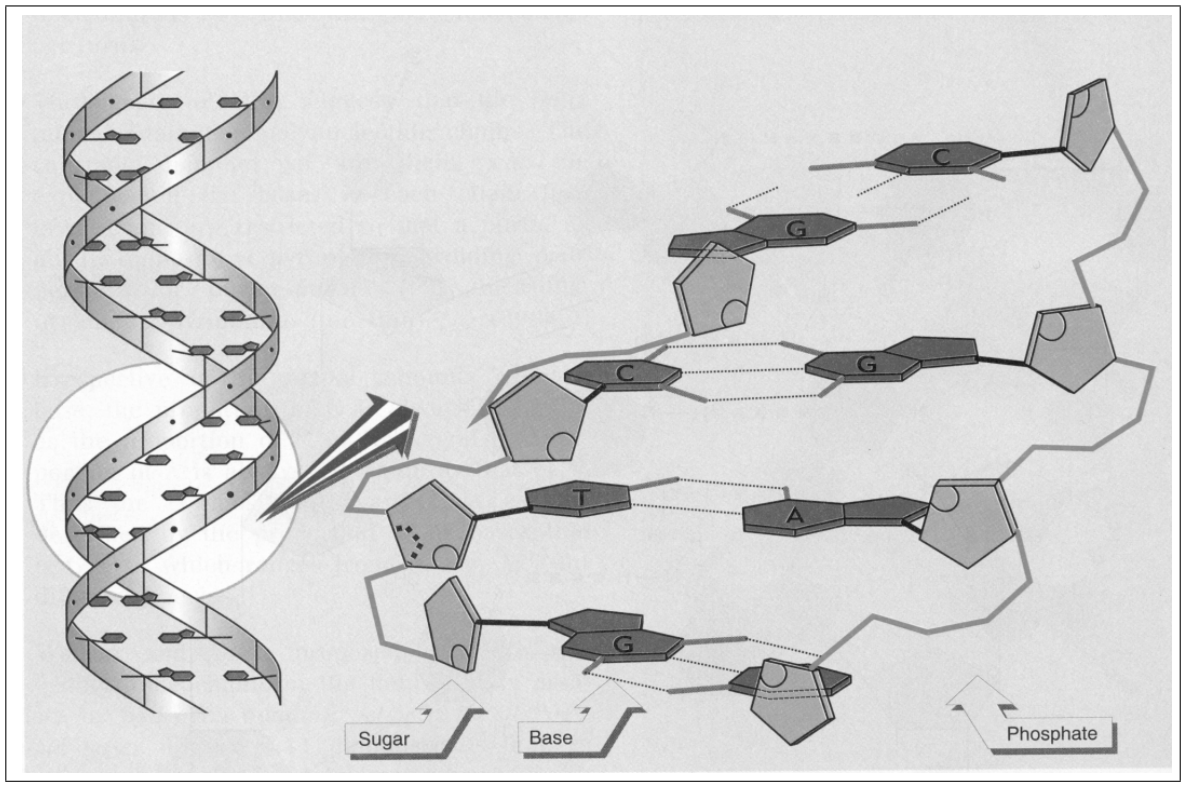


Figure 3.1. DNA. DNA consists of a chemically linked sequence of subunits. Each subunit contains a nitrogenous base (A, G, C, T), a sugar and a phosphate group. G can hydrogen bond specifically only with C, while A can bond specifically only with T. [Lewin, 1997]

3.1.1 DNA as the Genetic Material

The genetic material of all known organism and many viruses is *deoxyribonucleic acid (DNA)*. Some viruses use an alternative nucleic acid – *ribonucleic acid – (RNA)* as the genetic material. The general principle of the nature of genetic material is that it is always nucleic acid.

A nucleic acid like DNA consists of a chemically linked sequence of subunits. Each subunit contains a *nitrogenous base*, a sugar and a phosphate group. The nitrogenous bases fall into two types *pyrimidines* and *purines*. Each nucleic acid contains 4 types of base. The same two purines, adenine and guanine, are present in both DNA and RNA. The two pyrimidines in DNA are cytosine and thymine; in RNA uracil is found instead of thymine. The bases are usually referred to by their initial letters; so DNA contains A, G, C, T, while RNA contains A, G, C, U. Watson and Crick [Watson and Crick, 1953] proposed the double helix model for DNA, in which two polynucleotide chains in the double helix associate by *hydrogen (weak) bonding between the nitrogenous bases*. Figure 3.1 demonstrates that G can hydrogen bond specifically only with C, while A can bond specifically only with T. These reactions are described as *base pairing*, and the paired bases (G with C, or A with T) are said to be *complementary*.

3.1.2 Central Dogma

The *central dogma* defines the paradigm of molecular biology: genes are **perpetuated** as sequences of nucleic acid, but function by being **expressed** in the form of proteins. Three types of processes are responsible for the inheritance of genetic information and for its conversion from one form to another:

- Information is perpetuated by **replication**; a double-stranded nucleic acid is duplicated to give identical copies.

Information is expressed by a two stage process:

- **Transcription** generates a single-stranded RNA identical in sequence with one of the strands of the duplex DNA. Several different types of RNA are generated by transcription. The three principal classes involved in the synthesis of proteins are **messenger RNA (mRNA)**, **transfer RNA (tRNA)** and **ribosomal RNA (rRNA)**.

- **Translation** converts the nucleotide sequence of RNA into the sequence of amino acids comprising a protein. An mRNA is translated into a protein sequence; tRNA and rRNA provide other components of the apparatus for protein synthesis. The entire length of an mRNA is not translated, but each mRNA contains at least one **coding region** that is related to a protein sequence by the genetic code: each group of three nucleotides (codon) of the coding region represents one amino acid.

Only one strand of a DNA duplex is transcribed into a messenger RNA. The strand of DNA that directs synthesis of the mRNA via complementary base pairing is called the **template strand**. The other DNA strand bears the *same* sequence as the mRNA (except for possessing T instead of U) and is called the **coding strand**. Figure 3.2 illustrates the roles of replication, transcription and translation:

- The perpetuation of nucleic acid may involve either DNA or RNA as the genetic material. Cells use only DNA. Some viruses use RNA, and replication of viral RNA occurs in the infected cell.

- The expression of cellular genetic information is usually unidirectional. Transcription of DNA generates RNA molecules that can be used further only to generate protein sequences. Generally they cannot be retrieved for use as genetic information. Translation of RNA into protein is always irreversible.

The restriction to unidirectional transfer from DNA to RNA is not absolute. It is overcome by the **retroviruses**, whose genomes consist of single-stranded RNA molecules. During infection, the RNA is converted by the process of **reverse transcription** into a single-stranded DNA, which in turn is converted into a double-stranded DNA. This duplex DNA becomes part of the genome of the cell, and is inherited like any other gene. Reverse transcription therefore allows a sequence of RNA to be retrieved and used as genetic information.

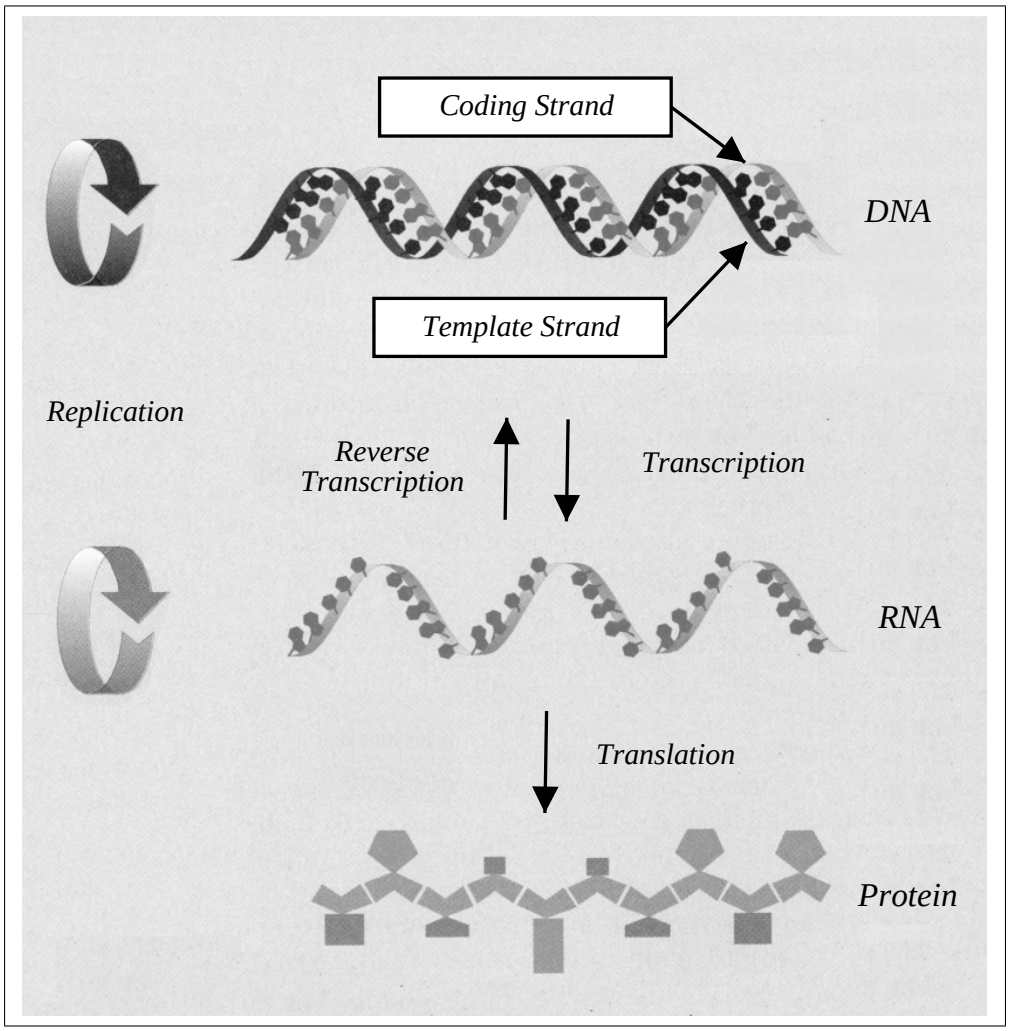


Figure 3.2. Central Dogma of Molecular Biology. Transcription generates an RNA which is complementary to the DNA template strand and has the same sequence as the DNA coding strand. Translation reads each triplet of bases into one amino acid. The central dogma states that information in nucleic acid can be perpetuated or transferred, but the transfer of information into protein is irreversible.

3.1.3 Hybridization

The same features that allow DNA to fulfill its biological role make it possible to manipulate the nucleic acid *in vitro*, and ultimately to isolate the segment of DNA that represents a particular protein. The hydrogen bonds that stabilize the double helix are disrupted by heating or by exposure to low salt concentration. The two strands of a double helix separate entirely when all the hydrogen bonds between them are broken. The process of strand separation is called **denaturation**.

Nucleic acid sequences can be assessed in terms of either similarity or complementarity. *Similarity* between two sequences is given in principle by the proportion of bases or base pairs that is identical. However, without determining the actual sequences, there is no direct way to measure similarity. *Complementarity* is determined by the rules for base pairing between

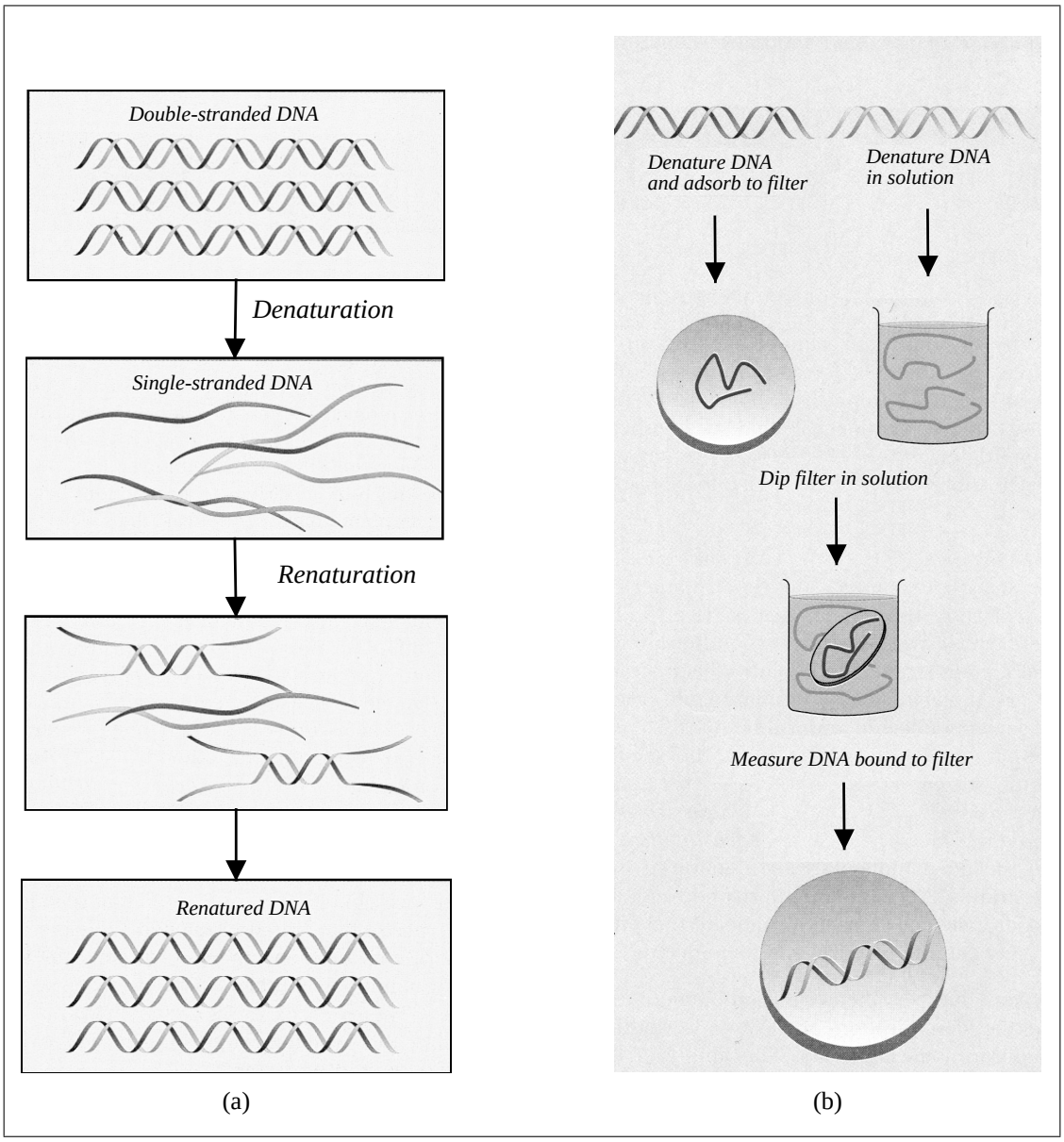


Figure 3.3. Denaturation and Renaturation. (a) Denatured single strands of DNA can renature to give the duplex form. (b) Filter hybridization establishes whether a solution of denatured DNA (or RNA) contains sequences complementary to the strands immobilized on the filter. [Lewin, 1997]

A $\leftrightarrow$ T and G $\leftrightarrow$ C . In a perfect duplex of DNA, the strands are precisely complementary. By comparing different but related double-stranded molecules, each strand of the first molecule will be similar to one strand of the second molecule and will be (partly) complementary to the other strand of the second molecule. Complementarity can be measured directly by the ability of two single-stranded nucleic acids to base pair with each other. If double-stranded molecules are denatured into single strands, the complementarity between the single strands can be used to indicate the similarity between the original molecules.

It is possible to measure complementarity because the denaturation of DNA is reversible under appropriate conditions. The ability of the two separated complementary strands to reform into a double helix is called **renaturation** (Fig. 3.3). Renaturation depends on specific base pairing between the complementary strands. The reaction takes place in two stages. First, single strands of DNA in the solution encounter one another by chance. If their sequences are complementary, the two strands base pair to generate a short double-helical region. Then the region of base pairing extends along the molecule by a zipper-like effect to form a long duplex molecule. Renaturation of the double helix restores the original properties that were lost when the DNA was denatured.

Renaturation describes the reaction between two complementary sequences that were separated by denaturation. However, the technique can be extended to allow any two complementary nucleic acid sequences to **anneal** with each other to form a duplex structure. The reaction is generally described as **hybridization** when nucleic acids from different sources are involved, as in the case when one preparation consists of DNA and the other consists of RNA. The ability of two nucleic acid preparations to hybridize constitutes a precise test for their complementarity since only complementary sequences can form a duplex structure.

The principle of the hybridization reaction is to expose two single-stranded nucleic acid preparations to each other and then to measure the amount of double-stranded material that forms. **Filter hybridization** uses the nitrocellulose filter property of adsorbing single strands of DNA but not RNA. Once a filter has been used to adsorb DNA, it can be treated to prevent any further adsorption of single strands. Figure 3.3b illustrates the resulting procedure in which a DNA preparation is denatured and the single strands are adsorbed to the filter. Then a second denatured DNA (or RNA) preparation is added. This material adsorbs to the filter only if it is able to base pair with the DNA that was originally adsorbed. The usual form of the experimental procedure is to add a radioactively labeled RNA or DNA preparation to the filter. This allows to measure the extent of reaction as the amount of radioactive label retained by the filter.

The extent of hybridization between two single-stranded nucleic acids can be taken in principle to represent their degree of complementarity. Two sequences need not be perfectly complementary to hybridize. If they are closely related but not identical, in imperfect duplex is formed in which base pairing is interrupted at positions where the two single strands do not correspond.

3.1.4 mRNA Extraction and Reverse Transcription

If mRNA is required for use in hybridization experiments, it must be purified from total cellular contents: mRNA accounts for only about 3 % of all RNA in a cell – the rest is tRNA and ribosomal RNA. Thus isolating it in sufficient quantity for an experiment can be a challenge. Common mRNA isolation methods take advantage of the fact that most mRNA's have a poly-adenine (poly(A)) tail. These poly(A)$^+$ mRNA's can be purified by capturing them using complementary oligo(dT) molecules bound to a solid support. Captured mRNA's are still difficult to work with because they are prone to being destroyed. In order to prevent the experimental samples from being lost, they are *reverse-transcribed* back into more stable DNA form. The products of this reaction are called **complementary DNA's** (**cDNA**'s) because their sequences are the complements of the original mRNA sequences [Lewin, 1997].

3.2 DNA Array Types

DNA arrays consist of large numbers of DNA molecules spotted in a systemic order on a substrate (such as a nylon membrane, glass slides, or silicon chip).

3.2.1 Mechanical Spotting

In a mechanical spotting approach, a prepared DNA sample (e.g. cDNA) is loaded into a spotting pin by capillary action, and a small volume is transferred to a solid surface by physical contact between the pin and the solid substrate. Fig. 3.4a). After the first spotting cycle, the pin is washed and a second sample is loaded and deposited to an adjacent address. Robotic control

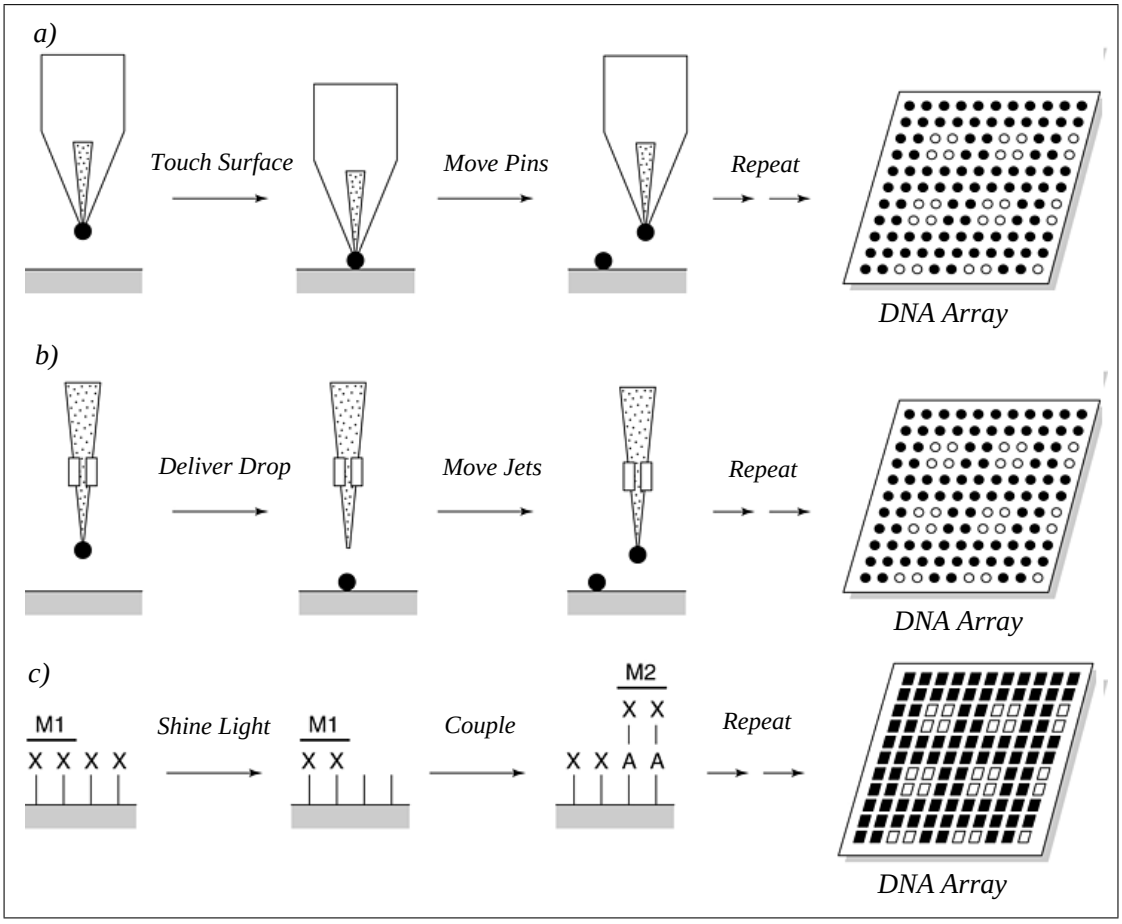


Figure 3.4. DNA Array Technologies. (a) In mechanical spotting approaches, a prepared DNA sample is loaded into a spotting pin by capillary action, and a small volume is transferred to a solid surface by physical contact between the pin and the solid substrate. (b) In ink jetting approaches, a DNA sample is loaded into a miniature nozzle equipped with a piezoelectric part which is used to expel a precise amount of liquid from the jet onto the substrate. (c) In the photolithographic process, a glass wafer modified with photolabile protecting groups is selectively activated for DNA synthesis by shining light through a photo-mask.

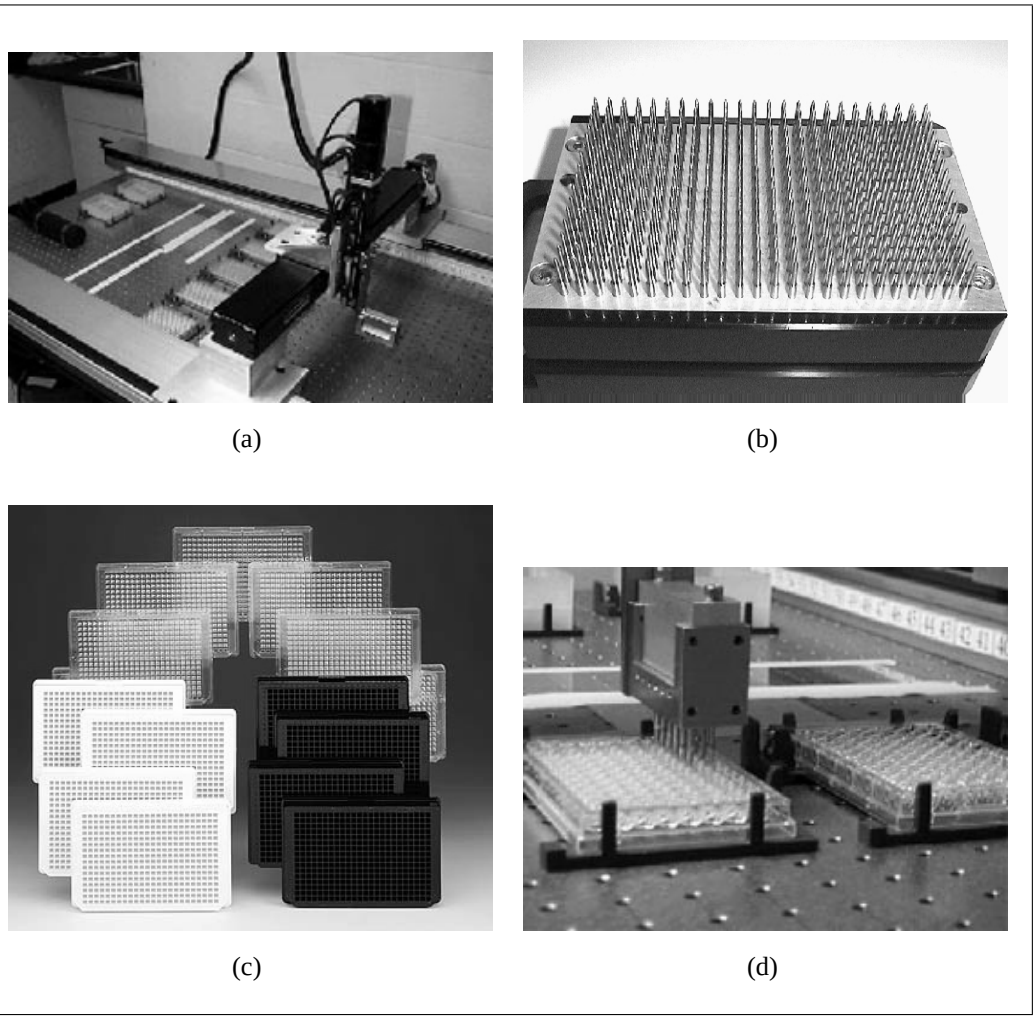


Figure 3.5. DNA Array Robot. The key component of the arrayer is the print-head containing pins witch a pitch of 4.5 mm (b and d). Samples are prepared and arrayed from micro-titer plates (c and d)

.

systems and multiplexed print heads allow automated DNA array fabrication (see Fig. 3.5). Mechanical spotting is easy to use, has low costs and is versatile. One disadvantage of mechanical spotting is that each sample must be synthesized, purified and stored prior to DNA array fabrication. The arrays currently manufactured contain 80.000 and more spots on the solid support. Nylon has a maximum resolution of 20 μm, where glass has a minimum resolution of 5 μm.

3.2.2 Ink Jetting

In this approach, a DNA sample is loaded into a miniature nozzle equipped with a piezoelectric part which is used to expel a precise amount of liquid from the jet onto the substrate (Fig. 3.4b). After the first jetting step, the jet is washed and a second sample is loaded and deposited to an adjacent address. A repeated series of cycles with multiple jets enables rapid microarray production. Ink-jetting technology has been used to prepare microarrays of single cDNA at a

density of 10,000 spots/cm^2.

3.2.3 Photolithography

The photolitography synthesis technology, developed by Fodor et al. [Fodor et al., 1991] and Pease et al. [Pease et al., 1994], combines photolithography technology from the semiconductor industry with DNA-synthetic chemistry to enable high-density oligonucleotide-microarray manufacture. In the process, a glass wafer modified with photolabile protecting group is selectively activated for DNA synthesis by shining light through a photomask (Fig. 3.4c). The wafer is then flooded with a photo-protected DNA base, resulting in spatially defined coupling on the chip surface. A second photomask is used to deprotect defined regions of wafer. Repeated deprotection and coupling cycles enable the preparation of high-density oligonucleotide DNA arrays.

A key advantage of this approach is that photo-protected versions of the four DNA building blocks allows chips to be manufactured directly from sequence databases, thereby removing the uncertain and burdensome aspects of sample handling and tracking. Another advantage of this technology is that the use of synthetic regents minimizes chip to chip variation by ensuring a high degree of precision in each coupling cycle. One disadvantage of this approach is, however the need for photomasks, which are expensive and time-consuming to design and build. At presence microarrays prepared by this approach contain as many as 400,000 groups of oligonucleotides or features in an area of around 1.6 cm2, with each feature containing approximately ten million oligonucleotides of a given sequence.

3.3 Applications of DNA Arrays

3.3.1 Gene Expression Studies

DNA microarrays are perfectly suited for *comparing* gene expression in different populations of cells. The goal of comparative cDNA hybridization is to compare gene transcription in two or more different kinds of cells. The following experiments are of particular interest.

Tissue-specific Genes

Cells from two different tissues (e.g. cardiac muscle and neuron) are specialized for performing different functions in an organism. Although we can recognize cells from different tissues by their phenotypes, it is not known just what makes one cell function as smooth muscle, another as a neuron, and still another as prostate. Ultimately, a cell's role is determined by the proteins it produces, which in turn depend on its expressed genes. Comparative hybridization experiments can reveal genes which are preferentially expressed in specific tissues. Some of these genes implement the behaviors that distinguish the cell's tissue type, while other controlling genes make sure that the cell *only* performs the functions for its type.

Regulatory Gene Defects in Cancer

Genetic disease is often caused by genes which are inappropriately transcribed – either too much or too little – or which are missing altogether. Such defects are especially common in cancers, which can occur when regulatory genes are deleted, inactivated or become permanently active [Levine, 1993]. Unlike genetic diseases in which a single defective gene is always responsible, cancers which appear clinically similar can be genetically heterogeneous. For example, prostate cancer may be caused by several different, independent regulatory gene defects even in a single patient. In a group of prostate cancer patients, every one may have different set of missing or damaged genes, with differing implications for prognosis and treatment of the disease.

Comparative hybridization can serve two purposes in studying cancer: it can reveal the transcription differences responsible for the change from normal to cancerous cells, and it can distinguish different patterns of abnormal transcription in heterogeneous cancers. Understanding the diverse basis of a cancer is crucial for inventing therapies targeted to the different varieties of the disease, so that each patient receives the most appropriate and effective treatment [DeRisi et al., 1996] .

Cellular Responses to the Environment

Cells survive in the face of changes in temperature and pH, changing nutrient availability and the presence of environmental toxins. Usually, a change in environment requires that expression of some genes be turned up or down so that the organism can respond appropriately. Comparative hybridization experiments can point out genes whose transcription changes in response to an environmental stimulus. In the simplest experiment, a population of cells is subjected to the stimulus and allowed to reach a steady state of transcription. Transcription levels in the altered cells can then be compared to those in a control population. A more informative experiment subjects cells to a change, then takes samples of the cell population at successive points in time. In this way, the experimenter can watch as the gene transcription patterns change from the old to the new steady state. This **expression profiling** can identify not only genes whose transcription changes but also the order of the changes, providing evidence about which genes control the response directly and which are only indirectly affected by it [Duggan et al., 1999].

3.3.2 Genomic Studies

The applications of arrays to genomic studies primarily involve identification and genotyping of genetic variations. Photolithographic oligonucleotide microarrays have largely been used for identification of novel DNA variants ([Chee et al., 1996].) With the ability to perform custom synthesis at high density, one can construct a tiling array to scan a target sequence for mutations. Each overlapping molecule consisting of 25 bases in the sequence is covered by four complementary oligonucleotide probes that differ only by having A, T, C or G substituted at the central position. An amplified product containing the expected sequence will hybridize best to the expect probe, whereas a sequence variation will typically alter the hybridization pattern. Such tiling arrays have been used to detect variants in such targets as the HIV genome [Kozal et al., 1996].

3.3.3 Protein Arrays

Protein Arrays containing spotted proteins (instead of DNA molecules) are the latest technical development. The objective is a functional study of thousands of proteins in parallel. MacBeath and Schreiber [MacBeath and Schreiber, 2000] printed more than 10.000 protein spots on a glass slide. The chip was used to identify protein-protein and protein-drug interactions.

3.4 Main Steps in a DNA Array Experiment

Typically, there are three steps in performing a DNA microarray experiment: Array Preparation, Array Experiment, and Array Analysis.

3.4.1 Array Preparation

DNA arrays are made from a collection of purified DNA's. DNA samples are prepared from the cells or tissues of interest. A drop of each type of DNA in solution (**probe**) is placed onto a specially-prepared glass slide or nylon array by the arraying machine. The choice of DNA's to be used in the spots on a DNA array determines which genes can be detected in a comparative hybridization experiment. For organisms whose genomes have been completely sequenced, including several bacteria, the yeast S. cerevisiae [Goffeau et al., 1996] and the human genome [Venter et al., 2001], [Lander, 2001],[Deloukas et al., 2001], it is possible to array genomic DNA from every known gene in the organism. Each gene is amplified from total genomic DNA by polymerase chain reaction (PCR) [Lewin, 1997], [Mullis, 1987], producing enough DNA to make unlimited numbers of arrays.

Another way to produce arrayable DNA even for unknown genes is to use amplified clones from cDNA libraries. A **clone** is defined as large number of cells or molecules all identical with an original ancestral cell or molecule. Alternatively, one can synthesize oligonucleotides directly from known expressed sequence information such as expressed sequence tags (EST's) [Boguski et al., 1993]. While neither of these methods will produce DNA's for every (human) gene, both can yield enough different expressed sequences to make substantial arrays.

3.4.2 Array Experiment

The array approach is based on the hybridization of the DNA probe sequences (cDNA or oligonucleotides) to a complex **target** of reversely transcribed cDNA representations of total RNA pools from test and reference cells. Once the cDNA probes have been hybridized to the array and any loose probe has been washed off, the array must be scanned to determine how much of each probe is bound to each spot. In order to detect hybridized cDNA's bound to the DNA array, one must **label** them with a reporter molecule that identifies their presence. Possible reporters include fluorescent dyes and radioactive phosphorus.

Fluorescent Labels

If light is shined on fluorescent reporter molecules, one observes light of a *different color* emitted from that molecule. The molecules adsorb high energy light (blue, for example). This

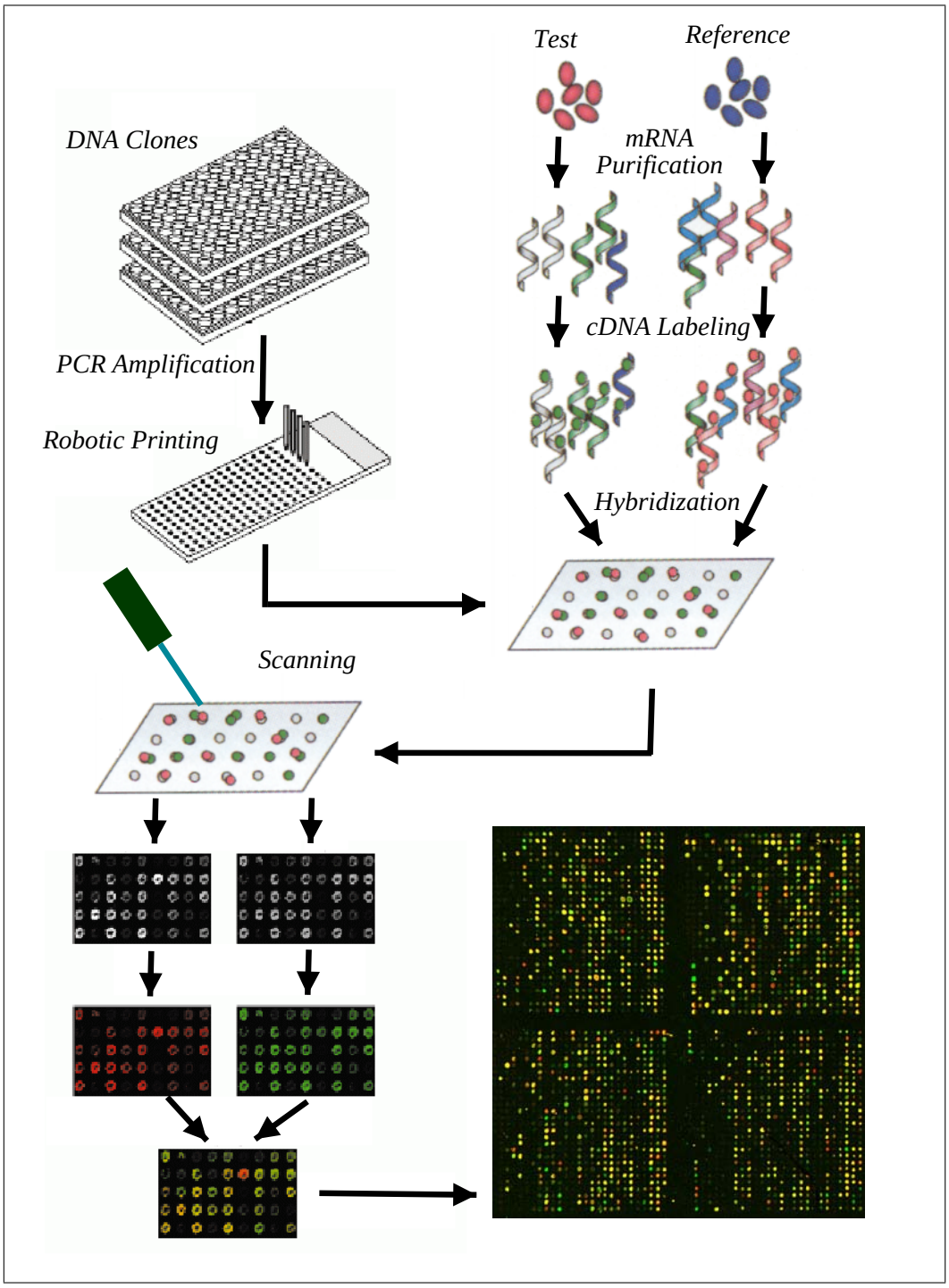


Figure 3.6. Comparative DNA experiment schema. Templates for genes of interest are obtained and amplified by Polymerase Chain Reaction (PCR) [Lewin, 1997]. The clones are printed on a solid support like glass or nylon using a computer-controlled, high-speed robot. RNA from both the test and reference sample is fluorescently labelled (green and red). The fluorescent targets are pooled and allowed to hybridize to the clones on the array. Laser excitation of the targets yields an emission with a characteristic spectrum, which is measured using a scanning confocal laser microscope. The monochrome images can be pseudo-colored and merged. Data from a single comparative hybridization experiment is viewed as a normalized ratio (green/red) in which significant deviations from 1 (no change) are indicative of increased (> 1) or decreased (< 1) levels of gene expression relative to the reference sample.

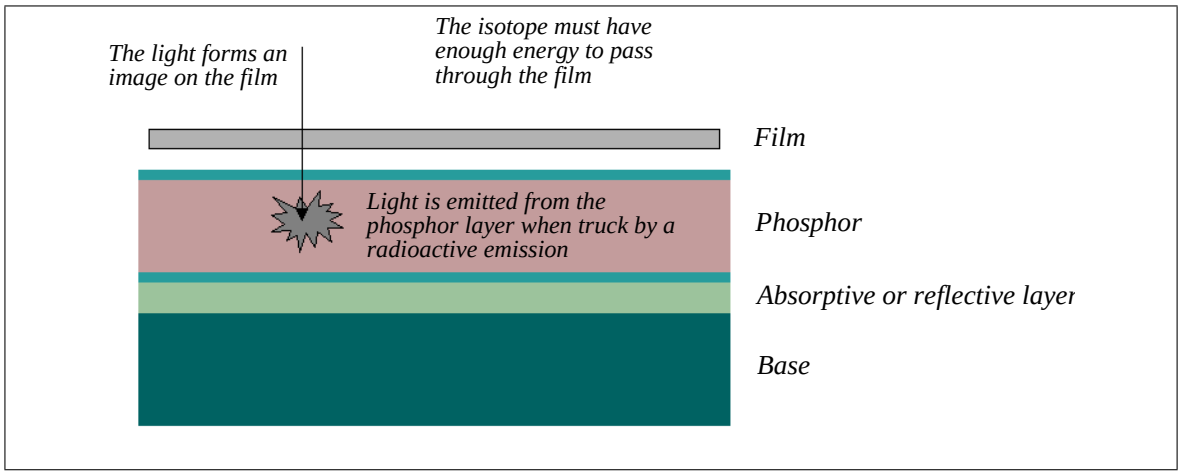


Figure 3.7. Itensifying screen for radioactive labels. Intensifying screens convert the radioactive emissions which pass through the film to visible light. Phosphors are compounds which adsorb radiation and emit visible light.

increases the energy of the molecules, and some of the energy from the blue photon is lost internally. The molecules then emit a photon with less energy, green, for example. The color of light emitted is material dependent, and likewise the excitation light wavelength depends on the material. The emitted light is captured by a detector, either a charge-coupled device (CCD) or a confocal microscope, which records its intensity. Spots with more bound probe will have more reporters and will therefore fluoresce more intensely. The advantage of fluorescent scanning is that more than one type of dye can be used. By changing the excitation light, one can cause one type of dye to fluoresce, and then another, in order to distinguish two different parts of the sample. This allows the determination of the relative amount of transcript present in the pool by the intensities of fluorescent signals generated. Relative message abundance is inherently based on a direct comparison between a test cell state and a reference cell state. Figure 3.6 illustrates the principles of a comparative hybridization experiment.

Radioactive Labels

Hybridized radioactively labeled cDNA's form spots on X-ray film or more sensitive intensifying screens (phosphorimager, Fig. 3.7). Intensifying screens convert the radioactive emissions which pass through the film to visible light. This increases speed and sensitivity of the development process. Phosphors are the substances which make intensifying screens work. They are compounds which adsorb radiation and emit visible light. Radioactivity labeling is being used in conjunction with nylon arrays. The scheme of the experiment is very similar to the one depicted in Fig. 3.6, but it is not possible to carry out simultaneous hybridization of test and reference samples. In such cases, serial or parallel hybridization is required and the probability of variability in comparisons of expression level is higher [Duggan et al., 1999]. An alternative approach to comparative gene expression analysis with radioactive labeling is *oligonucleotide fingerprinting*: Here, short oligonucleotide probes (octamers or decamers) are hybridized to the probes in order to derive a sequence dependent 'fingerprint'. This fingerprint

can identify new genes, as well as analyze their exact level of expression in different tissues [Meier-Ewert et al., 1993], [Meier-Ewert et al., 1998].

3.4.3 Array Analysis

The end product of a hybridization experiment is a scanned DNA array image. The spots provided by the array image must be correctly addressed (grid fitting) and quantified. From a computational point of view, the initial problem is to correctly map the pixel matrix of the spot array image $\mathbf{S}$ to a matrix of quantified fluorescent or radioactive spot intensies $\mathrm{I}(g_{ij})$, where g_{ij} belongs to the grid $\mathcal{G}$. In a two-color comparative hybridization experiment with a green channel G and a red channel R, the ratio

$$r_{ij} = \mathrm{I_G}(g_{ij}) \,/\mathrm{I_R}(g_{ij}) \tag{3.1}$$

of fluorescent intensities for a spot g_{ij} is interpreted as the ratio of concentrations for its corresponding mRNA in the two cell populations. Significant deviation from $r_{ij} = 1$ (no change) are indicative of increased ($r_{ij} > 1$) or decreased ($r_{ij} < 1$) levels of relative gene expression. Typically, the interpreted array data will highlight a relatively small number of spots representing very differentially-expressed mRNA's whose genes deserve further investigation.

Another way of looking at expression data from DNA arrays is to track expression levels of each gene across discrete time points $t_1, t_2, \ldots, t_k$, so that there are K measurements $\mathrm{I}(g_{ij}, t_k)$ corresponding to each gene g_{ij}. The aim is to group together genes whose expression levels exhibit similar behavior through time. Similarity indicates possible co-regulation. The observations (quantified intensity for gene g_{ij} at time t_k) can be expressed as a vector of numerical features. Each object can be represented as a point in a feature (vector) space. Statistical techniques are used to decompose the feature space into clusters. There are many publications about cluster analysis, for example [Kaufman and Rousseeuw, 1990], [Duda et al., 2000] and [Jain et al., 2000]. Recent publications about clustering techniques for gene expression patterns include [Herrero et al., 2001] and [Lukashin and Fuchs, 2001].

Interpreting the data from a DNA array experiment is challenging. Quantification of the intensities on each spot is subject to noise from irregular spots, dust on the solid support, and non-specific hybridization. Arrays printed on a stiff matrix, such as glass, render grid fitting a relatively easy task. In contrast, extraction of data from film or phosphor-image representations of radioactive hybridizations presents many difficulties for image analysis. If the array is on a membrane, there is frequently non-linear warping of the matrix on nylon membranes, which means that the observed array will not have the strict geometric regularity of an array printed on a stiff matrix, such as glass. This introduces difficulty in developing highly accurate grids to specify target locations [Duggan et al., 1999]. The following chapter describes how the general framework of spot array image analysis introduced in Chapter 2 can be adapted to DNA array images.

3.5 Chapter Summary

Alterations in gene expression pattern or in a DNA sequence can have profound effects on biological functions. These variations in gene expression are at the core of altered physiologic

and pathologic processes. DNA array technology provides rapid and cost-effective methods of identifying gene expression and genetic variations. Robotic technology is employed in the preparation of most arrays. The DNA sequences are bound to a surface such as a nylon membrane or glass slide at precisely defined locations on a grid. Using an alternative method, some arrays are produced using laser lithographic processes. DNA samples are prepared from the cells or tissue of interest. For expression analysis, the sample is cDNA, DNA copies of RNA. The DNA samples are tagged with a radioactive or fluorescent label and applied to the array. Single stranded DNA will bind to a complementary strand of DNA. At positions on the array where the immobilized DNA recognizes a complementary DNA in the sample, binding or hybridization occurs. The labeled sample DNA marks the exact positions on the array where binding occurs, allowing automatic detection. The output consists of a list of hybridization events, indicating the presence or the relative abundance of specific DNA sequences that are present in the sample.

Chapter 4

Case Study: DNA Array Image Analysis

This chapter deals with the adaption of the general framework for spot array image analysis developed in Chapter 2 to DNA array analysis. In order to cope with a broad class of DNA array images, one should take into account the following aspects of DNA array fabrication:

- As mentioned in Sect. 3.2, the pin resolution of a multiplexed print head (Fig. 3.5) is often increased on the DNA array by different spotting cycles: After one spotting cycle, the pin matrix is washed and further samples are loaded from a micro-titer plate and deposited at an adjacent location (shifted horizontally or vertically). The resulting sub-array which results from a single pin address is denoted here as a *block*. The DNA array image in Fig. 4.1 was produced by shifting the pin matrix 5 times in the horizontal direction and 5 times in the vertical direction, resulting in 5×5 block. Since it is possible that single pins are broken or bent, the blocks should be adequately represented.

- A DNA array is often divided into sub-units denoted here as *fields*. One field is spotted after the other and it may occur that the individual fields are not exactly aligned but shifted against each other. Fig. 4.1 contains a total of 6 fields, arranged in two columns and three rows. In order to successfully cope with field shifts, on should model them separately.

- If a DNA array is expected to have few hybridization events, one may introduce a sub-grid of reliable spots which should be guaranteed to always give a strong hybridization signal. Such spots are denoted as *guide spots*. The DNA array in Fig. 4.1 has a guide spot in every block center. If guide spots are present, the grid fitting should be performed first on the guide spot grid.

Further aspects of DNA array image analysis are discussed in the following section, which gives an overview of the state of the art. The general framework introduced in Chapter 2 is then adapted to the DNA array images. Section 4.2 describes the robust grid fitting and Sect. 4.3 described the robust spot fitting.

4.1 State of the Art

Several DNA microarray image analysis implementations have been described, both on rigid slides and on flexible membranes. Below, we review some essential features of these image

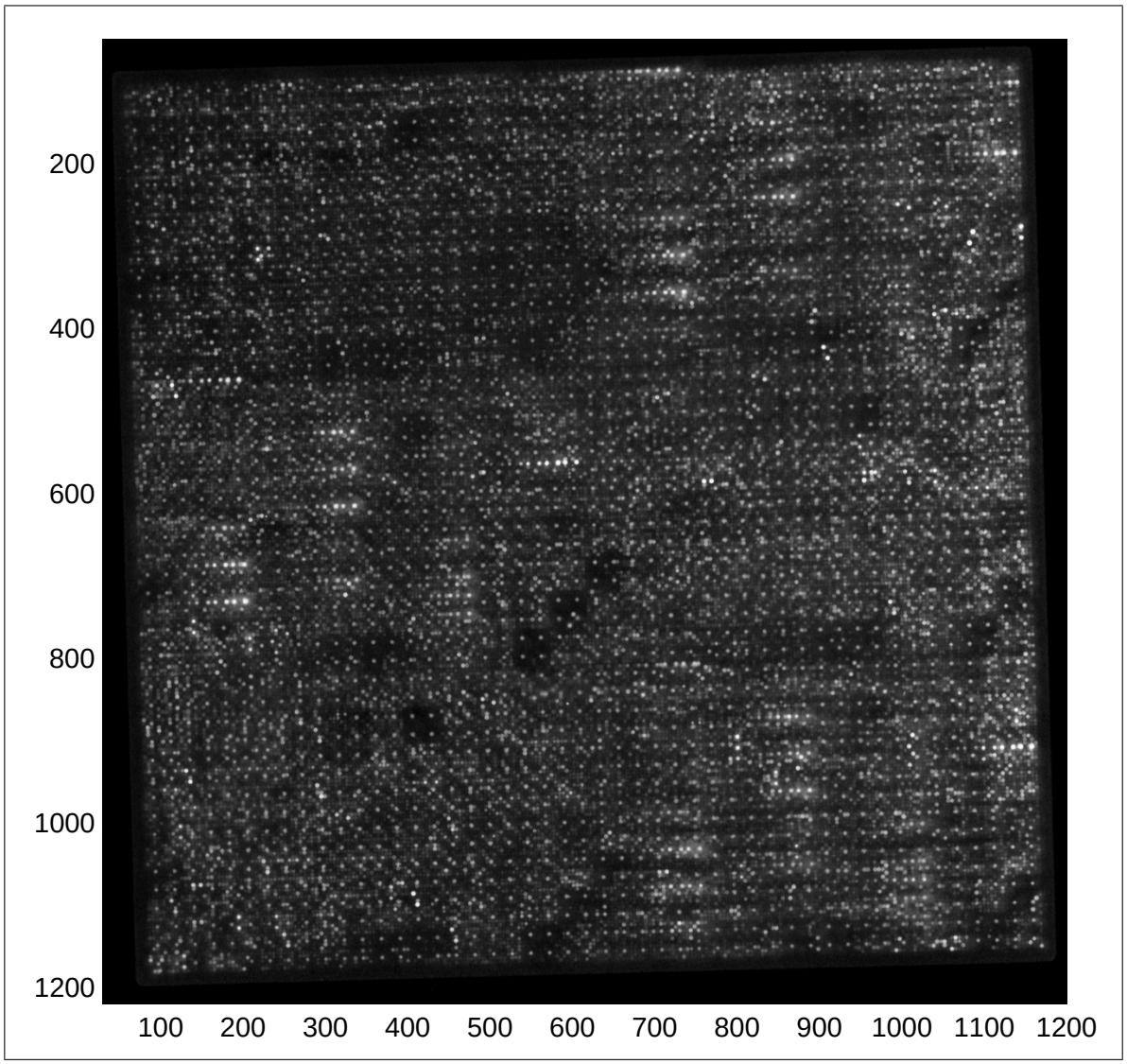


Figure 4.1. Oligonucleotide Fingerprint (ONF) Image. The intensity of every spot corresponds to the amount of radioactive label remaining after hybridizing a liquid containing the labeled targets and subsequently washing off targets not bound to the spotted probes. The ultimate image analysis goal is to automatically assign a quantity to every spot giving information about the hybridization signal (*quantification*). For a successful quantification of the hybridization signals it is necessary to assign to every grid node an image location (grid fitting).

analysis algorithms. Other implementations may be found in commercial array analysis packages but have not been publicly documented.

4.1.1 Semi-Automatic Spot Detection and Grid Fitting

The semi-automatic grid fitting method requires some level of user interaction. This approach typically uses algorithms for automatically adjusting the location of the grid lines or individual

grid points after the user has specified the approximate location of the grid. What the user needs to do is to tell the program where the outline of the grid is in the image. For example, the user may need to put down a grid and adjust the size of it to fit on the array of the spots, or to tell the program the location of the corners of the fields in the images. Then the spot finding algorithm adjusts the location of the grid lines, or grid points, to locate the arrayed spots in the image. User interface tools are usually provided by the software to allow for manual adjustment of the grid points if the automatic spot finding method has not correctly identified each spot. This section only reviews methods that include at least some sort of spot detection decoupled from spot characterization. Spot characterization methods will be reviewd in Sect. 4.1.3.

Granjeaud et al. implemented the HDG program [Granjeaud et al., 1996] to quantify radio-labeled arrays spotted on nylon membranes. HDG identifies candidate spots by tracing their edges. HDG does not use the geometry of the array to direct its search for spots; candidate spots may be found anywhere on the image in any arrangement. Only after this search are spots filtered based on a template of their expected positions, which may be warped interactively by the user to fit the image. Clearly, the edge-based segmentation will have problems to cope with artifacts and overlapping spots.

The DeArray package, described by Appel et al. [Appel et al., 1997], was developed for arrays on rigid glass slides. These slides do not suffer the distortions of membranes, so the program can safely divide an array image into rectilinear local target regions containing one spot and process each target region separately.

The Dapple technique described in [Buhler et al., 2000] adjusts initial grid locations by the center of intensity of the target regions, a mean value of all intensities above the median of the target region weighted by the image coordinates. The spots are then detected by convolving the image with a Laplacian (second derivative) filter detecting spot edges. No information about the indexing is given, it is therefore assumed that the grid is not distorted.

Hartelius [Hartelius, 1996] describes the spot array image generation process with the help of three mathematical models. First, a *point process* describes the relative spot location on the solid support as a Markov Random Field (MRF) [Winkler, 1995]. This means that the coordinates of every spot location are random variables which depend statistically on the coordinates of the neighbored locations. The deviation of an expected mean distance between grid nodes is given by a normal distribution with a given variance. Second, a *line process* describes the random deviation of the mean node distance in a local area of the solid support from a constant value. The line process can model nonlinear shrinking of a membrane an discontinuous node distances at field borders. The distance between two nodes is described as a random variable and assigned to a line connecting the node (hence the term line process). Since neighbored variables are considered statistically dependent, the line process is a MRF as well. Finally, Hartelius uses a *observational model* to describe the image intensity given the node location. The probability of the realization of the intensity (quantification) in a small neighborhood (e.g. 3×3 pixel) depends on the mean intensity of this neighborhood. The likelihood values of the point process, the line process and the observational model are combined in a global target function and the maximum with respect to the parameters is found with the help of Simulated Annealing [Winkler, 1995]. The found optimal parameters define the grid position of the given image and the intensity of the spots. The grid model of Hartelius is suited for problems in which the object is described by many statistically dependent variables, where no information about the analytical dependence is available. The parameters of the probability distribution must therefore be

empirically estimated. For grid fitting it is necessary to optimize a cost function with respect to all variables, where the optimization problem may become very complicated. For grid models in microarray images, however, knowledge about analytical correlation between different model parameters is availabe and should be taken into account.

4.1.2 Automatic Spot Detection and Grid Fitting

An automatic grid fitting algorithm with no user interaction is given in the description of the GLEAMS package [Zhou et al., 2001]. The authors do not take into account prior knowledge about expected pixel distance and estimate the inter-node distance with the help of the auto-correlation function of the spot array image. The auto-correlation function has a maximum value at the image center which shows that (obviously) an image is most similar to itself when it has not been shifted. For periodic functions like (full) spot arrays, further maxima appear for a multiple of this period. The distance from the image center to the first peak in the correlation function corresponds to the (constant) inter-node distance in x-direction. A template image containing a grid of Gaussian spots with the computed distance is then correlated with the original spot image. They place the origin of the template at its geometric center, then each local maximum points represents a possible location for the geometrical center of the array. The number of such possible locations can be significantly reduced if maxima are discarded which are not the abolute maximum in an area of a few grid nodes around. This automatic grid fitting approach has the drawback of not taking into the account the rotation of the grid. The rotation could be theoretically extracted from the discrete Fourier transform of the spot array images. The authors have, however, not found a reliable way of uniquely identifying the peak in the frequency spectrum resulting from the grid nodes. Furthermore, the computational costs of the algorithm appear to be relatively high due to the autoccorrelation and the template matching with a template as large as the spot array itself. Finally, the algorithm still needs prior knowledge about the dimensions of the spot array. Information about pin distance and scanner resolution are normally stored at the same place as the grid dimension, and it appears questionable that the grid distance computed by auto-correlation will significantly differ from the pre-comuted distance.

Bergemann et al. [Bergemann et al., 2001] finds a grid based on robust marginal distributions of the image, similar to the projections described in Sect. 2.3 (they use the median of the image rows and columns). The first projection value above the 65% percentile of the projections is defined as the upper left hand corner, and the remaining locations are searched with the prior knowlage about the spot distance. Their approach has some similarities with the approach described in this work, but clearly fails to cope with rotated grids.

Steinfath et al. [Steinfath et al., 2001] describe an automatic grid fitting algorithm divided into three steps. The preprocessing step convolves the spot array image with a matched filter and sets the matched filter response C to zero if $C < 0.6$ and to $2^N C$ otherwise, where N is the radiometric resolution in bits. The preprocessing step also includes a rotation estimation which is selected as the median of four angles between the image coordinates and four estimated straight lines representing the filter borders. The image is then rotated by the negative rotation angle to reverse the filter rotation. The automatic corner detection step uses the intersection points of the four estimated straight lines and shifts them half a block size. The final spot finding step first transforms the rotated matched filter output to a unit square by the inverse perspective trans-

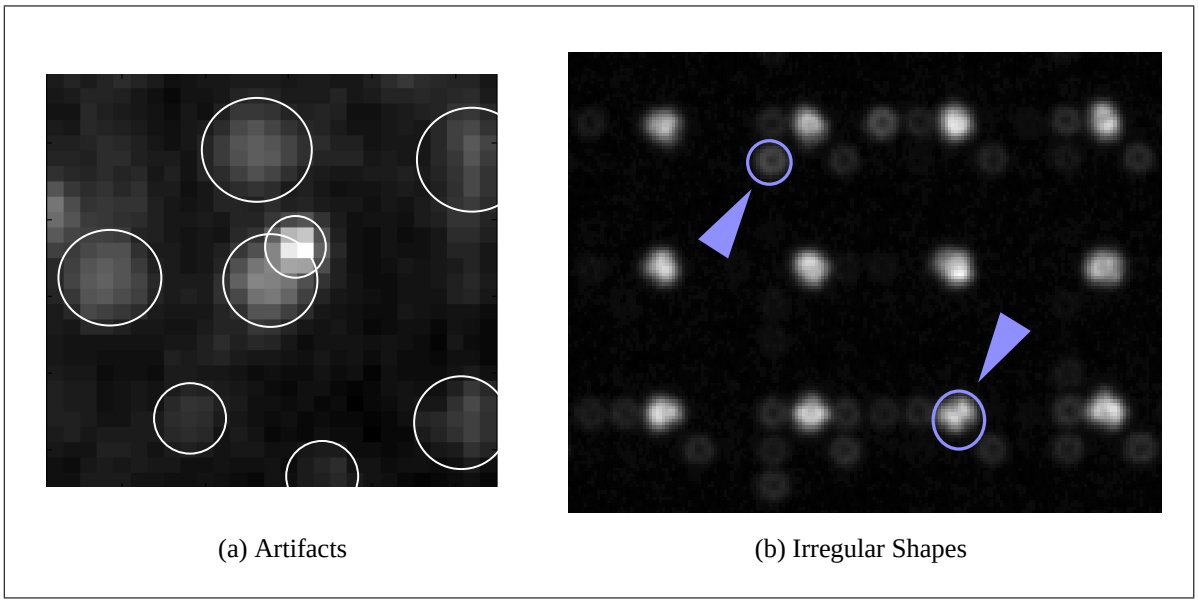
formation defined by the quadriliteral of the four estimated straight lines representing the filter borders. The unit square is discretized into small rectangles the intensities of which are determined by their shape and intensities in the original quadriliteral. The rows and colums are then further coarsened by an empiciral value in order to compute horizontal and vertical projection values. They then find a consistent set of local maxima and fit straight lines to the grid nodes in order to be able to assign locations of lacking spots. They report that the algorithm performed successfully on a set of 2000 images. While their grid fitting approach apparently works, the following points are problematic with their method, especially with regard to the subsequent spot quantification. Transforming the spot array to a rectangular grid might facilitate subsequent grid finding and quantification algorithms, but also means loss of information. In general, any image transform requires intensity *interpolation* in order to determine the intensity values of the pixels. This will unnecessarily affect the accuracy of the spot quantification. They even transform the image twice, where the parameters do not seem to be well-founded. The choice of the median of the four angles to the image border seems arbitary. Even more problematic seems to be to assume a perspective transform of a scanned image, involving significant orientations into the third dimension.

4.1.3 Segmentation Methods

After the spot location is determined in the image, a small patch around that location (target region) can be used to quantify the spot intensity level. Many approaches segment the target region into signal and background. At this stage, size and shape irregularities of the spots and any artifact problem in the images are the major concerns to the algorithm design. A number of solutions have been provided with different levels of sophistication. Their advantages and disadvantages are described below.

Circle and Ellipse segmentation

Fixed circle segmentation fits a circle with a constant diameter to all the spots in the image, in order to refine the regular grid positions of the grid provided by the user. In adaptive circle segmentation, the circle's diameter is estimated separately for each spot. The fitted circular masks separate the signal from background. It is assumed that the pixels inside the circle are due to the true signal and that those outside are background. Spot intensity measurements are then performed on these classified pixels. These types of methods are optimal when the spot shapes are close to perfect circles and no contamination is present. However, when shape irregularities occur, the accuracy of the measurements is largely compromised. In addition, spot contamination is still an issue in many DNA array images. An example is shown in Fig. 4.2a, where a small bright contamination heavily affects the measurements of the spot intensity and background. In addition to quantification inaccuracies generated due to the presence of artifacts, problems arise due to variation in shapes of the spots. There are two broad classes of spot shapes that will cause problems with pure space-based circle segmentation (Fig. 4.2b). First, doughnut-shaped spots, which are often seen with many arraying systems, contain many non-hybridized pixels within the circular spot area. These pixels will be mistakenly classified as signal pixels. Second, non-circular spots (e.g. more elliptical spots) cannot be fit perfectly with a round circle, thus causing some signal pixels to be considered as background and vice versa.



(a) Artifacts (b) Irregular Shapes

Figure 4.2. Circle segmentation. All the pixels inside the circles are considered signal according to this method. When contaminants exist (smallest circle), measurement errors are introduced into both signal and background values.

The circle segmentation method is implemented in many spot array image analysis software packages (e.g. [Eisen, 1999], GenePix[Axon Instruments, 1999]), where information about the used circle fitting algorithms is not available. Only Kegelmeyer et al. [Kegelmeyer et al., 2001] state that they use a circular Hough transform to find the best fit circle.

Bergemann et al. [Bergemann et al., 2001] describe a method for ellipse fitting based on the marginal distributions of the target regions. It is the same marginal distribution as used in their grid fitting step, i.e. the median value of the pixel rows and columns. The values of the marginal distributions above an empirically chosen threshold define the principal axes of an ellipse aligned with the image coordinates. While this method increases flexibility in of the spot shape, elliptical spots not oriented with the image coordinates will still lead to inaccurate results.

Purely intensity-based histogram segmentation

This type of method uses a target mask which is chosen to be larger than any spot. For each spot, foreground and background intensity estimates are determined in some fashion from the histogram of pixel values for pixels within the masked area: It is assumed that the brightness intensities of pixels are statistically higher than that of the background pixels.

Kaifel et al. [Kaifel et al., 2000] provide a purely intensity based segmentation approach in which the histogram of the target mask is approximated by step functions with different step widths, resulting in a quantization of the histogram at different scales. The different approximating step functions are checked for stable minima, rated by the width and depth of the surrounding minimum region. The set of stable minima is reduced by minima that appear to be caused by noise within (rather than between) background and foreground. Therefore, they em-

pirically set the background to 20% of the width of the histogram and remove all stable minima in this area. The same procedure is performed for the high end of the histogram. Of the remaining minima, the one with the best score is chosen as the histogram threshold. Kegelmeyer et al. [Kegelmeyer et al., 2001] find the low-intensity peak of the histogram representing the background values and fit a line to the falling slope on the right side, forming a triangle with the y-axis [Ballard, 1981]. The intensity threshold is where the line crosses the x-axis.

When the signal intensity is low, the intensity distribution of the signal overlaps largely with background. The signal and background pixels are not separable based on their intensity values alone. Applying purely intensity-based histogram methods will produce biased estimates of the signal and background intensity values. This shortcoming can be remedied by exploiting spatial information.

Histogram Segmentation with Spatial Information

Chen et al. [Chen et al., 1997] use a nonparametric statistical method, the Mann-Whitney test, to segment out signal pixels from target regions. After the grid fitting, a circle is placed in the target region to demarcate the spatial region from the spot. Because the pixels outside the circle are assumed to be background, the statistical properties of these background pixels can be used to determine which pixels inside the circle are signal pixels. The Mann-Whitney test is the method used to obtain a threshold intensity level. Pixels inside the circle that exceed the threshold intensity are identified as signal. This method works well when the spot location is found correctly and when there is no contamination in the image. However, when contaminated pixels exist inside of the circle, they are incorrectly scored as signal pixels. If there are contaminated pixels outside the circle or if the spot location is not found correctly such that some of the signal pixels are outside the circle, these high-intensity pixels will raise the intensity threshold level. Signal pixels with intensities lower than the threshold will be incorrectly scored as background. This method also has its limitations when dealing with weak signals and noisy images. When the intensity distribution functions of the signal and background are largely overlapping, classification of pixels based on an intensity threshold is prone to errors, resulting in measurement biases. The Mann-Whitney test was also used in [Appel et al., 1997].

Trimmed measurement [Zhou et al., 2000] is another method that combines both spatial and intensity information in segmenting signal pixels from background. The logic of this method proceeds as follows: After the spot is localized and a target circle is placed in the target region, *most* of the pixels inside the circle are signal pixels, and *most* of the background pixels are outside the circle. Due to shape irregularities, some signal pixels may lie outside the circle, and some background pixels may lie inside the circle. These pixels are considered outliers in the sampling of the signal and background pixels. Similarly, contaminant pixels can also be considered outliers in the intensity domain. These outlier will severely change the measurement of the mean and total signal intensity. To remove the impact of the outliers on these measurements, one may simply trim off a certain percentage of signal and background pixels. Typically, a certain percentage of pixels from the high end of the intensity distribution, for example 10%, is trimmed off from the pixels inside the circle, due to the high possibility that such pixels may be contaminants. A certain percentage of pixels at the low end of the intensity distribution, for example 20%, is trimmed off from the pixels inside the circle, due to the high possibility that such pixels may be background. Whatever remains inside of the circle after trimming is used

for quantification. The exact amount to be trimmed depends on the effectiveness of the grid fitting process and the quality of the image, such as the extent of size and shape of irregularities. A good estimate for these thresholds is determined empirically. This method is effective as long as the image quality does not significantly change. The major advantage of the trimmed measurement method is the robustness of the measurement against outliers, at the expense of accuracy: When there are no outliers, trimming will reduce the measurement accuracy by reducing the number of pixels used in the calculation of the mean signal value. However, if there is a significant number of pixels per spot and only a small percentage of the pixels are removed, this method can yield highly accurate values.

The GLEAMS approach [Zhou et al., 2001] computes a threshold which is constrained to fall within a range determined from an estimate of the local background's mean and variance [Otsu, 1979]. The background estimation uses pixels that fall outside fitted circles within an area of 5×5 target masks centered around the target mask in question. Clearly, a Gaussian distribution of the background is assumed, which may not be the case, especially for spots with low intensity.

Adaptive shape segmentation

Segmentations of this class include region-based segmentation methods like watersheds and Seeded Region Growing (SRG), which both have been introduced in Sect. 1.3.2. As for SRG, the seed locations for the spots are provided are by the grid fitting procedure. Yang et al. [Yang et al., 2000] choose a spot seed location from the intersections of the horizontal and vertical grid lines of the fitted spot grid. It is possible, particularly when the spot is small, that this intersection location may not be inside the spot because of local irregularities or errors in the grid fitting. To overcome this problem, they chose a squared seed region centered at the local maximum pixel in a small neighborhood. Background seeds are cross-shaped regions from the fitted background grid. A drawback of SRG is that is cannot cope with overlapping spots, since two overlapping spots will be merged into a single region.

4.1.4 Data Quantification

On a single DNA array for expression profiling, the expression levels of many genes are in parallel. Under the proper conditions, the total fluorescence intensity or radioactive intensity of a spot is proportional to the expression level of a gene. These conditions are as follows [Zhou et al., 2000]:

1. The preparation of the probe cDNA (through reverse transcription of the extracted mRNA) solution is performed such that the probe cDNA concentration in the solution is proportional to the mRNA in the tissue.

2. The hybridization experiment is performed such that the amount of cDNA binding to each spot is proportional to the partial concentration of each cDNA species in the probe solution.

3. The amount of cDNA target deposited at each spot during the array fabrication is constant and in approximately 10-fold excess relative to the most abundant species in the probe solution.

4. There is no contamination on the spots.

5. The signal pixels are correctly identified by the grid fitting.

In the following discussion of existing quantification methods, it is assumed that conditions 1 and 2 are satisfied. Whether these two conditions are truly satisfied is determined through the design of the experiments. For the quantification measurements, the more closely conditions 1 to 5 are followed the better. Often, conditions 3,4, and 5 are violated to varying degrees. The DNA concentrations in the spotting procedure may vary from time to time and spot to spot. Higher or lower concentrations may result in altered signals. When adjacent spots overlap, the signal intensity corresponding to the contaminated region is not measurable in a direct manner. The grid fitting may not correctly identify all the signal pixels; The quantification methods should therefore be designed to address these problems. The commonly used methods are total, mean, median, mode, volumen, intensity ratio and the correlation ratio across two channels.

Total

The total signal intensity is the sum of the intensity values of all the pixels in the signal region. Total intensity is sensitive to variations in the amount of DNA deposited on the surface and the existence of contamination. Because these problems occur frequently, this measurement may not be accurate.

Mean

The mean signal intensity is the average intensity of the signal pixels. This method has certain advantages over the total. The spot size correlates very often with the samples and pins used in the arraying step. Measuring the mean will reduce the error caused by the variation of the amount of DNA deposited on the spot.

Median

The median of the signal intensity is the intensity value that splits the distribution of the signal pixels such that the number of pixels above the median intensity is the same as the number below the median intensity. The advantage of choosing this measurements derives from the resistance of the median value to outliers. An alternative to the median measurement is to use a trimmed, as was discussed in the section above. The trimmed mean estimate is obtained by trimming a certain percentage of pixels from the high- and low-intensity sides of the distribution.

Mode

The mode of the signal intensity is the “most likely” intensity value and can be measured as the intensity level corresponding to the peak of the intensity histogram. It enjoys the same robustness against outliers as the median. The tradeoff is that the mode will be more unstable than the median when the distribution is multimodal. This is because the mode value will be equal to one of the modals in the distribution, depending on which is the highest.

Volume

The volume of signal intensity is the sum of the signal intensity above the background intensity. It may be computed as:
(signal mean - background mean) $\times$ signal area. This method is based on the argument that the measured signal intensity has an additive component due to the nonspecific binding, and this additive compoment is the same as that of the background.

Intensity Ratio

If the hybridization is two-color and the scanning measurements are taken in two channels, then the intensity ratio between the channels is often an important quantified value of interest. This value will be insensitive to variations in the exact amount DNA spotted since the ratio between the two channels is being measured. This ratio can be obtained from the mean, median, or mode of the intensity measurement, obtained as discussed above, for each channel.

4.2 Robust Grid Fitting

This section adapts the grid fitting tools developed in Chapter 2 to DNA array images according to the aspects outlined in the beginning of this chapter.

4.2.1 DNA Array Representation

In addition to the general grid representation defined Sect. 2.1.2, we formally divide the grid $\mathcal{G}$ into subunits to represent the nature of the spotting cycles.

A field $\mathcal{F}_{pq} \subset \mathcal{G}$ is a subunit of a grid $\mathcal{G}$ and is a set of nodes in $\{1 \dots I_{\text{F}}\} \times \{1 \dots J_{\text{F}}\}$ with I_{F} as the number of field rows and J_{F} as the number of field columns. A grid $\mathcal{G}$ is partitioned into $F_{\text{I}} * F_{\text{J}}$ fields such that

$$\bigcup_{p=1}^{F_{\text{I}}} \bigcup_{q=1}^{F_{\text{J}}} \text{L}(\mathcal{F}_{pq}, \mathbf{S}) = \text{L}(\mathcal{G}, \mathbf{S}) \tag{4.1}$$

and

$$\text{L}(\mathcal{F}_{pq}, \mathbf{S}) \cap \text{L}(\mathcal{F}_{rs}, \mathbf{S}) = \emptyset, \qquad \forall (p,q) \neq (r,s). \tag{4.2}$$

The row extraction and column extraction relations $\text{R}_{\mathcal{F}}$ and $\text{C}_{\mathcal{F}}$ are defined similarly to (2.2) and (2.3). A *block* $\mathcal{B}_{ij} \subset \mathcal{F}$ is a subunit of a field $\mathcal{F}$ and is a set of nodes in $\{1 \dots I_{\text{B}}\} \times \{1 \dots J_{\text{B}}\}$ with I_{B} as the number of block rows and J_{B} as the number of block columns. A field $\mathcal{F}$ is partitioned into $I_{\text{W}} * J_{\text{W}}$ blocks such that

$$\bigcup_{i=1}^{I_{\text{W}}} \bigcup_{j=1}^{J_{\text{W}}} \text{L}(\mathcal{B}_{ij}, \mathbf{S}) = \text{L}(\mathcal{F}, \mathbf{S}) \tag{4.3}$$

and

$$\text{L}(\mathcal{B}_{ij}, \mathbf{S}) \cap \text{L}(\mathcal{B}_{kl}, \mathbf{S}) = \emptyset \qquad \forall (i,j) \neq (k,l). \tag{4.4}$$

The dimensions of a field counted in blocks corresponds to the dimensions of the micro-titer plate. The following equations hold:

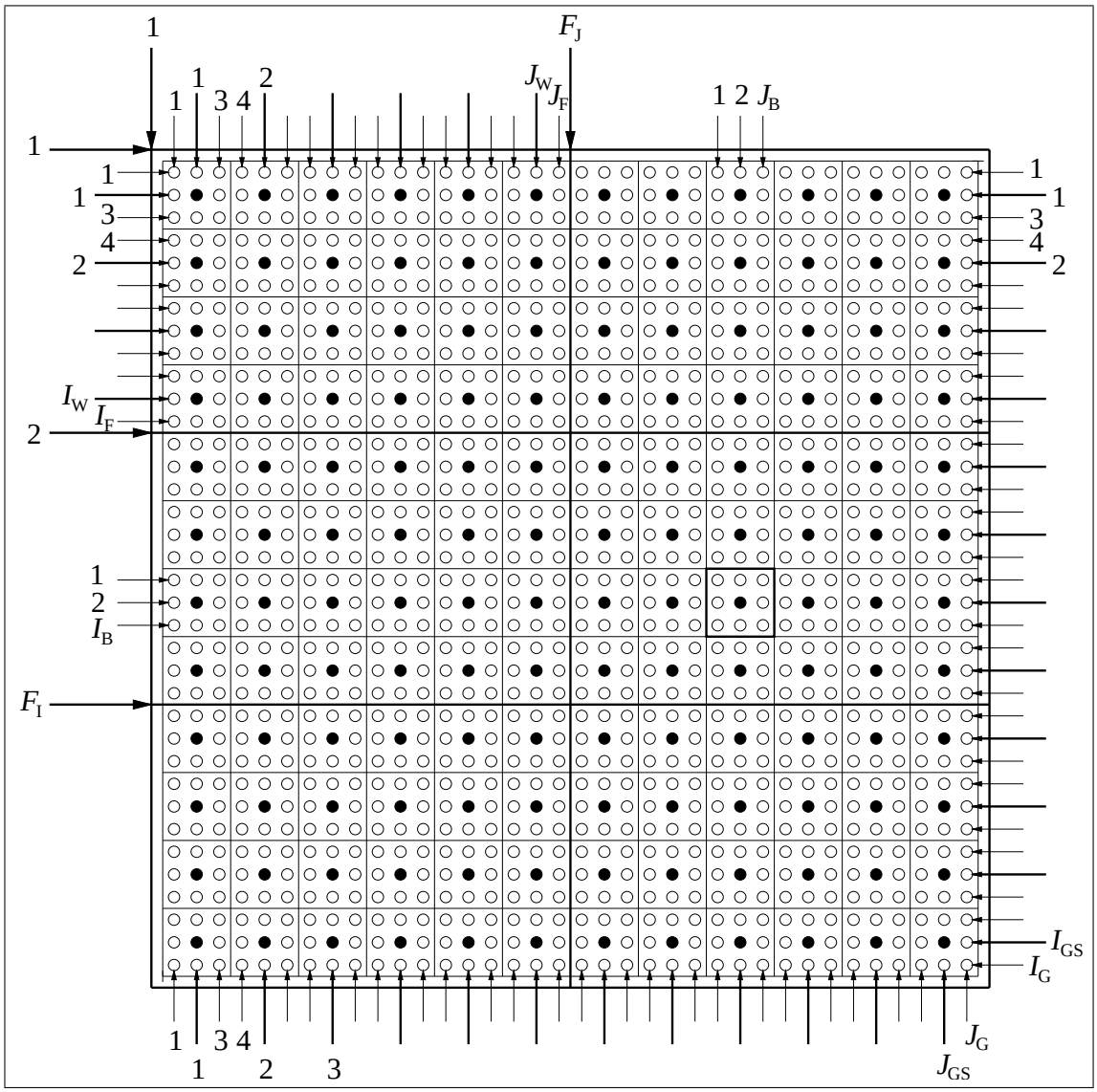


Figure 4.3. An example grid and its subunits: The grid consists of $I_G \times J_G = 36 \times 36$ spots and $F_I \times F_J = 3 \times 2$ fields. One field consists of $I_F \times J_F = 12 \times 18$ spots and of $I_W \times J_W = 4 \times 6$ blocks. One block consists of $I_B \times J_B = 3 \times 3$ spots. Guide spots are centered in the blocks and marked black. The guide spots define a $I_{GS} \times J_{GS} = 12 \times 12$ guide spot grid.

$$I_F = I_W * I_B \qquad \text{and} \qquad J_F = J_W * J_B. \tag{4.5}$$

A *guide spot grid* $\mathcal{G}^\star$ is a set of nodes in $\{1 \dots I_{GS}\} \times \{1 \dots J_{GS}\}$, with I_{GS} as the number of guide spot grid rows and J_{GS} as the number of guide spot grid columns, where

$$I_{GS} = F_I * I_W \qquad \text{and} \qquad J_{GS} = F_J * J_W \tag{4.6}$$

The row extraction and column extraction relations $R_{\mathcal{G}^\star}$ and $C_{\mathcal{G}^\star}$ are defined similarly to (2.2) and (2.3).

Not.	Description	Not.	Description
I_{B}	# block rows	J_{B}	# block columns
I_{W}	# micro-titer plate rows # guide spot field rows	J_{W}	# micro-titer plate columns # guide spot field columns
I_{F}	# field rows $I_{\mathrm{F}} = I_{\mathrm{W}} * I_{\mathrm{B}}$	J_{F}	# field columns $J_{\mathrm{F}} = J_{\mathrm{W}} * J_{\mathrm{B}}$
F_{I}	# fields in vertical direction	F_{J}	# fields in horizontal direction
I_{G}	# grid rows $I_{\mathrm{G}} = F_{\mathrm{I}} * I_{\mathrm{F}}$	J_{G}	# grid columns $J_{\mathrm{G}} = F_{\mathrm{J}} * J_{\mathrm{F}}$
I_{GS}	# guide spot grid rows $I_{\mathrm{GS}} = F_{\mathrm{I}} * I_{\mathrm{W}}$	J_{GS}	# guide spot grid columns $J_{\mathrm{GS}} = F_{\mathrm{J}} * J_{\mathrm{W}}$
Δy	vertical scanner resolution	Δx	horizontal scanner resolution
N_y	vertical spot distance [mm]	N_x	horizontal spot distance [mm]
S_y	vertical spot distance (theoretical) [pixel] $S_y = N_y/\Delta y$	S_x	horizontal spot distance (theoretical) [pixel] $S_x = N_x/\Delta x$
B_y	vertical block distance (theoretical) [pixel] $B_y = S_y * I_{\mathrm{B}}$	B_x	horizontal block distance (theoretical) [pixel] $B_x = S_x * J_{\mathrm{B}}$
M	vertical spot array image size	N	horizontal spot array image size

Table 4.1. Notation. Overview of the notation of the spot array image and the spot array. The left part of the table describes "vertical" entities, the right part of the table contains "horizontal" entities.

Prior knowledge

In addition to the theoretical horizontal and vertical spot distance S_x and S_y (2.4), we introduce the *theoretical horizontal block distance* $B_x \in \mathbb{R}$ and the *theoretical vertical block distance* $B_y \in \mathbb{R}$. They are the distances in pixels between two adjacent block in the horizontal and vertical direction computed as

$$B_x = S_x * J_{\mathrm{B}} \qquad \text{and} \qquad B_y = S_y * I_{\mathrm{B}} \tag{4.7}$$

Table 4.1 provides an overview of the notation. The different dimensions of a DNA array are illustrated in Fig. 4.3

4.2.2 Spot Amplification

The first step of DNA array image analysis is the spot amplification with a matched filter as described in Sect. 2.2. Figure 4.4b is the matched filter response image $\mathbf{R}^{\mathrm{M}}$ of the image in Figure 4.4a.

If a guide spot grid is present, one can first try to amplify the guide spot locations in order to subsequently span a guide spot grid. The main idea to amplify the locations of potential guide spots is to consider the matched filter response values at the theoretical guide spot neighborhood

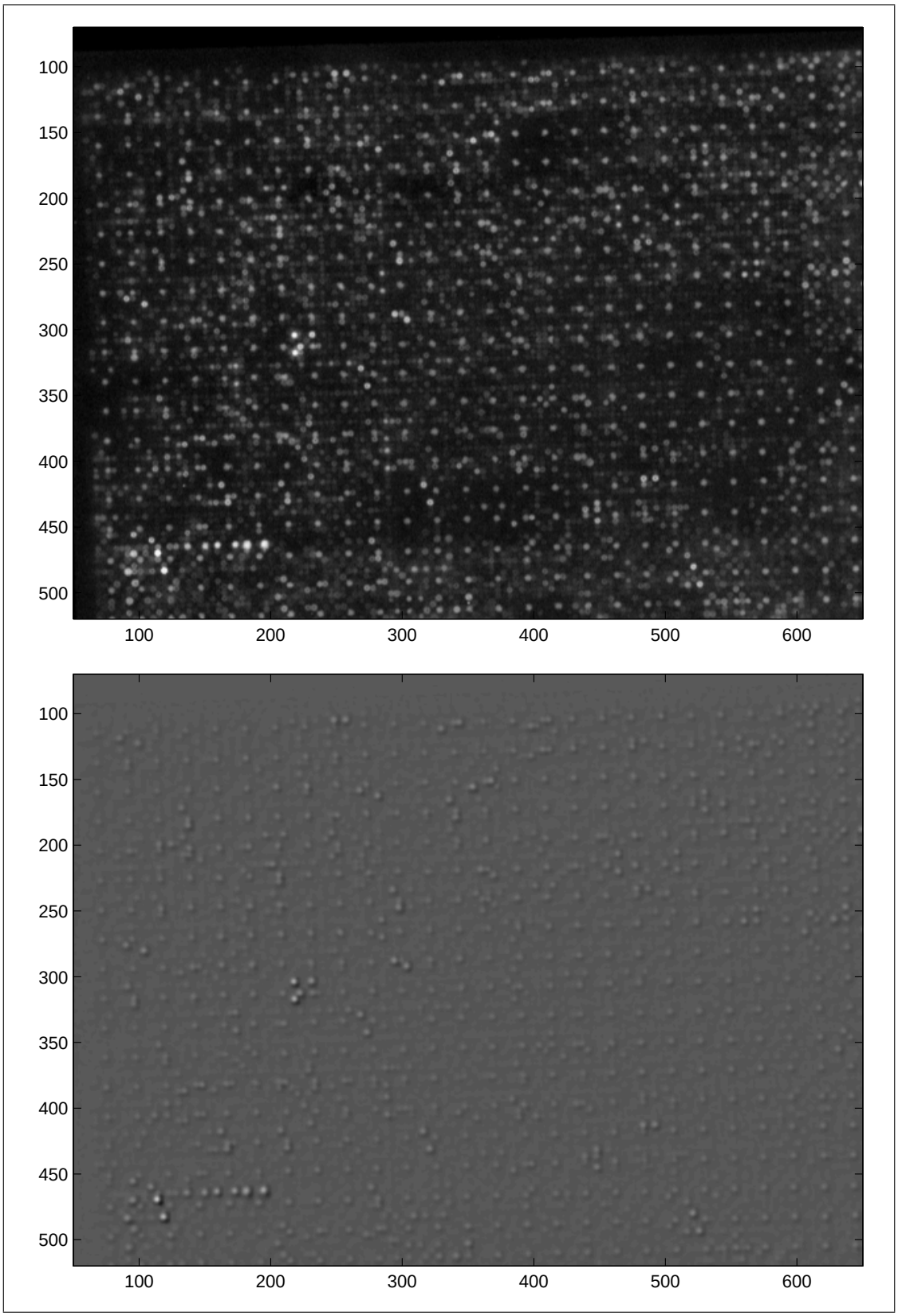


Figure 4.4. Filtering with a matched filter. A high pixel value at the matched filter response image (b) indicates a high similarity to the matched filter of the guide spots and is proportional to the probability of a guide spot location in (a). Responses of regular spots can be stronger than the guide spots responses.

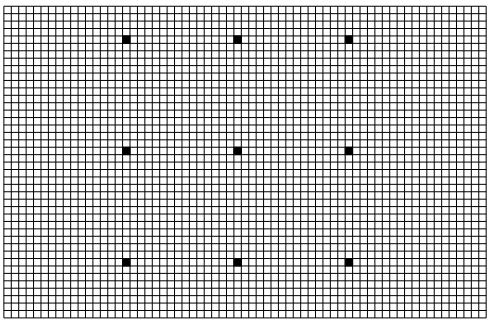


Figure 4.5. Example for guide spot location amplification (GSLA): The black pixels are considered in the computation of the GSLA response value for the center black pixel (assuming theoretical block distances of 15 pixels). The median of the intensities at the 9 locations is taken as the response value of the center pixel. If the center black pixel is a guide spot location, the grid neighborhood locations will have high response values and the median of the response values will be high.

locations. Since a guide spot location is part of a grid, its grid neighborhood locations must also have high matched filter response values. If this is not the case, it is likely that the location is not a guide spot.

We formally define the set $\mathcal{T}_{[m,n]}$ which includes the response value $\mathbf{R}^{\text{M}}[m,n]$ and the response values of the theoretical guide spot neighborhood locations of (m,n) in $\mathbf{R}^{\text{M}}$ as

$$\mathcal{T}_{[m,n]} = \{\mathbf{R}^{\text{M}}[m+k, n+l] \mid k \in \{0, \circ(B_y), \circ(-B_y\}) \wedge l \in \{0, \circ(B_x), \circ(-B_x\})\}. \quad (4.8)$$

with B_y and B_x as the theoretical block distances defined in (4.7). Figure 4.5 illustrates the neighborhood set $\mathcal{T}_{[m,n]}$ for the center black pixel assuming theoretical block distances $B_y = B_x = 15$. The GSLA response value $\mathbf{R}^{\text{A}}[m,n]$ is determined as

$$\mathbf{R}^{\text{A}}[m,n] = \text{median}(\mathcal{T}_{[m,n]}). \quad (4.9)$$

If a MF response value at a location (m,n) and all the response values in the theoretical neighborhood are high it is likely that (m,n) is a guide spot location. This is not the case for locations where the neighborhood response values are low. Note that the median value is more robust for the guide spot location amplification than the mean value. In case of a regular spot (or an artifact) with a very high MF impulse response compared to the theoretical neighbors, the mean value measure would propagate high values to the theoretical grid neighbors and generate new local grid structures. Fig. 4.6 shows the GSLA response image of the matched filter response image in Fig. 4.4b.

4.2.3 Rotation Estimation

The rotation estimation described in Sect. 2.3 can be directly applied to DNA array images. In the presence of guide spots, the only adaption one has to make refers to the set $\mathcal{M}_r$ in the median computation of (2.19): After the application of GSLA filter the number of bright spots in a row should correspond to the number J_{GS} of guide spots belonging to a row. The number J_{G} of spots in a row for the computation of $\mathcal{M}_r$ is therefore replaced by replaced by J_{GS}.

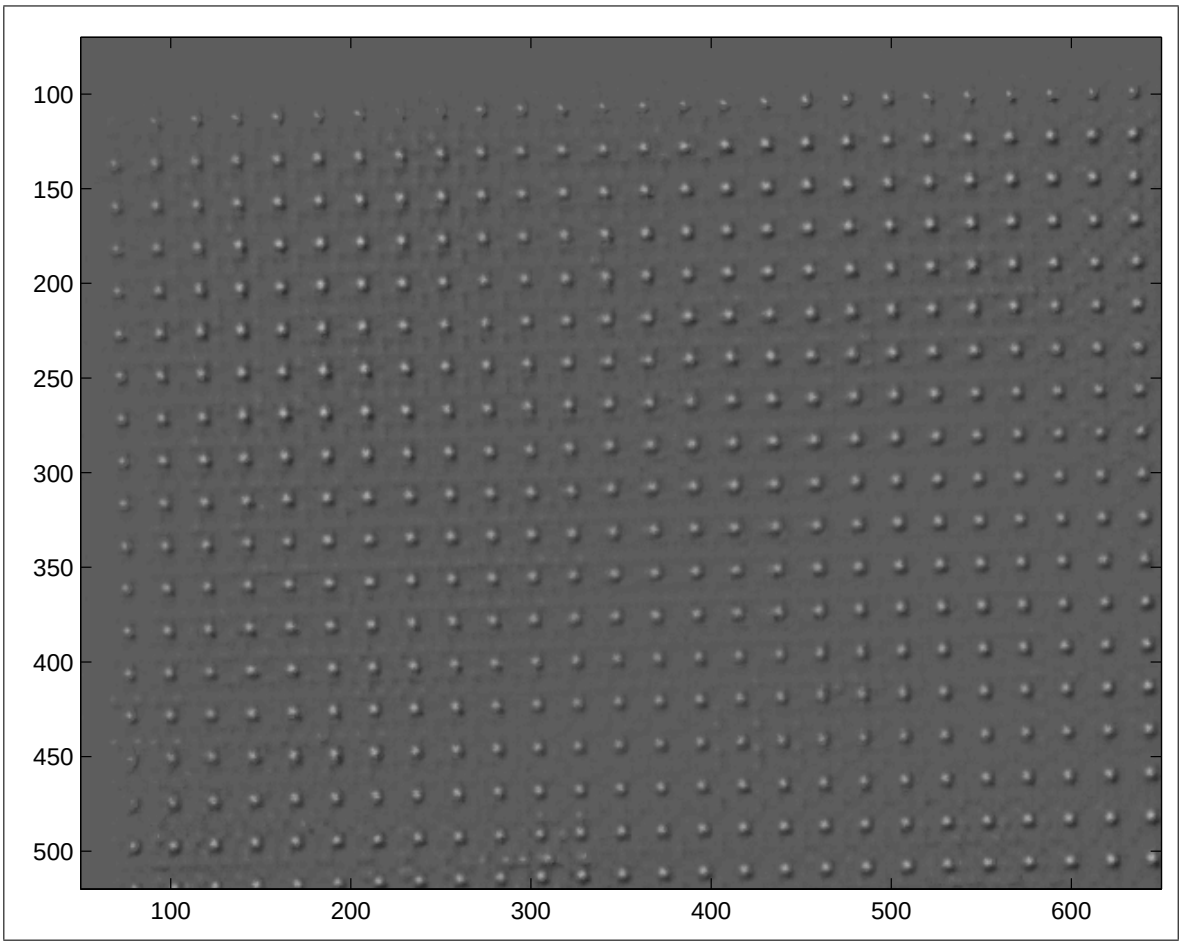


Figure 4.6. Guide spot location amplification (GSLA) filtered reponse image of the matched filter response image in Fig. 4.4b. High intensity values indicate that locations of the theoretical guide spot neighborhood also have high intensities. The guide spot location at the upper left grid corner is missing because the median of 8 neighborhood positions is computed. A corner location only has 3 guide spot neighbors.

4.2.4 Grid Spanning

In the following, the grid spanning algorithms described in Sect. 2.4 are adapted to DNA arrays, taking into account the additional subgrids.

Initial Grid after Maximum Search

If a guide spot grid is present and the GSLA filter has been applied, two adaptions to the methods described in Sect. 2.4.1 are necessary. First, the windows size for the maximum search is now the next smaller number of the theoretical *block* distance B_y rather than the theoretical spot distance S_y. This is because the regular spot locations are attenuated in the GSLA image. The second change if guide spots are present is of course the definition of a prior guide spot grid rather than a prior spot grid. The prior guide spot grid is transformed in the same way as

in (2.23).

Initial Grid before Maximum Search

If the initial grid is computed by the inverse Radon transform and the input is a GSLA image highlighting guide spot locations, the hypotheses as shown in Fig. 2.9 must be based on *guide* spot rows and column.

Spot Grid Parameterization

As described in Sect. 2.4.3, grid parameterization deals with fitting straight lines to every row and column of the grid. If a DNA array is divided into different fields, there might be a significant shift between adjacent fields. In order not to bias the fitted straight lines towards the field shift, we define parameter sets $\mathcal{P}_{pq}$ for every (guide) spot field $\mathcal{F}_{pq}$ as

$$\mathcal{P}_{pq} = \{((a_{r_i}, b_{r_i}), (a_{c_j}, b_{c_j})) \mid 1 \leq i \leq I_{\text{F}}, 1 \leq j \leq J_{\text{F}}\}, \tag{4.10}$$

with a_{r_i} and b_{r_i} as the parameters of a row straight line model (2.44) .

Initialization of regular spots

If guide spots are present in the image, the following principle is used to initialize the locations of regular non-guide spots. Their exact positions is estimated in the quantification step. Locations belonging to regular spots in a block $\mathcal{B}$ are inferred from the guide spot location with the help of the prior knowledge of the theoretical spot distances S_x and S_y and the block rotation. The rotation $\theta_{\mathcal{B}}$ of a block $\mathcal{B}$ in row i of a field $\mathcal{F}$ is given by the slope b_{r_i} of the field grid parameter set (2.43) as follows:

$$\theta_{\mathcal{B}} = \arctan(b_{r_i}). \tag{4.11}$$

Given the guide spot $(i, j) \in \mathcal{B}$ and its location $\text{L}((i, j)) = [x_{\text{G}}\ y_{\text{G}}]^T$, the regular spot locations $\text{L}((m, n))$ with $(m, n) \in \mathcal{B}$ and $(m, n) \neq (i, j)$ are initialized as follows:

$$\text{L}((m,n)) = \begin{bmatrix} x_{\text{G}} \\ y_{\text{G}} \end{bmatrix} + \begin{bmatrix} \cos\theta_{\mathcal{B}} & \sin\theta_{\mathcal{B}} \\ -\sin\theta_{\mathcal{B}} & \cos\theta_{\mathcal{B}} \end{bmatrix} \begin{bmatrix} (m-i)S_x \\ (n-j)S_y \end{bmatrix} \quad \forall (m,n) \neq (i,j). \tag{4.12}$$

Equation (4.12) is valid for blocks with guide spots residing at an arbitrary position of the block. The regular spot locations (4.12) are used as initial estimates of of the center of a parametric spot model.

4.2.5 Experimental Results

The quality of the grid fitting cannot be assessed for the high-density test images with a simple location distance measure, since *no ground-truth data* is available for the spot array images. We haven chosen two other ways to demonstrate the effectiveness of the grid fitting presented in this work. We first show five examples of image types originating from different hybridization experiments, having different quality, resolution and size. We then show that the grid fitting success is correlated with the image quality.

Visual Examples

We present five examples of images for which the grid fitting was successful. In order to demonstrate the different scanning resolutions of the images, the image parts in Fig. 4.7- 4.11 are of the same size and in scale, meaning that the spots of a high-resolution image are displayed larger than the spots of a low-resolution image.

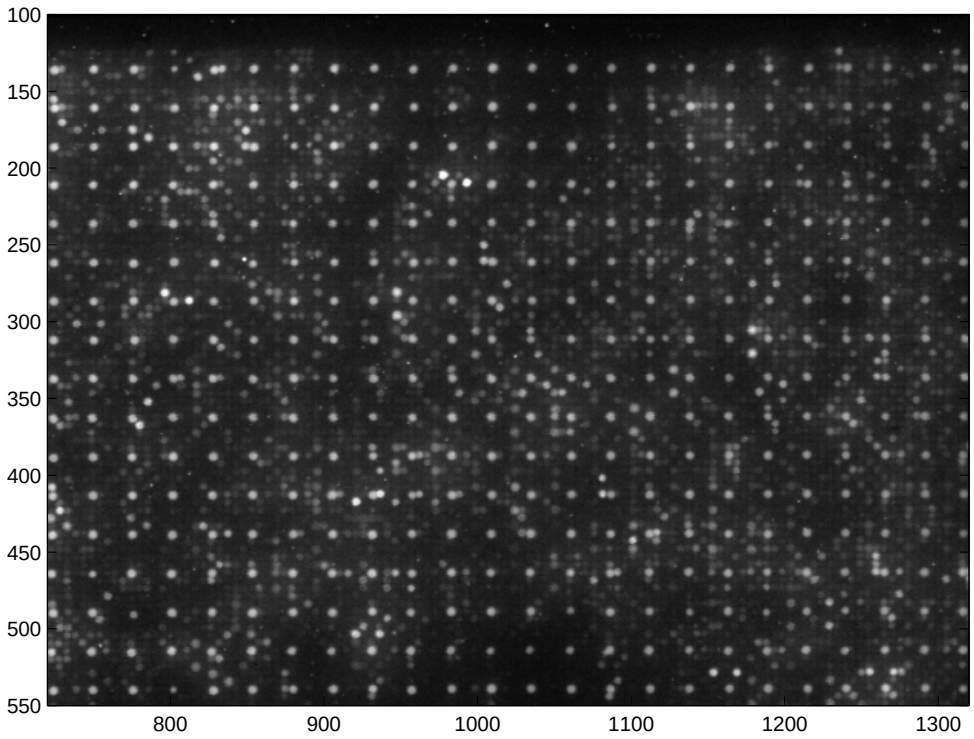
Figure 4.7a shows a part of a 1596×1482 ONF image which has been scanned at a resolution of $175\mu m$ at the Max Planck Institute for Molecular Biology (MPIMG) Berlin. The guide spots at the center of the 5×5 blocks in Fig. 4.7a are bright and clearly identifiable. Figure 4.7b shows the same image with the computed guide spot locations superimposed as cross-hairs. For the sake of overview, the initialized locations of the regular spots are not shown – they are simply derived from the guide spot locations as demonstrated in (4.12). Please note that the locations need not necessarily be right in the center of the (guide) spot: They are just the initializations for the center of a parametric spot model.

Figure 4.8a shows a part of a 1300×1200 ONF image which has been scanned at a resolution of $200\mu m$ at the Novartis Forschungsinstitut (NFI) Vienna. The signals of the hybridization signals of the guide spots at the center of the 5×5 blocks are relatively low in comparison with the signals of the hybridized regular spots, for example at the lower right corner of the filter. Furthermore, there are regions in which the signal-to-noise ratio is very low. It can be seen in Fig. 4.8b that our algorithm is able to restore the guide spot grid. In this example, the border between two fields can be noticed by comparing guide spot columns 4 and 5 of Fig. 4.8b: There is a leap in the y-coordinates indicating a field shift and therefore justifying the parameterization (2.43) of the fields.

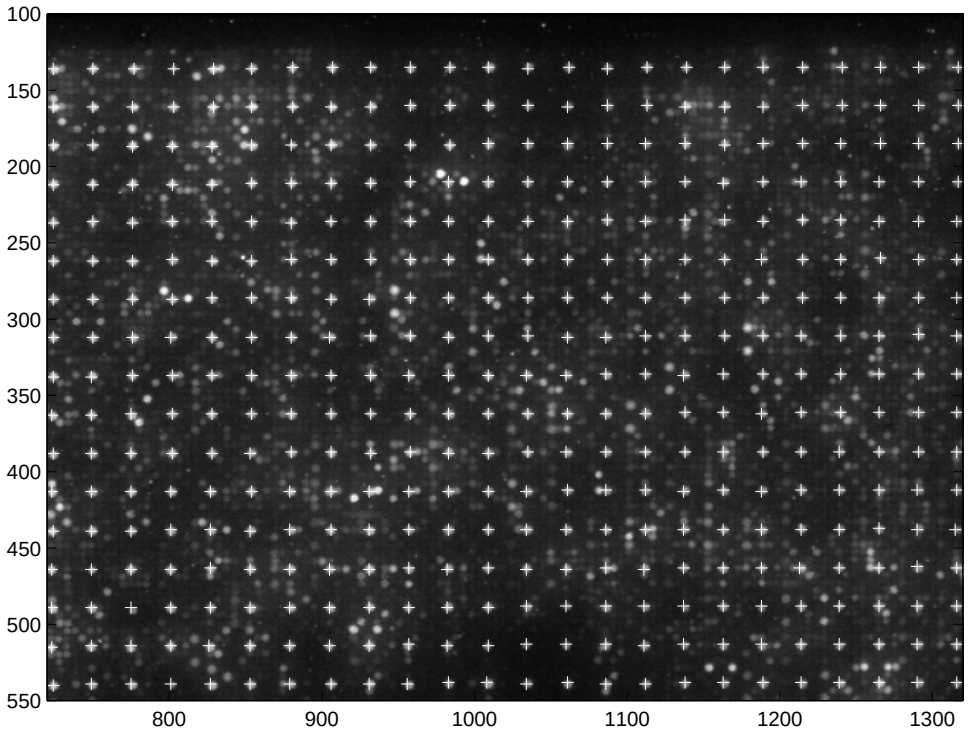
Figure 4.9a shows a part of a 1300×1586 image originating from hybridizations of complex cDNA samples. It has been scanned at a resolution of $200\mu m$ at the NFI Vienna. The grid is nearly full, but the guide spots at the center of the 5×5 blocks are brighter than the majority of the regular spots. Note the visible vertical field shift in pixel row 800 of the image. The correct grid fitting output can be seen in Fig. 4.9b.

Figure 4.10a shows a part of a 1300×1486 image originating from colony filter hybridizations. The image has been scanned at a resolution of $200\mu m$ at the NFI Vienna. The upper left field is clearly noticeable. The intensities of the of the guide spots at the center of the 5×5 blocks differ significantly: They are very high at the first row and first column of the guide spot grid and are partly not distinguishable from the regular spots in regions within the field. The dark rectangular region around pixel $(375, 250)$ in the image indicates that a needle was lacking (broken) on the needle matrix (Fig. 3.5a). Due to the parameterization (2.43) of the fields, such lacking guide spot grid information can be easily restored, as is demonstrated in Fig. 4.10b.

Fig. 4.11a shows a part of a 2400×3544 image originating from hybridizations of complex cDNA samples. If was scanned at a resolution of $100\mu m$ at the NFI Vienna. It is an example for a low signal-to-noise ratio hybridization image. Due to the high resolution of the image, Fig. 4.11b also shows the computed locations of the regular spots superimposed as dots.
Table 4.2.5 summarizes the image information and indicates the names of the image files.

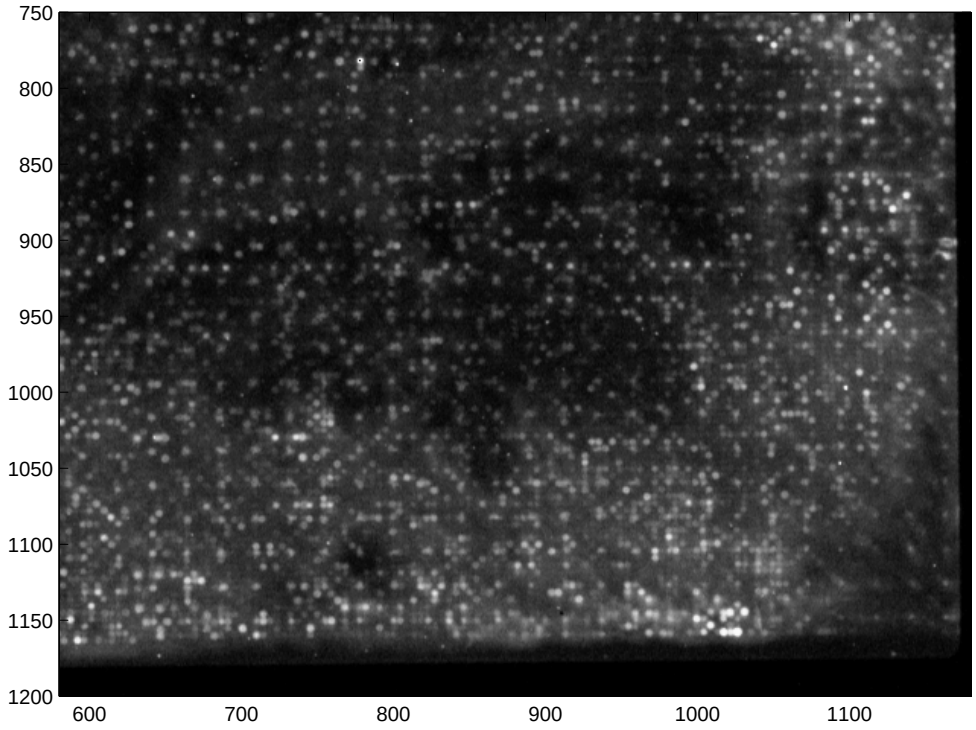


(a) Intensity Data

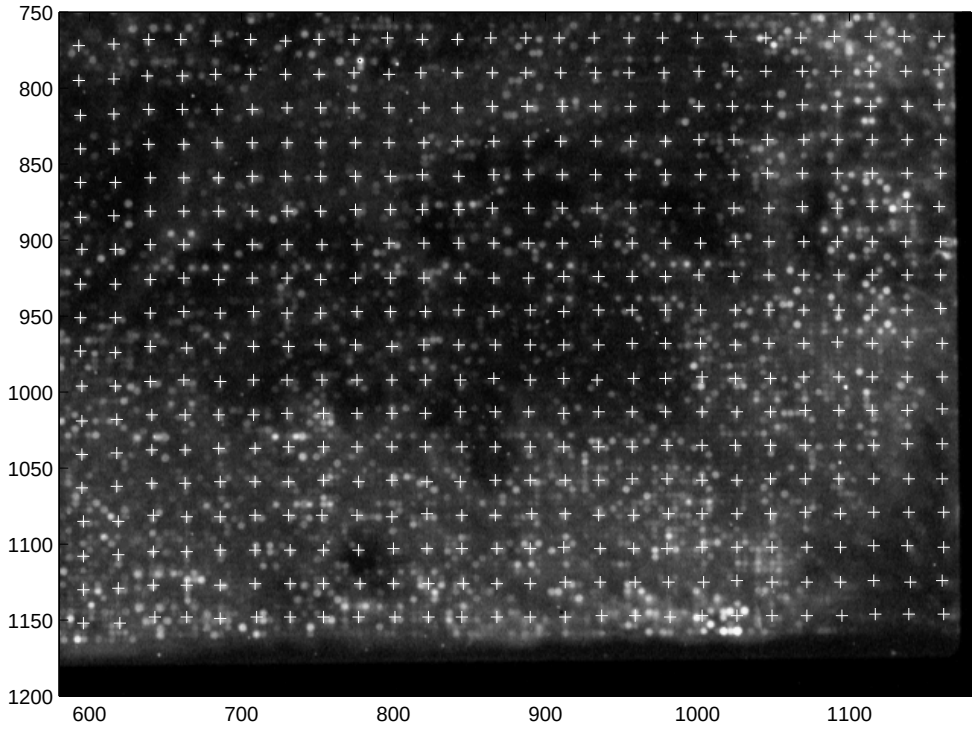


(b) Intensity Data + Guide Spot Locations

Figure 4.7: Part of a 1596×1482 ONF Image with $175\mu m$ resolution scanned at MPIMG Berlin.

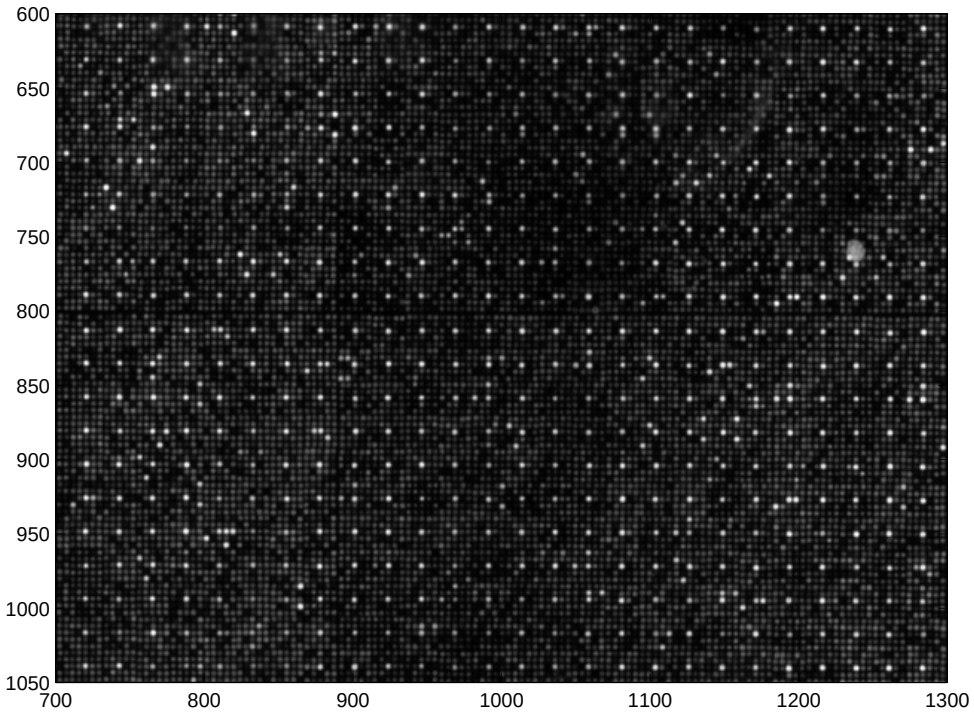


(a) Intensity Data

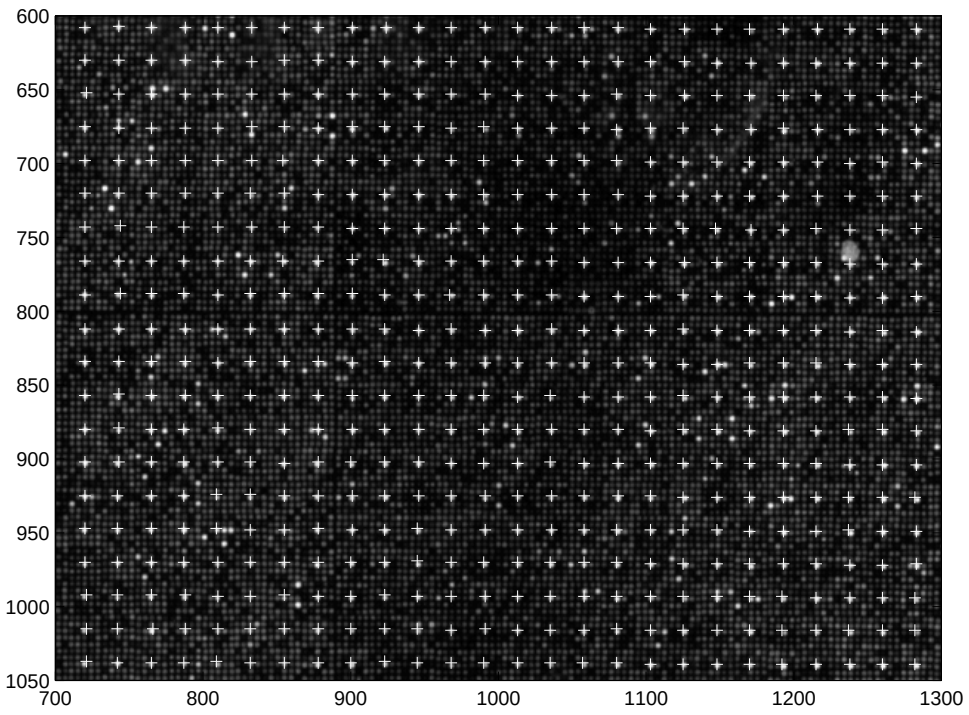


(b) Intensity Data + Guide Spot Locations

Figure 4.8: Part of a 1300×1286 ONF Image with $200\mu m$ resolution scanned at NFI Vienna.

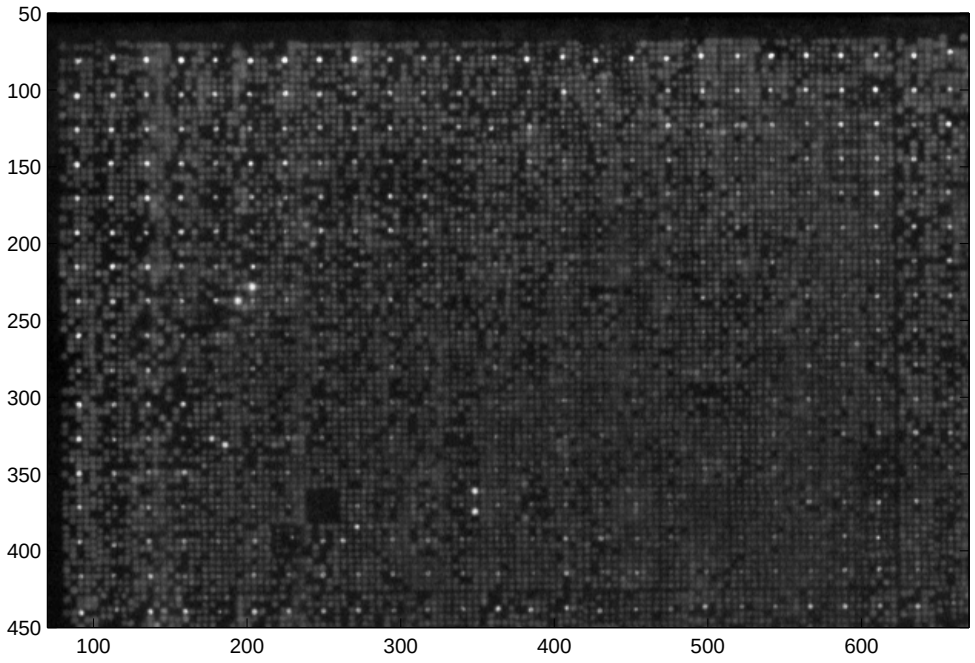


(a) Intensity Data

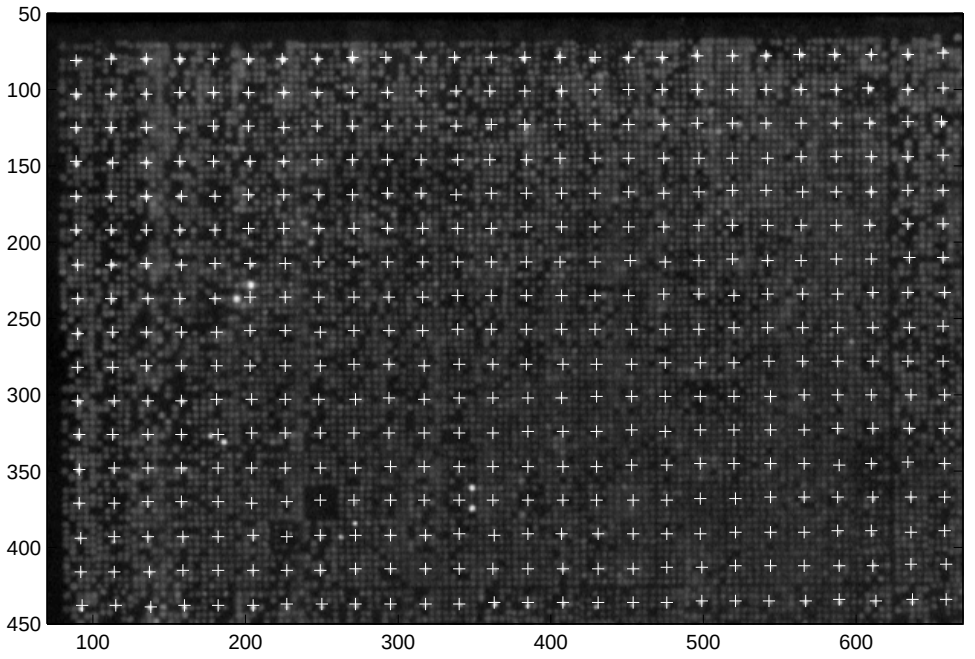


(b) Intensity Data + Guide Spot Locations

Figure 4.9. Part of a 1300×1586 ComplexHyb Image with $200\mu m$ resolution scanned at NFI Vienna.

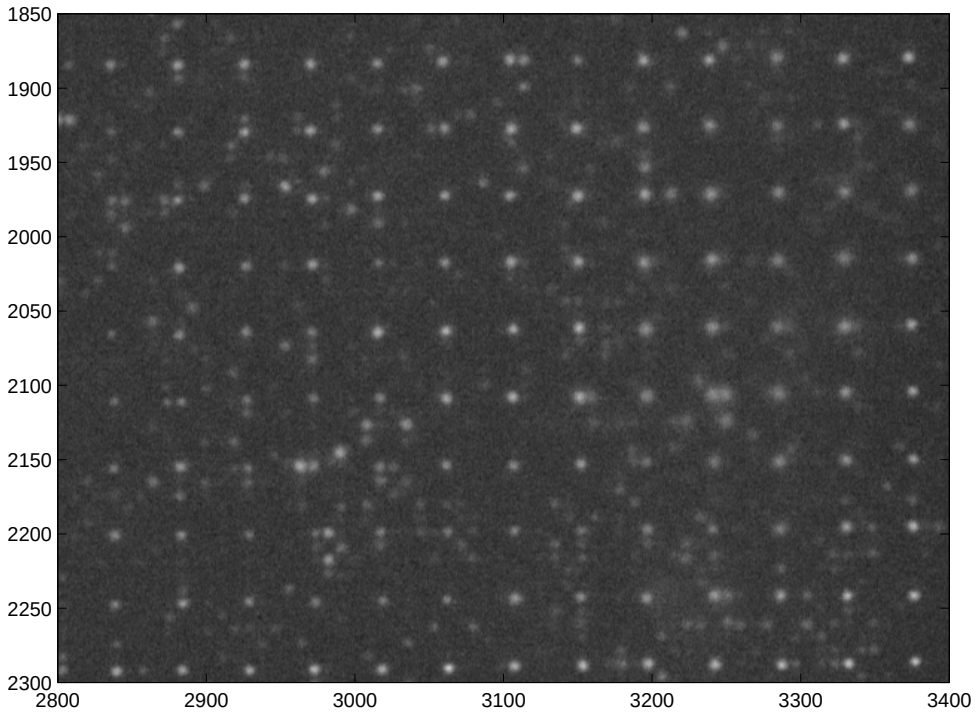


(a) Intensity Data

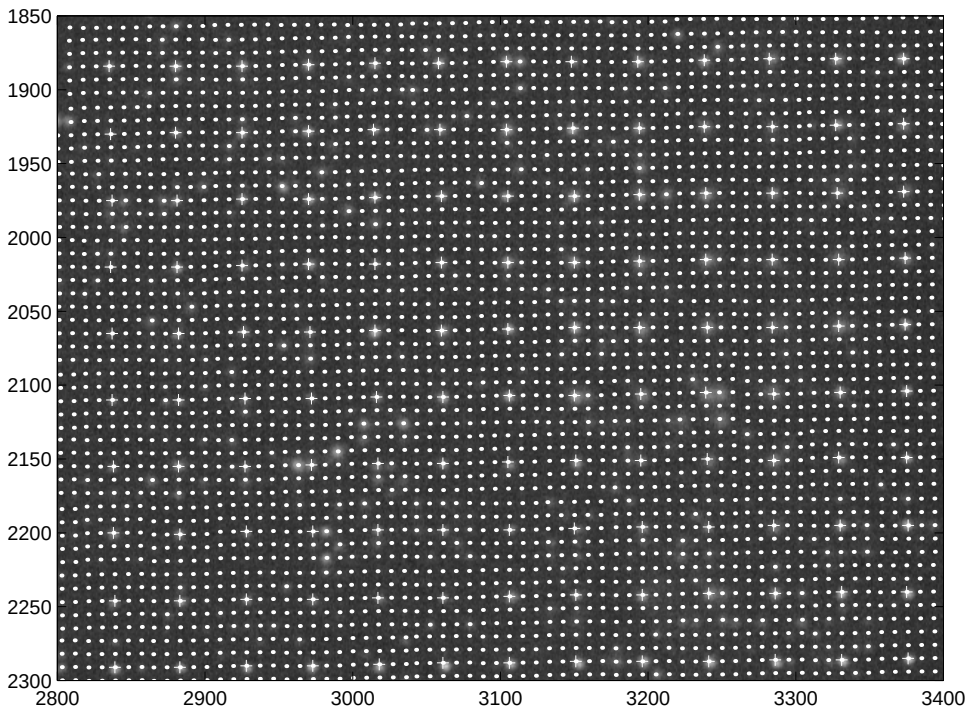


(b) Intensity Data + Guide Spot Locations

Figure 4.10. Part of a 1300×1486 ColonyHyb Image with $200\mu m$ resolution scanned at NFI Vienna.



(a) Intensity Data



(b) Intensity Data + Guide Spot Locations

Figure 4.11. Part of a 2400×3544 ComplexHyb Image with $100\mu m$ resolution scanned at NFI Vienna.

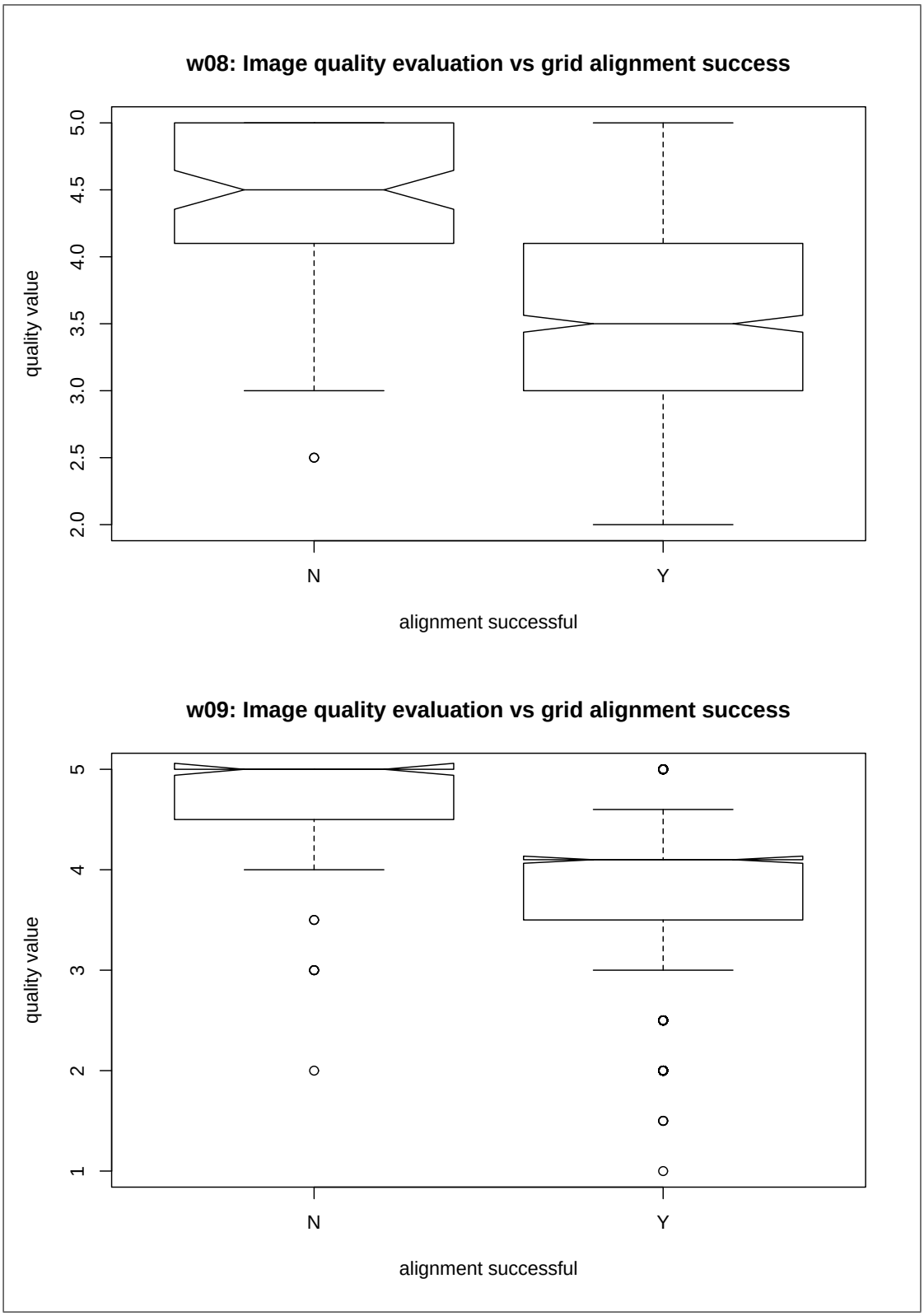


Figure 4.12. Box plots of image qualities versus grid fitting success. The grid fitting is successful for images with good quality. The algorithm can also cope with some images the quality of which was rated as very bad. Most of the images for which the grid fitting fails are rated as bad. There are, however, some outliers of images with good quality which do not meet the expectations. Applying Algorithm II of Sect. 2.4.2 completeley solved this problem.

Figure	Image Name	Experiment	Resolution [μm]	Size
4.7	o163_04260_A1	ONF	175	1596×1482
4.8	o100_295107_y1	ONF	200	1300×1286
4.9	coctail2_c64102_y1	ComplexHyb	200	1300×1586
4.10	990401ptaa_b111dk_y1	ColonyHyb	200	1300×1486
4.11	u266-1-99031_c64120_x1	ComplexHyb	100	2400×3544

Table 4.2: Overview of the image examples demonstrated in Fig. 4.7–4.11.

4.2.6 Evaluation of Image Quality versus Grid Fitting Success

The largest test set consisted of ONF images from two cDNA-libraries named w08 (855 images) and w09 (885 images). These ONF images had been scanned at the Novartis Research Institute Vienna at a resolution of $\Delta x = \Delta y = 200\mu m$. The image quality of every image in w08 and w09 had been rated by a human with numbers between 1 (very good) and 5 (very bad). Hence it is at least possible to investigate the correlation between the image quality and the success of the grid fitting (abortion criterion (2.46)). Figure 4.12 shows both for w08 and w09 images the Box plot of image qualities for which the grid fitting fails (left hand side) and the Box plot of image qualities for which the grid fitting was successful (right hand side) - the initial grid was defined after maximum search (Algorithm I in Sect. 2.4.1. Figure 4.12 can be interpreted as follows:

- As expected, the grid fitting is successful for images with good quality. However, the algorithm can also cope with some images the quality of which was rated as very bad.
- Most of the images for which the grid fitting fails are rated as bad. There are, however, some outliers of images with good quality which do not meet the expectations.

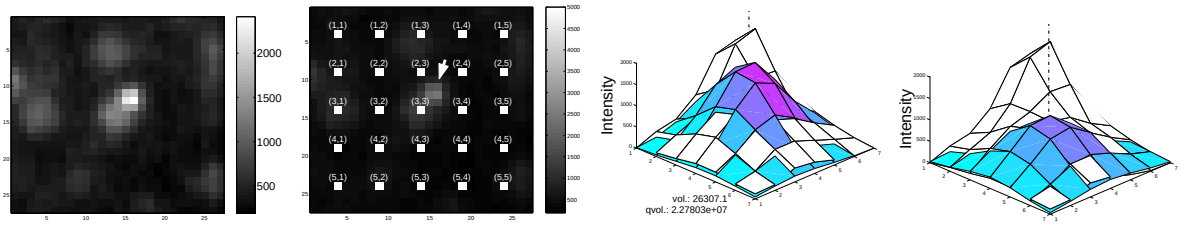
The quality of images was rated by humans only with respect to the level of noise and the shape of the spots. Some apparently good quality images failed in the grid fitting process, because the spot distances have been too irregular. This was due to non-linearities in the step-motor of the scanner and therefore led to problems with the prior guide spot grid covering the whole spot area. Applying the second grid spanning spanning algorithm (Sect. 2.4.2)led to satifactory results.

4.3 Experimental Results of Robust Spot Fitting

This section shows the results of applying the robust parametric spot fitting developed in Sect. 2.6 to DNA array images. Statistical models have been chosen in order to cope with the heavy spot overlap. The spot intensities are quantified as the volume of the Gaussian spot shape according to (2.91).

4.3.1 Artifacts

Consider the image patch of a 5×5 block $\mathcal{B}$ in Fig. 4.13a. The initial spot locations after grid fitting are shown in Fig. 4.13b. The spot $(3, 3)$ in the block center is distorted by an artifact.



(a) Image Data (b) Initial Spot Locations (c) Non-robust Gaussian Fit (d) Robust Gaussian fit

Figure 4.13: Spot Fitting for a Spot with an Artifact.

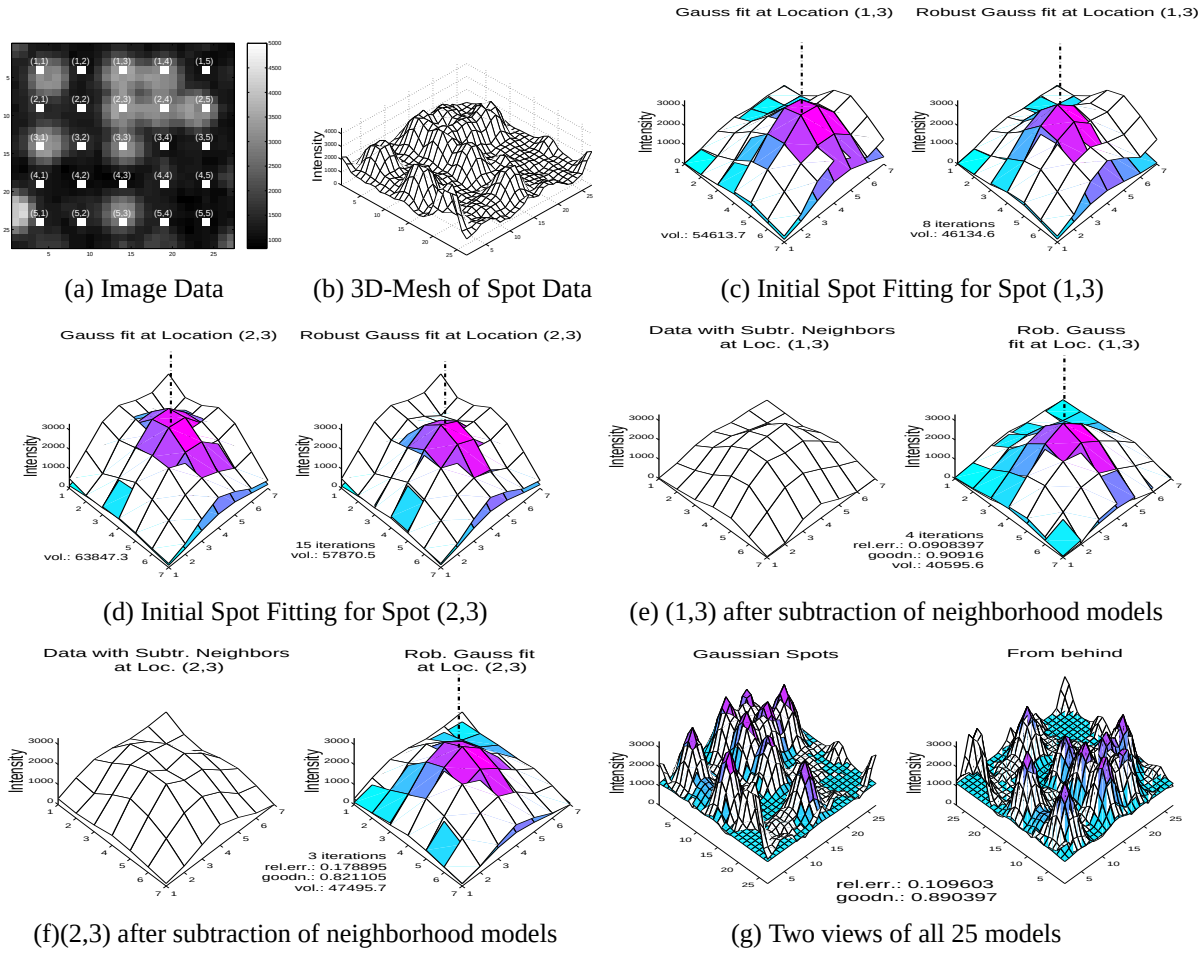
As can be seen in Fig. 4.13c, a simple non-robust Gaussian fit will fail, because the location is biased towards the location of the artifact. The robust Gaussian fit can overcome the outlier (Fig. 4.13d).

4.3.2 Spot Overlap

Figure 4.14 demonstrates how the robust Gaussian fit works on image data with overlapping spots. Figure 4.14a shows a 5×5 block originating from an ONF image with low resolution, together with the initial spot locations after grid fitting. Figure 4.14b shows the image data as a 3D-meshgrid. Spots $(1,3)$, $(2,5)$ and $(3,3)$ have up to three overlapping neighbors, here the robust estimator can recover the original spot location quite well, especially for $(1,3)$ and $(3,3)$. Spot $(1,3)$ is plotted in Fig. 4.14c. The non-robust Gaussian fit is biased towards the neighboring spots, whereas the location of robustly fitted Gaussian spot is more plausible. Spots $(1,4)$, $(2,3)$ and $(2,4)$ have over four overlapping neighbors and are therefore difficult cases, but still some improvements can be achieved by robust fitting. The non-robust and robust Gaussian fit of spot (2,3) are plotted in Fig. 4.14d. After the first robust Gauss fit we refit on every location with subtracted neighborhood models. The centers computed during the first fit are taken as the initial centers for the second fit. When taking a look at the new patches with subtracted neighbors (see Fig. 4.14e) one will notice that the patches are now less distorted than the previous patch and are more "spot like" – an indication that the situation has improved.

When investigating the goodness of fit and the patch shapes, the first robust fitting resolved the overlaps at spots $(1,3)$ (see Fig. 4.14e) and $(3,3)$ very well. The results for the spots (1,4) and (2,5) are good, the results for (2,3) (see Fig. 4.14f) are acceptable, and the results for (2,4) are not good enough. Generally, on can say that the robust estimation will perform well up to four overlapping neighbors while more than four will make problems. This is can be explained by the fact that highest possible breakdown point of a robust estimator is $\epsilon^* = 0.5$. If more than 50% of the input data are false the situation cannot be recovered directly by a robust estimator. An overview of the fitted models can be seen in Fig. 4.14g.

Figures 4.15a and b show a volcano spot with an overlap from the right hand side. An ordinary Gaussian fit would be biased to the right neighbor, but a robust estimator recovers the location easily (Fig. 4.15c). After performing a robust Gaussian fit on both sides, we subtract the neighborhood spot model from the patch receiving the corrected data (see Fig. 4.15d). The initial volume estimation after a Gaussian refit can be observed in Fig. 4.15e, but the estimated volume is not very reliable due to the high relative error rate. Using the center and dispersion we



(a) Image Data (b) 3D-Mesh of Spot Data (c) Initial Spot Fitting for Spot (1,3)

(d) Initial Spot Fitting for Spot (2,3) (e) (1,3) after subtraction of neighborhood models

(f)(2,3) after subtraction of neighborhood models (g) Two views of all 25 models

Figure 4.14: Spot Fitting of Overlapping Spots.

performed a semi parametric fit (see Fig. 4.15f). We smoothed the profile points by replacing each point (except at the border) with the weighted sum over the left, the point itself and right neighbor with the weights 3,6, and 2. The left neighbor received higher weights, because the points on the left hand side are more reliable since they are closer to the center. The goodness of fit improved and a more reliable quantification is done.

Figure 4.16a shows of a part of an ONF image as a 3D-surface. The fitted Gaussian models after three neighborhood subtraction iterations are shown in Fig. 4.16b. Overlapping spots are well-separated (the individual contributions of the neighborhood models to an image location are not added in Fig. 4.16b). Some spots do not have a Gaussian shape, so a semi-parametric fit would eventually be more appropriate.

4.3.3 Complexity

Table 4.3.3 shows the CPU-time costs for each method per fit in flops (the methods have been implemented in Matlab™). The values should be interpreted as follows:

1. A (non robust) Gaussian fit in low resolution requires approximately 10.000 flops.

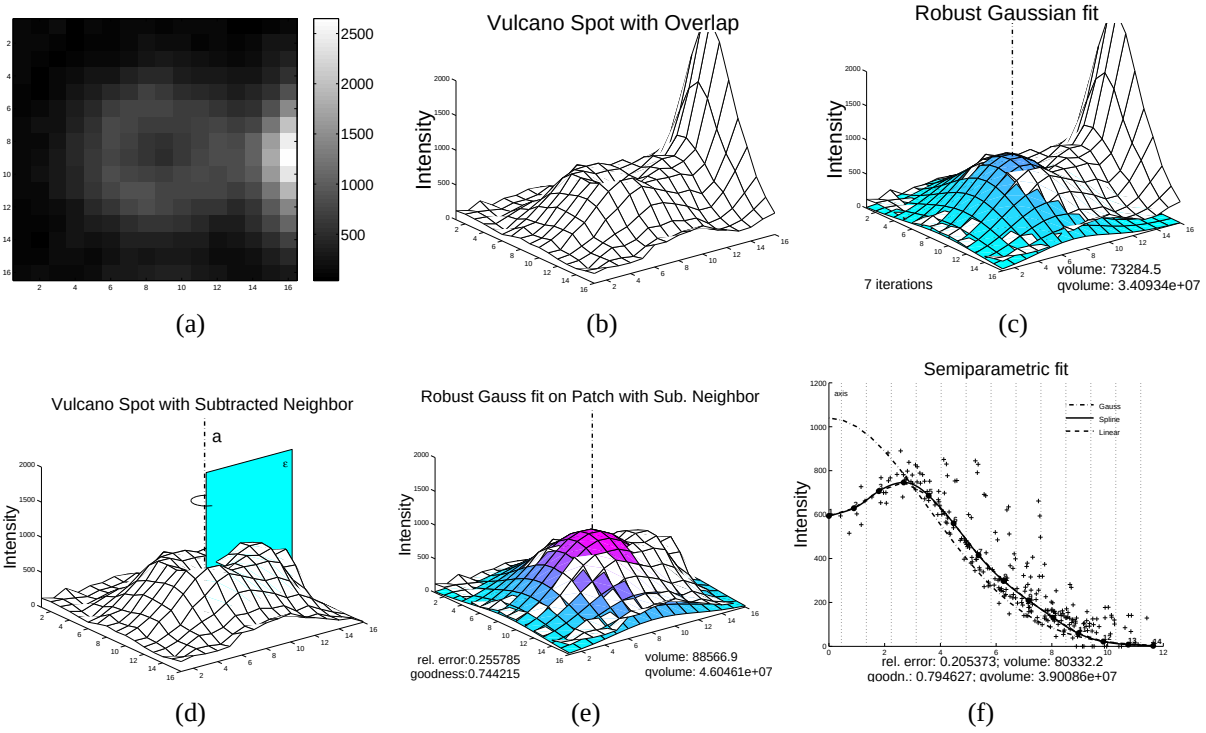


Figure 4.15: Volcano Spot with Overlapping Neighbor.

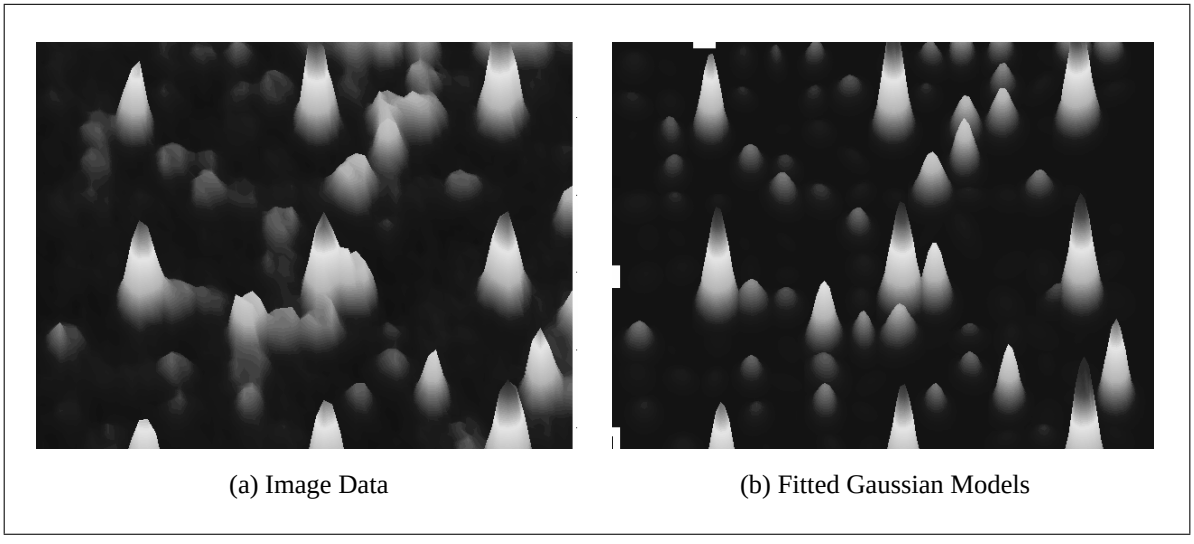


Figure 4.16: Three-dimensional Illustration of Parametric Spot Fitting.

Resolution → Method ↓	Low Res. 7x7 flops/per fit	High Res. 16x16 flops/per fit
Gaussian Fit	10.000	47.000
Semi-param. Fit	2.000	15.000

Table 4.3: CPU-time in Flops

2. A robust Gaussian fit with k iterations requires approximately $(k+1) \times 10.000$ flops (1 fit for the initial guess and k remaining fits for each iteration).

3. A semi-parametric fit with 5 "profile points" costs 2.000 flops in low resolution, while in high resolution 14 "profile points" are computed requiring 15.000 flops.

4. A single semi-parametric fit is approximately four times faster than a Gaussian fit in low and high resolution. However, one should keep in mind that a semi-parametric fit in general can not be performed directly without any preceding center search by a M-estimator of location.

5. Let $n \times n$ be the dimension of the input patch, i.e. $n = 7/n = 16$ for low/high resolution. While the computing time for the Gaussian fit will increase with $O(n^2)$, the computing time for a semi-parametric fit will increase with $O(n^2 \cdot \log(n))$. The reason is that a Gaussian fit basically sums over all data points while sorting algorithms are needed for a semi-parametric fit.

6. An implemented C-version (KHOROS toolbox) of the non-robust spot fitting with subtracting neighbors for all spots needs about four minutes on the same machine (including the grid fitting).

4.3.4 Comparison to Existing Approach

We have compared the performance of the parametric spot fitting with Gaussian models (in the following denoted as SPOTFIT) to an approach which simply averages a rectangular pixel area around the spot center. Such a simple quantification is used by Hartelius [Hartelius, 1996], where the whole image analysis including grid fitting is denoted as hybridization fingerprint analysis (HFA). Our results stem from a test image with 200 μm resolution for which the HFA quantification is the mean pixel intensity of a 3×3 rectangular window around the spot center. Fig. 4.17a and b show the histograms of the log-intensities of the quantified spots with HFA and SPOTFIT, respectively. The HFA intensities are lower since the spots for the test image have a theoretical (mean) expansion of $S_y \times S_x = 4.5 \times 4.5$ pixels according to (2.4). This expansion is not covered by the 3×3 pixel HFA approach and therefore leads to lower quantifications. Furthermore, as opposed to the Gaussian-shaped intensity distribution of Fig. 4.17b, the distribution in Fig. 4.17a shows a tail in the direction of higher intensities. This imbalance is also illustrated in Fig. 4.18a, where the spot intensities of both methods are plotted: The scatter shows that there is a tendency of HFA spot intensities – especially darker ones –to have lower quantification values in SPOTFIT.

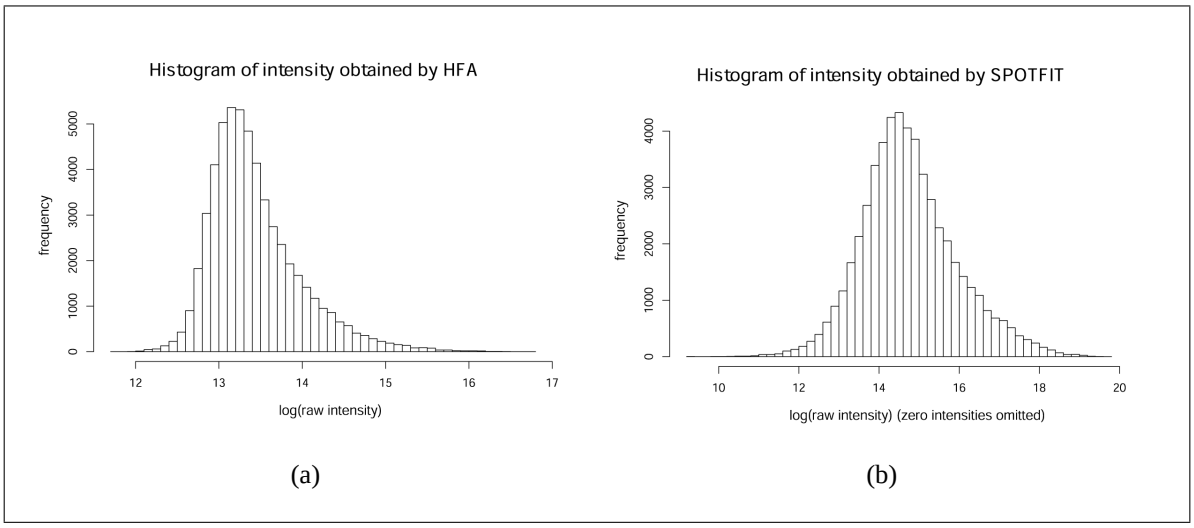


Figure 4.17. Intensity comparison of parametric spot fitting (SPOTFIT) with simple quantification by averaging (HFA).

This phenomenon can be explained by looking at the the spotting pattern of a block $\mathcal{B}$ of the investigated microarray image: In order to increase the reliability of the hybridization signals, every probe was spotted twice within a 5×5 block, except of the guide spot at the block center. The spotting pattern matrix is

$$\begin{pmatrix} 1^{(a)} & 2^{(g)} & 3^{(i)} & 4^{(a)} & 5^{(b)} \\ 6^{(e)} & 7^{(h)} & 8^{(f)} & 9^{(j)} & 10^{(f)} \\ 11^{(i)} & 12^{(k)} & 13^{(-)} & 14^{(d)} & 15^{(l)} \\ 16^{(d)} & 17^{(c)} & 18^{(g)} & 19^{(k)} & 20^{(b)} \\ 21^{(h)} & 22^{(e)} & 23^{(j)} & 24^{(i)} & 25^{(c)} \end{pmatrix},$$

where the superscripts of the probe numbers denote the duplicate to which to probe belongs. For example, the probe pairs $(1, 4)$, $(2, 18)$ and $(8, 10)$ form duplicates (Fig. 4.18b). Fig. 4.18c shows the box plots of the quantified HFA intensities, where the horizontal numbers 1–25 indicate the pattern position in the block according to Fig. 4.18a. The box plots show an intensity imbalance of those duplicates which include a neighbor of the guide spot (which has always a high hybridization signal). For example, the spot intensity of block position 18 is significantly higher than the spot intensity of position 2, because the pixel intensity data of position 18 also contain parts of the bright upper guide spot neighbor 13 (Fig. 4.18b). In contrast, the HFA intensities of duplicate $(1, 4)$ are balanced, since neither position is a neighbor of the guide spot. The box plots of the SPOTFIT intesnties in Fig. 4.18d show that this imbalance problem is relaxed. This is mainly due to the robust estimation and the subtraction of neighborhood models.

4.4 Chapter Summary

DNA array image analysis serves as a case study for applying the general framework developed in Chapter 2. The grid fitting of existing published and commercial methods can be divided

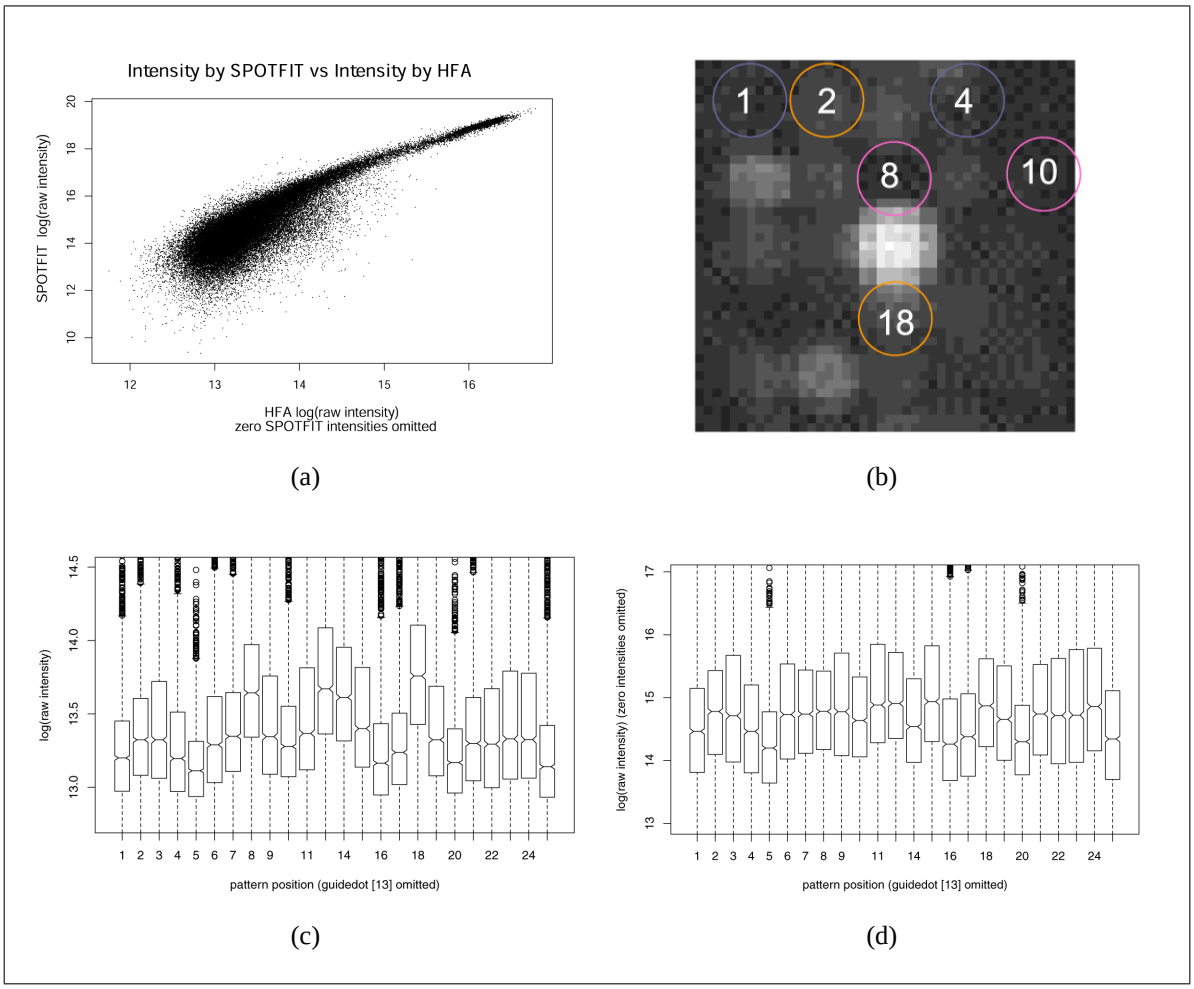


Figure 4.18. Positional intensity bias. (a) Many high spot intensities in HFA have a lower intensity value in SPOTFIT. This can be explained by regarding the spotting pattern. (b) The spotting pattern for a 5×5 block consists of 12 spot duplicates and a guide spot in the center. The guide spots have always a high intensity and therefore have an influence to the adjacent neigbhors. (c) When quantifying with a simple mean value like in HFA, the spot intensities for duplicates with an adjacent guide spot neighbor are imbalanced. (d) When using robust statistical models, the situation is relaxed.

into semi-automatic and automatic grid fitting. The semi-automatic grid fitting methods require some level of user interaction. The user needs to tell the program where the outline of the grid is in the image. Automatic grid fitting algorithms do not need user interaction and can be used for batch processing. If an existing automatic grid fitting method takes into account grid rotation, the whole spot image is rotated back in order to facilitate grid spanning. This generally means a loss of information and impacts the quantification of the spot intensities.

Current quantification methods for DNA arrays are based on trying to segment the signal area from the background area. The simplest approaches fit a circle with a certain diameter to all the spots in the image in order to refine the regular grid positions of the grid provided by the user. More advanced methods base the segmentation on the histogram of the intensities belonging to

a spot and may take into account spatial information. Adaptive shape segmentation includes region-based segmentation methods like watersheds and Seeded Region Growing. Based on the segmentation, the intensity of the spots representing the amount of bound DNA or mRNA can be estimated with the following methods: Total intensity, mean intensity, median intensity, the mode of the intensity histogram and the volume. If the hybridization is two-color and the scanning measurements are taken in two channels, then the intensity ratio between the channels is a possible alternative.

We applied the framework described in Chapter 2 to the analysis of DNA array images. Some minor adaptions for the grid fitting have to be made in order to cope with a possible spot field shift and the subgrid called guide spot grid. Due to the heavy spot overlap we used a Gaussian parametric spot model as described in Sect. 2.6. The quantified intensities are defined as the volume under the Gaussian model. The approach proved to be successful on thousands of images. The grid fitting success correlates with the quality of the images (which was assessed by a human). The robust quantification was demonstrated to be superior to a simple quantification by the means of spot duplicates. This led to more plausible and reliable data for subsequent array analysis steps like clustering and expression profiling.

Chapter 5

Conclusion and Outlook

This work described a novel general framework for the analysis of spot array images. DNA array images – an increasingly important tool in biotechnology – served as an example to demonstrate that the approach can deal with high spot density and possibly multiple overlapping spots. Furthermore, the approach is able to cope with lacking spots in the array and contaminations in the spot array image.

The proposed general framework for spot array image analysis is composed of a set of robust tools. The analysis starts with an amplification of the spot locations with the help of matched filters built by averaging a number of representative training spots. The matched filter response is expected to have maximum values at the spot locations. The grid rotation is estimated with the help of projections of matched filter responses along different directions. The projection at the correct rotation angle is expected to have maximum projection values. Two alternative grid spanning modules enable to span a consistent grid with the help of the matched filter response and the estimated rotation angle. The first grid spanning method transforms a prior spot grid according to the grid corner locations. The grid corner locations are estimated by robustly fitting straight lines to the first and last row and column of the grid. If the spot distances are too irregular, an alternative grid spanning projects intersects back-projected straight lines based on the inverse Radon transform. The spot characterization step consists of the interdependent background estimation and spot fitting. The background estimation uses a hierarchical pyramidal approach in order to yield smooth backgrounds. At a high pyramid level with low resolution, a synthetic image based on spot fitting is subtracted and the result is interpolated to the original spot image resolution.

The core spot fitting approach in the framework is based on statistical spot models. The parameters of the statistical model describe the observed distribution of pixel intensities and must be fitted to the observed data. Maximum Likelihood (ML) estimators find the parameters that best describe the observed data based on the squared error between the data and the model. The parameters of a Gaussian model consist of an overall height (amplitude), two parameters for the spot center (mean) and three parameters for the dispersion of the spot (covariance matrix). The squared error of the ML estimator is non-robust against outliers. Robust M-estimators weight the contribution of a data point according to the residual error between the data and the model: A large residual will be scaled down or even be truncated in order not to bias the estimator. The semi-parametric approach is based on the robust parametric fit and then performs a dimension reduction by rotating a plane around the spot center. A robustly fitted curve in the

plane then allows to model deviation the Gaussian form, e.g. a volcano spot. Non-parametric approaches include the classical segmentation methods to segregate signal and background.

The general framework has two main characteristics: Firstly, the tools do not need critical thresholds provided by human users. The only user-information is the array configuration. Secondly, for the quantification step, the approach keeps as much information as possible. We avoid, for example, rotating the spot array image since this would mean an information loss caused by gray level interpolation.

The approach was demonstrated to work on challenging DNA array images containing up to 57.600 spots. The Gaussian spot model proofed to be useful for the images in the test set. However, evolving technology will yield spot array images with increasing resolution. Increasing resolution images contain more and more volcano spot shapes: Owing to the impact of the robot finger on the solid support, a spot has a small inner core that is not illuminated. One the other hand, the increased input data will allow for statistical models with more parameters than the Gaussian. Note, however, that for two-color hybridization images the key information is the ratio between the two color channels. One can observe that when immediately dividing the green channel image by the red channel image, the volcano shape vanishes. Nevertheless, more sophisticated models might be necessary in order to characterize the underlying data. Bettens et al. [Bettens et al., 1996] provided a spot model for electrophoresis gels based on diffusion principles. They assume two main direction of diffusion, with different diffusion properties for each direction. Furthermore, the initial distribution is not concentrated in one point but occupies a finite region. Balagurunathan et al. [Balagurunathan et al., 2001] provide a random model for the generation of fluorescent cDNA array images. They model the inner core of a spot by an ellipse with random horizontal and vertical axes. Since on glass slides the liquid placed on the spot tends to accumulate towards the outer edge, the edge is randomly enhanced. The contaminations occurring due to various physical effects at the hybridization step, chords are randomly cut from the circular spots. The number of chords follows a Poisson distribution. Their overall spot model contains over twenty parameters.

Another extension to the current approach would be to employ machine learning techniques [Cherkassky and Mulier, 1998]. When analyzing sets of spot array images, the computer performs millions of fits. It would be possible to learn to assess the goodness-of-fit, to discriminate between spots and non-spots. If a spot model does not reflect the underlying distribution of the spot intensities, the machine could develop heuristics that quantify systematic deviations from the true spot intensity

Appendix

Hierarchical Radon transform It is useful to compute $\mathbf{R}^{\mathrm{T}}$ in a *hierarchical* manner with increasing angle resolutions $\Delta\theta$:

1. Start at the initial resolution, for example $\Delta\theta_0 = 1.0°$, and compute a Radon transform for the maximum rotation angle between $-\theta_{\mathrm{M}}$ and $+\theta_{\mathrm{M}}$ in order go get a coarse rotation estimate $\theta_{\mathcal{G}_0}$.

2. Double the angle resolution to $\Delta\theta_1 = \Delta\theta_0/2 = 0.5°$. As the accuracy of the initial angle estimate $\theta_{\mathcal{G}_0}$ is $\pm\Delta\theta_0$, or $\pm 1°$, it is sufficient to compute 5 projections for the angle set $\{\theta_{\mathcal{G}_0} - 2\Delta\theta_1,\ \theta_{\mathcal{G}_0} - \Delta\theta_1,\ \theta_{\mathcal{G}_0},\ \theta_{\mathcal{G}_0} + \Delta\theta_1,\ \theta_{\mathcal{G}_0} + 2\Delta\theta_1\}$.

3. Repeat step 2 until the desired angle resolution $\Delta\theta$ is reached. Every step except of the first one only requires the computation of 5 projections.

Table 5 illustrates the speedup that can be gained using the hierarchical approach.

Maximum Rotation Angle. For some spot array images it is feasible to compute a a maximum possible rotation which depends on the size of the physical filter and the size of the digital spot array image. However, the pixel dimensions of some spot array images are much larger than the pixel dimensions of the physical filter they comprise. As a consequence, a filter could

Non-Hierarchical Radon transform			
Iteration	$\Delta\theta$	$\theta_{\mathcal{G}}$	Projections
0	$0.125°$	$1.625°$	54
Total number of projections:			54

Hierarchical Radon transform			
Iteration	$\Delta\theta$	$\theta_{\mathcal{G}}$	Projections
0	$1.0°$	$2°$	9
1	$0.5°$	$1.5°$	5
2	$0.25°$	$1.75°$	5
3	$0.125°$	$1.625°$	5
Total number of projections:			24

Table 5.1. Speedup for the hierarchical Radon transform with am maximum rotation angle of $\theta_{\mathrm{M}} = 4°$. For an angle resolution of $\Delta\theta = 0.125°$, the non-hierarchical Radon transform needs 54 projections according to (2.16). the hierarchical approach gradually refines an initially coarse angle resolution ($\theta_{\mathcal{G}}$ shows the intermediate results for the grid rotation angle). Every refinement of the initial results only needs 5 projections, the total number of projections for the desired angle resolution therefore sums up to only 24 projections.

have theoretically any rotation $\theta_{\mathcal{G}}$ in the digital image. We found empirically that no filter was rotated more than $3°$. In order to have an additional tolerance we set $\theta_{\text{M}} = 4°$.

Angle Resolution. The necessary angle resolution $\Delta\theta$ depends on the size of the $M \times N$ spot array image $\mathbf{S}$. The bigger the image, the more orientations a straight line can have in the digital image. In the central coordinate system with the origin at the image center, the straight line with the minimal possible rotation has a Δx of $N/2$ pixels and a Δy of 1 pixel. Since we are dealing with digital images with $M > 1000$ and $N > 1000$, there is barely a difference between a straight lines with $\Delta y = 1$ and a straight line with $\Delta y = 2$. We therefore fix the minimal Δy to 2 pixels and have

$$\Delta\theta = \frac{180}{\pi} \arctan\left(\frac{2}{N/2}\right) = \frac{180}{\pi} \arctan\left(\frac{4}{N}\right). \tag{5.1}$$

Applying the estimated rotation $\theta_{\mathcal{G}}$ to the set of prior guide spot locations $\text{L}_{\text{P}}(\mathcal{G}^{\star})$ yields a set $\text{L}_{\theta}(\mathcal{G}^{\star})$ of rotated prior guide spot locations.

Bibliography

[Abdel-Aziz and Karara, 1971] Abdel-Aziz, Y. L. and Karara, H. M. (1971). Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. In American Society of Photogammetry, Falls Church, V., editor, *Proc. Symposium on Close-Range Photogrammetry*, pages 1–18. 10

[Adams and Bischof, 1994] Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(6):641–647. 9

[Antoine et al., 1993] Antoine, J. P., Carette, P., Murenzi, R., and Piette, B. (1993). Image analysis with two-dimensional continuous wavelet transform. *Signal Processing*, 31:241–272. 8

[Appel et al., 1997] Appel, R. D., Vargas, J. R., Palagi, P. M., Walther, D., and Hochstrasser, D. F. (1997). Melanie II - a third-generation software package for analysis of two-dimensional electrophoresis images. II: algorithms. *Electrophoresis*, 18(15):2735–2748. 65, 69

[Axon Instruments, 1999] Axon Instruments (1999). *GenePix 400A User's Guide*. Foster City, CA. 68

[Balagurunathan et al., 2001] Balagurunathan, Y., Dougherty, E. R., Chen, Y., Bittner, M. L., and Trent, J. M. (2001). A Random Signal Model for cDNA Microarrays. In *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE - Progress in Biomedial Optics and Imagin*, pages 163–170. SPIE, International Society for Optical Engineering. 95

[Ballard, 1981] Ballard, D. H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13:111–122. 69

[Batista et al., 1999] Batista, J., Araùjo, H., and de Almeida, A. T. (1999). Iterative Multistep Explicit Camera Calibration. *IEEE Trans. on Robotics and Automation*, 15(5). 10

[Bayer and Mogg-Schneider, 1997] Bayer, T. A. and Mogg-Schneider, H. U. (1997). A Generic System for Processing Invoices. In IEEE, editor, *Proc. Fourth International Conference on Document Analysis and Recognition*, volume 2. 1

[Bergemann et al., 2001] Bergemann, T., Quiaoit, F., Delrow, J., and Zhao, L. P. (2001). Statistical Issues in Signal Extraction from Microarrays. In *Microarrays: Optical Technologies*

and Informatics, volume 4266 of *Proceedings of SPIE - Progress in Biomedial Optics and Imaging*, pages 24–34. SPIE, International Society for Optical Engineering. 11, 66, 68

[Bertsekas, 1999] Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific. 27

[Bettens et al., 1996] Bettens, E., Scheunders, P., Sijbers, J., Van Dyck, D., and Moens, L. (1996). Automatic Segmentation and Modelling of Two-Dimensional Electrophoresis Gels. In *Proceedings IEEE International Conference on Image Processing, September 1996, Lausanne Switzerland, Vol. II*, pages 665–668. 95

[Beucher and Meyer, 1993] Beucher, S. and Meyer, F. (1993). The morphological approach to segmentation: the watershed transformation. In *Mathematical morphology in image processing*, volume 34 of *Optical Engineering*, chapter 12, pages 433–481. Marcel Dekker, New York. 10

[Bishop, 1995] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press. 35, 44, 45

[Blanford and Tanimoto, 1988] Blanford, R. and Tanimoto, S. (1988). Bright-Spot Detection in Pyramids. *Computer Vision, Graphics and Image Processing*, 43(2):133–149. 2, 8

[Boccignone et al., 2000] Boccignone, G., Chianese, A., and Picariello, A. (2000). Multiresolution spot detection by means of entropy thresholding. *Journal of the Optical Society of America*, 17(7):1160–1171. 3, 8

[Boguski et al., 1993] Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). dbEST-databse for ”expressed sequence tags”. *Nature Genetics*, 4:332–333. 58

[Brändle et al., 2001] Brändle, N., Bischof, H., and Lapp, H. (2001). A Generic and Robust Approach for the Analysis of Spot Array Images. In *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE - Progress in Biomedical Optics and Imaging*, pages 1–12, San Jose, California, USA. SPIE International Society for Optical Engineering.

[Brändle et al., 2000] Brändle, N., Chen, H.-Y., Bischof, H., and Lapp, H. (2000). Robust Parametric and Semi-parametric Spot Fitting for Spot Array Images. In *ISMB-2000 8th Intl. Conference on Intelligent Systems for Molecular Biology, August 20–23*, pages 46–56, La Jolla, California, USA.

[Brändle et al., 1999] Brändle, N., Lapp, H., and Bischof, H. (1999). Automatic Grid Fitting for Genetic Spot Array Images Containing Guide Spots. In Solina, F. and Leonardis, A., editors, *Computer Analysis of Images and Patterns, CAIP99*, volume 1689 of *LNCS*, pages 357–366. Springer.

[Bresenham, 1965] Bresenham, J. E. (1965). Algorithm for Computer Control of a Digital plotter. *IBM Systems Journal*, 4(1):25–30. 23

[Broadhurst and Cipolla, 1999] Broadhurst, A. and Cipolla, R. (1999). Calibration of Image Sequences for Model Visualisation. In *Proc. Conf. Computer Vision and Pattern Recognition*, volume I, pages 100–105. 10

[Bryant et al., 2000] Bryant, M., Wettergreen, D., Abdallah, S., and Zelinsky, S. (2000). Robust Camera Calibration for an Autonomous Underwater Vehicle. In *Australian Conference on Robotics and Automation, Melbourne, Australia*. 11

[Buhler et al., 2000] Buhler, J., Ideker, T., and Haynor, D. (2000). Dapple: Improved Techniques for Finding Spots on DNA Microarrays. Technical Report UWTR 200-08-05, University of Washington. 65

[Canny, 1986] Canny, A. (1986). A computational approch to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6):679–698. 10

[Chee et al., 1996] Chee, M., Yang, R., Hubbell, E., and Berno, A. (1996). Accessing Genetic Information with High-Density DNA Arrays. *Science*, 274:610–614. 4, 57

[Chen et al., 2001] Chen, B., Stoughton, C., and Smith, J. A. (2001). Stellar Population Studies with the SDSS. I. The Vertical Distribution of Stars in the Milky Way. *The Astrophysical Journal*, 553(1):184–197. 4

[Chen et al., 2000] Chen, H.-Y., Brändle, N., Bischof, H., and Lapp, H. (2000). Robust spot fitting for genetic spot array images. In Society, I. S. P., editor, *ICIP-2000 Intl. Conference on Image Processing*, volume 3, pages 412–415, Vancouver, Canada.

[Chen et al., 1997] Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio-Based Decicions and the Quantitative Analysis of cDNA Microarray Images. *Journal of Biomedical Optics*, 2(4):364–374. 69

[Cherkassky and Mulier, 1998] Cherkassky, V. S. and Mulier, F. M. (1998). *Learning from Data : Concepts, Theory, and Methods*. John Wiley and Sons. 24, 95

[Collins and Schneider, 1998] Collins, S. and Schneider, J. (1998). *Braille for the Sighted*. Garlic Press. 4

[Cooper, 1979] Cooper, D. (1979). Maximum Likelihood Estimation of Markov-Process Blob Boundaries in Noisy Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1(4):372–384. 9

[Danker and Rosenfeld, 1981] Danker, A. and Rosenfeld, A. (1981). Blob Detection by Relaxation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 3(1):79–92. 2, 9

[Deloukas et al., 2001] Deloukas, P., Matthews, L., and Ashurst, J. (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature*, 414(20):865–872. 58

[DeRisi et al., 1996] DeRisi, J. L., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer. *Nature Genetics*, 14(4):457–460. 57

[Duda et al., 2000] Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification and Scene Analysis*. John Wiley and Sons. 45, 61

[Duggan et al., 1999] Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics Supplement*, 21:10–14. 57, 60, 61

[Eisen, 1999] Eisen, M. B. (1999). *ScanAnalyze documentation*. http://rana.Standford.edu/software. 68

[Eye Institute, 1996] Eye Institute, N. (1996). Haptic Display of Computer Graphics for the Blind. Health and Human Services grant:SBIR 5-R44-EY0S166-04. 4

[Faugeras, 1993] Faugeras, O. (1993). *Three-Dimensional Computer Vision. A Geometric Viewpoint*. MIT Press. 10

[Fodor et al., 1991] Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-Directed, Spatially Addressable Parallel Chemical Synthesis. *Science*, (251):767–773. 56

[Gennery, 1979] Gennery, D. (1979). Stereo-camera calibration. In *Proc. 10th Image Understanding Workshop*, pages 101–108. 10

[Goffeau et al., 1996] Goffeau, A., Barrell, B. G., and Bussey, H. (1996). Live with 6000 Genes. *Science*, 274:546–567. 58

[Granjeaud et al., 1996] Granjeaud, S., Nguyen, C., Rocha, D., Luton, R., and Jordan, B. R. (1996). From hybridization image to numerical values: a practical, high throughput quantification system for high density filter hybridization. *Genetic Analysis: Biomolecular Engineering*, 12:151–162. 65

[Halousek, 1999] Halousek, J. (1999). Embossed Braille Advancements: Automatic ”Reading” by a New Optical Braille Recognition System ”OBR” and Objective Dot and Paper Quality Evaluation. In *CSUN’99 Technology and Persons with Disabilities, Los Angeles, March 15–20*. http://www.dinf.org/csun_99/csun99.htm. 4

[Haralick et al., 1991] Haralick, R. M., , and Shapiro, L. G. (1991). Glossary of computer vision terms. *Pattern Recognition*, 24:69–93. 8, 31

[Haralick and Shapiro, 1992] Haralick, R. M. and Shapiro, L. G. (1992). *Computer and Robot Vision*, volume 1. Addison Wesley. 10

[Hartelius, 1996] Hartelius, K. (1996). *Analysis of Irregularly Distributed Points*. PhD thesis, Institute of Mathematical Modelling, Technical University of Denmark. 11, 65, 90

[Hartung, 1989] Hartung, J. (1989). *Multivariate Statistik*. R. Oldenburg Verlag München Wien. 42

[Heikkilä and Silvén, 1997] Heikkilä, J. and Silvén, O. (1997). A Four-step Camera Calibration Procedure with Implicit Image Correction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1106–1112. 11

[Herrero et al., 2001] Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierachical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126–136. 61

[Huber, 1964] Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, pages 73–101. 38

[Huber, 1981] Huber, P. J. (1981). *Robust Statistics*. John Wiley and Sons. 36

[Illingworth and Kittler, 1988] Illingworth, J. and Kittler, J. (1988). A Survey of the Hough Transform. *Computer Vision, Graphics and Image Processing*, 44(1):87–116. 9

[Jain, 1986] Jain, A. K. (1986). *Fundamentals of Digital Image Processing*. Prentice-Hall. 16

[Jain et al., 2000] Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition, a review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(1):4–37. 61

[Jain et al., 1995] Jain, R., Kasturi, R., and Schunck, B. B. (1995). *Machine Vision*. McGraw-Hill. 10

[Johnston et al., 1990] Johnston, R. J., Picket, S. C., and Barker, D. L. (1990). Autoradiography using storage phosphor technology. *Electrophoresis*, 11:355–360. 3

[Jolion and Rosenfeld, 1994] Jolion, J. M. and Rosenfeld, A. (1994). *A Pyramid Framework for Early Vision*. Kluwer. 30

[Jureckova, 1984] Jureckova, J. (1984). *Robust Statistical Procedures*, volume 4 of *Handbook of Statistics*, chapter M-, L- and R-estimators, pages 463–485. Elsevier. 36

[Jureckova and Sen, 1996] Jureckova, J. and Sen, P. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations*. John Wiley and Sons. 36

[Kaifel et al., 2000] Kaifel, T., Schiekel, C., and Kämpke, T. (2000). Spotting approaches for biochip arrays. In *15th International Conference on Pattern Recognition*, volume 4, pages 356–361. IAPR, IEEE Computer Society. 68

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons. 61

[Kegelmeyer et al., 2001] Kegelmeyer, L. M., Tomascik-Cheeseman, L., Burnett, M. S., van Hummelen, P., and Wyrobek, A. J. (2001). A groundtruth approach to accurate quantitation of fluorescence microarrays. In *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE - Progress in Biomedial Optics and Imaging*, pages 35–45. SPIE, International Society for Optical Engineering. 68, 69

[Kozal et al., 1996] Kozal, M. J., Shah, A., and Shen, N. (1996). Extensive polymorphism observed in hiv-1 clade protease gene using high-density oligonucleotide arrays. *Nature Medicine*, 2:753–759. 57

[Lander, 2001] Lander, E. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921. 58

[Levine, 1993] Levine, A. J. (1993). The Tumor Suppressor Genes. *Annual Reviews of Biochemistry*, 62:653–683. 57

[Lewin, 1997] Lewin, B. (1997). *Genes VI*. Oxford University Press. 3, 48, 49, 52, 53, 58, 59

[Lukashin and Fuchs, 2001] Lukashin, A. V. and Fuchs, R. (2001). Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, 17:405–414. 61

[MacBeath and Schreiber, 2000] MacBeath, G. and Schreiber, S. L. (2000). Printing Proteins as Microarrays for High-Throughput Function Determination. *Science*, 289(5485):1760–1763. 4, 57

[Mallat, 1999] Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic Press. 8

[Maronna, 1976] Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1):51–67. 40

[Maybank and Faugeras, 1992] Maybank, S. J. and Faugeras, O. D. (1992). A theory of self-calibration of a moving camera. *International Journal of Computer Vision*, 8(2):123–152. 10

[Meier-Ewert et al., 1998] Meier-Ewert, S., Lange, J., Gerst, H., Herwig, R., Schmitt, A., and *et al.*, J. F. (1998). Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Research*, 26(9):2216–2223. 60

[Meier-Ewert et al., 1993] Meier-Ewert, S., Maier, E., Ahmadi, A. R., Curtis, J., and Lehrach, H. (1993). An automated approach to generating expressed sequence catalogues. *Nature*, 361:375–376. 60

[Microsoft, 2001] Microsoft (2001). Microsoft® Encarta® Online Encyclopedia 2001. http://encarta.msn.com. 2

[Minor and Sklansky, 1981] Minor, L. and Sklansky, J. (1981). The Detection and Segmentation of Blobs in Infrared Images. *IEEE Trans. Systems, Man and Cybernetics*, 11:194–201. 2, 9

[Mori et al., 1992] Mori, S., Suen, C. Y., and Yamanoto, K. (1992). Historical review of OCR research and development. *Proceedings IEEE*, 80:1029–1058. 1

[Mullis, 1987] Mullis, K. B. (1987). Process for amplifying nucleic acid sequences. *U.S. Patent*, 4,683,202. 58

[Nalwa, 1993] Nalwa, V. S. (1993). *A Guided Tour to Computer Vision*. Addison Wesley. 3

[Nayar and Poggio, 1996] Nayar, S. K. and Poggio, T., editors (1996). *Early Visual Learning*. Oxford University Press. 16

[Noordmans and Smeulders, 1998] Noordmans, H. J. and Smeulders, A. W. M. (1998). Detection and Characterization of Isolated and Overlapping Spots. *Computer Vision and Image Understanding*, 70(1):23–35. 7, 8, 33

[Otsu, 1979] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Systems, Man and Cybernetics*, 9(1):62–66. 70

[Pease et al., 1994] Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., and Fodor, S. P. A. (1994). Light-Generated Oligonucleotide Arrays for Rapid DNA Sequence Analysis. *Proceedings of thge National Academy of Sciences of the USA*, 91:5022–5026. 56

[Pratt, 1991] Pratt, W. K. (1991). *Digital Image Processing*. Wiley. 7

[Press et al., 1992] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press. 24, 31, 32

[Rewo, 1984] Rewo, L. (1984). Enhancement and Detection of Convex Objects Using Regression Models. *Computer Vision, Graphics and Image Processing*, 25(2):257–269. 3, 7

[Rimberg et al., 1997] Rimberg, A. J., Ho, T. R., Kurdak, C., and Clarke, J. (1997). Dissipation-Driven Superconductor-Insulator Transition in a Two-Dimensional Josephson-Junction Array. *Phyical Review Letters*, 78(13):2632–2635. 4

[Rissanen, 1987] Rissanen, J. (1987). Minimum description length principle. *Encyclopedia of Stastistic Sciences*, (5):523–527. 38

[Rothwell, 1995] Rothwell, C. A. (1995). *Object Recognition Through Invariant Indexing*. Oxford University Press, New York. 11

[Sahoo et al., 1988] Sahoo, P. K., Soltani, S., Wong, A. K. C., and Chen, Y. C. (1988). Survey of thresholding techniques. *Computer Vision, Graphics, and Image Processing*, 41(2):233–260. 9

[Schena, 2000] Schena, M., editor (2000). *Microarray Biochip Technology*. Eaton Publishing, BioTechniques Books Division. 4

[Schmidt, 1990] Schmidt, W. A. C. (1990). Modified Matched Filter for Cloud Clutter Suppression. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(6):594–600. 7

[Sher and Rosenfeld, 1989] Sher, C. and Rosenfeld, A. (1989). Detecting and Extracting Compact Textured Regions Using Pyramids. *Image and Vision Computing*, 7:129–134. 2

[Shneier, 1983] Shneier, M. (1983). Using Pyramids to Define Local Thresholds for Blob Detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 5(3):345–349. 2, 9

[Slama, 1980] Slama, C. C., editor (1980). *Manual of Photogrammetry*. American Society of Photogrammetry, Falls Church, Virginia, fourth edition. 10

[Sonka et al., 1999] Sonka, M., Hlavac, V., and Boyle, R. (1999). *Image Processing, Analysis, and Machine Vision*. Brooks/Cole Publishing Company, second edition. 1, 8, 9, 10

[Steinfath et al., 2001] Steinfath, M., Wruck, W., Seidel, H., Lehrach, H., Radelof, U., and O'Brien, J. (2001). Automated image analysis for array hybridization experiments. *Bioinformatics*, 17(7):634–641. 11, 66

[Strickland and Hahn, 1996] Strickland, R. N. and Hahn, H. I. (1996). Wavelet Transforms for Detecting Microcalcifications in Mammograms. *IEEE Trans. Medical Imaging*, 15(2). 8

[Takahashi et al., 1997] Takahashi, K., Nakazawa, M., and Watanabe, Y. (1997). DNAinsight: An Image Processing System for 2-D Gel Electrophoresis of Genomic DNa. In *Genome Informatics*, volume 8, pages 135–146. Universal Academy Press. 3

[Takahashi and Watanabe, 1998] Takahashi, K. and Watanabe, Y. (1998). An Analysis System for Two-Dimensional Gel Electrophoresis Images of Genomic DNA. In *Proc. International Conference on Pattern Recognition*, page SA12. 3

[Thacker and Lacey, 2000] Thacker, N. and Lacey, T. (2000). *Tina 4.0 User's Guide*. Neuro-Imaging Analysis Centre, Medical School, University of Manchester. 10

[Trier et al., 1996] Trier, Ø. D., Jain, A. K., and Taxt, T. (1996). Feature-Extraction Methods for Character-Recognition: A Survey. *Pattern Recognition*, 29(4):641–662. 1

[Trucco and Verri, 1998] Trucco, E. and Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall. 10

[Tsai, 1987] Tsai, R. Y. (1987). A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf cameras and lenses. *IEEE Trans. Robotics and Automation*, 3(4). 10

[Ullman, 1996] Ullman, S. (1996). *High-level Vision*. A Bradford Book. MIT Press. Object Recognition and Visual Cognition. 1

[van der Heijden, 1995] van der Heijden, F. (1995). Edge and Line Feature Extraction Based on Covariance Models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(1). 7

[van der Heijden et al., 1997] van der Heijden, F., Apperloo, W., and Spreeuwers, L. (1997). Numerical Optimization in Spot Detector Design. *Pattern Recognition Letters*, 18:1091–1097. 2, 7

[Venter et al., 2001] Venter, J. C., Adams, M. D., and Myers, E. W. (2001). The Sequence of the Human Genome. *Science*, 291(5507). 58

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. C. (1953). A structure for DNA. *Nature*, 171:737–738. 49

[Watt, 1994] Watt, A. (1994). *3D Computer Graphics*. Addison-Wesley. 23

[Winkler, 1995] Winkler, G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer Verlag. 65

[Yang et al., 2000] Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2000). Comparison of methods for image analysis on cDNA microarray data. Technical Report 584, Department of Statistics, University of California at Berkeley. 70

[York et al., 2000] York, D. G., Adelman, J., and Anderson, J. E. (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3):1579–1588. http://www.sdss.org. 4

[Zhang, 1995] Zhang, Z. (1995). Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting. Technical Report 2676, INRIA -Institut de Recherche en Informatique et en Automatique. 38, 39

[Zhang, 1999] Zhang, Z. (1999). Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In *Proc. Seventh IEEE Conf. on Computer Vision*, volume 1, pages 666–673. 10

[Zhou et al., 2000] Zhou, Y.-X., Kalocsai, P., Chen, J.-Y., and Shams, S. (2000). Information Processing Issues and Solutions Associated with Microarray Technology. In *Microarray Biochip Technology*, chapter 8. Eaton Publishing, BioTechniques Books Division. 69, 70

[Zhou et al., 2001] Zhou, Z. Z., Stein, J. A., and Ji, Q. Z. (2001). GLEAMS: A Novel Approach to High Throughput Genetic Micro-Array Image Capture and Analysis. In *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE - Progress in Biomedial Optics and Imaging*, pages 13–23. SPIE, International Society for Optical Engineering. 11, 66, 70