Bayesian Learning and Regularization for Unsupervised Image Restoration and Segmentation

vorgelegt von
Master of Science
Hongwei Zheng
aus China

von der Fakultät IV - Elektrotechnik und Informatik
der Technische Universität Berlin
zur Erlangung des Akademischen Grades

Doktor der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr.-Ing. Thomas Sikora
Gutachter:	Prof. Dr.-Ing. Olaf Hellwich
Gutachter:	Prof. Dr. Manfred Opper

Tag der wissenschaftlichen Aussprache: 20 Juli 2007

Berlin 2007
D 83

Acknowledgements

During the course of my PhD studies I have been fortunate enough to benefit from inspiring interactions with people in the mathematical image processing, computer vision, artificial intelligence and machine learning communities and I am delighted to be able to acknowledge these here.

My thanks go first of all to my advisor, Professor Dr.-Ing Olaf Hellwich, for his wisdom, support and encouragement. Professor Hellwich guided me through my Phd research, and whose comments often sparked my deeper interest in directions which I would have ignored otherwise, or forced me to re-consider intuitive argumentations more carefully. I have also benefited from his wide overview of relevant literatures and many novel strategies that I have been interested in. I hope that in my own career I will be able to follow the role model that he sets – not only his endless curiosity about relevant research areas, his brilliant insights into complicated phenomena but also in terms of his integrity and perseverance. I wish to thank Professor Dr. Manfred Opper for agreeing to be a reviewer of my thesis. His lecture on "Künstliche Intelligenz" inspired me to apply some novel statistics strategies and approaches into the research of computer vision. His suggestions and comments significantly improved the quality of this work. I am also very grateful to Professor Dr. Harmut Ehrig for introducing his new book of algebraic graph transformation and related discussions for the application of this theory into computer vision.

An important role was played also by several professors that I was lucky to meet during my previous studies, professor Dr.-Ing Michael Hahn, professor Dr.-Ing Detrich Schröder in Stuttgart and professor Dr.-Ing Kurt Kubik in Queensland, Professor Dr.-Ing Förstner in Bonn, Professor Dr.-Ing Jörg Albertz in TU-Berlin. In particular, I feel also grateful to my colleagues for their support, help and useful comments: Stephan Gehrke, Gerhard König, Gerold Baumhauer, Iliana Theodoropoulou, Hartmut Lehmann, Yasemin Kuzu, Olaf Sinram, Anke Bellmann, Esra Erten, Oliver Gloger, Stéphane Guillaso, Matthias Heinrichs, Marc Jäger, Sandra Mannheim, Maxim Neumann, Andreas Reigber, Volker Rodehorst, Saquib Sarfraz, Adam Stanski, Tim Suthau, Ulas Yilmaz, Wenju He, Ronny Hnich, especially Adam and I share the same work room and we have a lot of interesting discussions. I will never forget the joyful and rewarding days spent with them.

I want to thank a number of researchers for hospitality and stimulating discussion during my study, wherever they are, Eckart Michael, Tomas Minka, Andrew Blake, Yunmei Chen, Stacey Levine, Luminita Vese, Thomas Brox, Stephan Didas, Wei Du, Xiaojin Zhu, Zhuowen Tu, Yongsheng Pan, Guoyan Zheng, Dengyong Zhou. There are also a lot of friends have given me invaluable encouragements, Thomas Huang, Zhengyou Zhang, Xiaoyi Jiang, Zheming Lu, Tieniu Tan and Heung-Yeung Shum when we met in different conferences and places.

I should also thank all my friends. Last, but by no means least I would like to thank my parents for their love and nurture in all my years, especially my wife for her love, encouragement and endless patience with me.

The work in this thesis was carried out at our computer vision and remote sensing group. I am also grateful to our institute and TU-Berlin for generous travel grants.

Zusammenfassung

Die Herausforderung bei der blinden Bildrestauration ist, aus einem beobachteten Bild das ursprüngliche Signal eindeutig wieder herzustellen, ohne Nutzung einer zusätzlichen Informationsquelle. Die Schwierigkeit liegt vor allem bei den notwendigen Statistiken und Optimierungen und der praktische Nutzen in Anwendungen der Bildanalyse und Bildverarbeitung. Diese Arbeit leistet drei wichtige Beiträge zur blinden Bildrestauration und Segmentierung, die im Folgenden aufgeführt sind.

Der erste Teil dieser Arbeit beschäftigt sich mit der systematischen Integration von statistischer Modellselektion, Bayesschem Lernen und Regularisierungstheorie in streng konvexen Optimierungsfunktionalen. Der vorgeschlagene Ansatz der Bayesschen Schätzung basiert auf Doppelregularisierung. Er integriert globale nicht-parametrische Modellselektion, lokal parametrische Unschärfekernoptimierung für parametrische Unschärfeidentifikation und Dekonvolution. Ein guter initialer Unschärfekern wird durch eine konvexe Regularisierung geschätzt. Während der iterativen Doppelregulierung wird die geschätzte Pointspread-Funktion als Vorwissen für die nachfolgende iterative Schätzung des Bildes und umgekehrt verwendet. An dieser Stelle werden auch einige neue Ideen vorgestellt, welche die Qualität der Unschärfeerkennung in Bezug auf unterschiedliches Rauschen in den einzelnen Bildern oder in großen Videodateien verbessern.

Der zweite Teil dieser Arbeit widmet sich der Verbesserung der Wiedergabetreue und Qualität von wiederhergestellten Bilder, speziell in entrauschten Bildern. Hierbei werden verschiedene lineare Wachstumsfunktionale zur Bildverarbeitung genauer behandelt und auf den Raum von Funktionen mit beschränkter Variation angewendet. Basierend auf diesen Funktionalen wird eine Bayessche Schätzung zur datengetriebenen Bildrekonstruktion durch Variationsrechnung entwickelt und implementiert. Die Performanz wird über die numerische Approximation von hyperbolischen Erhaltungssätzen, selbstregelnde Diffusionsoperatoren, adaptive Anpassung von Regularisierungsparametern und optimale Stoppzeiten des Prozesses kontrolliert. Dieser Ansatz übertrifft nicht nur die meisten bisher bekannten Ansätze, sondern erlaubt auch eine hochpräzise und nach menschlichen Kriterien exakte Bildwiederherstellung.

Der dritte Teil dieser Arbeit beschäftigt sich mit dem allgemeineren Unschärfeproblem unter realen Bedingungen, beispielsweise für nur teilweis unscharfe Bilder einschließlich stationärer und nicht-stationärer Unschärfekerne. Eine Vielzahl vorhandener Segmentierungsverfahren erfüllt die Aufgabe der Identifikation und Segmentierung von unscharfen Regionen nicht zufriedenstellend. In Anlehnung an spektrale Bildsegmentierungskonzepte durch Clusteranalyse und deren zugrunde liegende Verbindung zur Regularisierungstheorie, wurde ein regularisierter spektraler Clusteringansatz auf diskreten Graphenräumen entwickelt, der gute Ergebnisse erzielt. Infolgedessen können die identifizierten und segmentierten unscharfen Regionen in einem auf Variationsrechnung basierendem Bayesschen Lernframework mit einem Prior aus natürlichen Bildstatistiken wiederhergestellt werden. Das üblicherweise nicht berechenbare inverse Lernproblem wird durch die variationale Bayessche Lernmethode berechenbar. Nicht-uniforme unscharfe Bilder können optimal rekonstruiert werden, ohne scharfe Regionen und Objekte zu zerstören.

Um den vorgeschlagenen Ansatz zu validieren, wurde die Leistungtähigkeit an unterschiedlichen Bildern demonstriert. Die Resultate zeigen, dass die vorgeschlagenen Algorithmen robust und leistungsfähig gegenüber Bildern sind, die in verschiedenen Umgebungen, mit unterschiedlichen Arten von Unschärfe und Rauschen, erzeugt wurden. Außerdem können diese Methoden auf Grund ihrer Flexibilität leicht angewendet werden, um verschiedene Probleme in der Bildverarbeitung und Bildanalyse zu lösen.

Abstract

The challenge of blind image restoration is to uniquely define the restored signals from only the observed images and without any other information. It gives opportunities not only for valuable contributions in the theoretical statistics and optimization but also for the practical demands in image processing and computer vision. The main contribution of this thesis is in the fields of image deblurring, denoising, image reconstruction and segmentation in low level vision.

The first part of this thesis is dedicated to the systematic integration of statistical model selection, Bayesian learning and regularization theory in a strictly convex optimization functional. The proposed approach is in a double regularized Bayesian estimation framework for parametric blur identification and image deconvolution. A good initial point spread function (PSF) blur kernel is estimated for convex regularization. During the iterative double regularization, the estimated PSF is prior knowledge for the next iterative estimation of the image, and vice visa. In this context, we also introduce several new ideas that improve the quality of blur identification with respect to other sources of image degradation.

The second part of this thesis is devoted to improving the fidelity and quality of restored images, especially in the context of image denoising and deblurring. It is in this part that we introduce and extend several linear growth functionals to the space of functions of bounded variation (BV) for image processing. Based on these functionals, a data-driven variational image restoration functional in a Bayesian learning framework has been designed and implemented in the BV space. The performance is controlled via numeric approximation in terms of hyperbolic conservation laws, self-adjusting diffusion operators, adaptive adjustment of regularization parameters and optimal stopping time of process. The approach does not only outperform most approaches in the literature, but also allows to achieve high-fidelity and human perceptual image deblurring, denoising and image reconstruction.

The third part of this thesis considers a more general blur problem in the real world, i.e., nonuniform blurred (e.g., partially-blurred) images including stationary and nonstationary blur kernels. There are numerous existing segmentation approaches that do not achieve satisfactory results for the identification and segmentation of blurred regions or objects. Inspired by spectral graph theory and their underlying connections with regularization theory, we develop a regularized spectral clustering approach on discrete graph spaces that achieves good performance. Also, the blur kernel can be identified in high-accuracy in a tractable variational Bayesian learning framework. The generalized parametric PSF prior and natural image statistics based image prior distribution are used for blur kernel estimation. As a consequence, nonuniform blur degraded images can be optimally restored without degrading unblurred regions and objects.

In order to validate the proposed approaches, we demonstrate good experimental performance in a number of contexts. The results show that the proposed algorithms are robust and efficient in that they can handle images that are formed in various environments with different types of blur and noise. Furthermore, because of the flexibility of these methods, they can be easily applied to solve a number of other problems in image processing and computer vision.

Contents

1 Introduction

Vision is the art of seeing thing invisible. - Jonathan Swift

1.1 Motivation

In the modern digital imaging world, CCD, CMOS, tomography, MRI, microscope, space telescope and radar data are often degraded due to blur and noise. These degradations heavily influence the implementation, automation, robustness and efficiency of many visual systems. For example, in visual surveillance systems, blurred frames or blurred objects in video sequences influence the efficiency of such systems. Moreover, these degradations also influence the reconstruction of high-resolution and high-fidelity images for display systems, image matching and registration for 3D reconstruction systems, and low-level vision processing for classification and recognition systems etc.

The primary goal of image restoration is to recover lost information from a degraded image and obtain the best estimate to the original image. The challenge of blind image restoration is to uniquely define the convolved signals only from the observed images without any other information. It gives opportunity for valuable contributions in the theoretical statistics and optimization for ill-posed inverse problems but also for the practical demands in image processing and computer vision. Its applications include tomography, stereology, medical imaging, multimedia processing. Compared with classical approaches, blind image restoration for entirely- and partially-blurred (stationary and non-stationary blur) noisy images becomes an important research topic, e.g., shown in Fig. 1.1. Its growing desirable features pose many new challenges to researchers in the field of mathematics, pattern recognition and computer vision.

Hadamard introduced the notion of ill-posedness in the field of partial differential equations [102]. A problem is well-posed when a solution exists, is unique and depends continuously on the initial data. It is ill-posed when it fails to satisfy at least one of these criteria. Ill-posed problems have been a mathematical curiosity for many years. Nowadays they arouse great interest since many problems of practical interest turned out to be ill-posed inverse problems, e.g., such as blur identification, image restoration, segmentation, and the under-constrained scene analysis, object reconstruction.

Most inverse problems are ill-posed. According to Keller and Engl's description [129], [68], one calls *two* problems *inverse* to each other if the formulation of one problem involves the other one. Due to historic reasons, one might call one of these problems the *direct* problem, the other one the *inverse problem*. The *direct* problem is usually the simpler one or the one which was studied earlier. In the real world, if one wants to predict the future behavior of a physical system from knowledge of its present state and the physical laws (including concrete values of all relevant physical parameters), one will call this direct problem. Possible *inverse problems* are the determination of the present state of the system from future observations (i.e., the calculation of

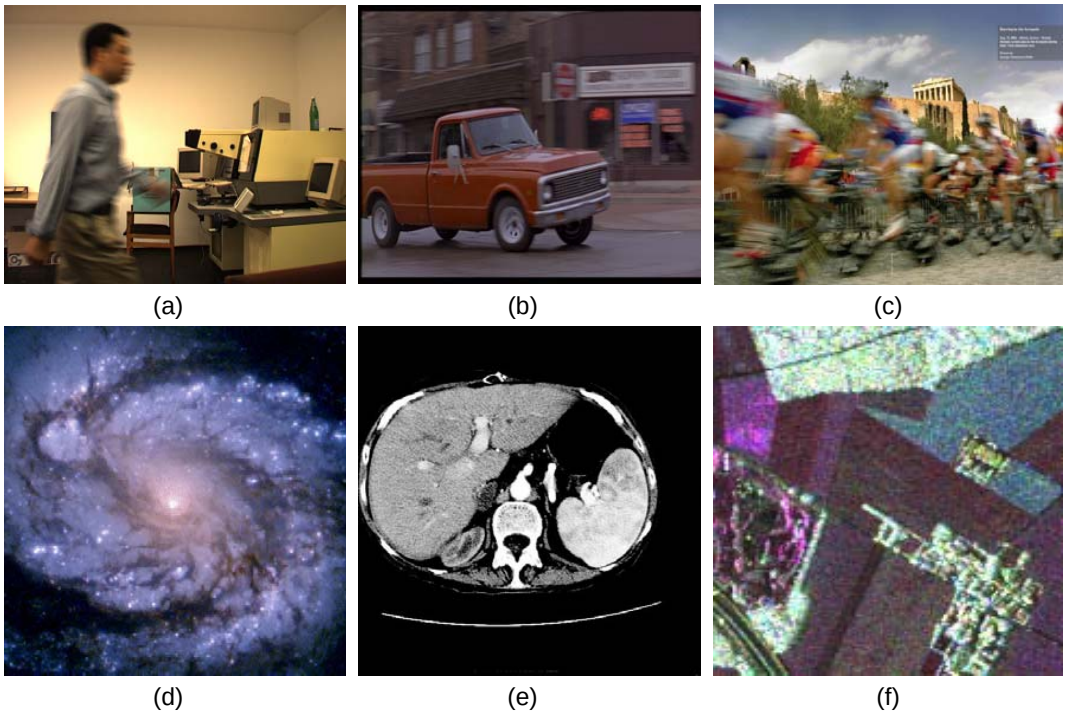


Figure 1.1: Entirely- and partially-blurred noisy images in real life environments. (a)(b)(c) Video data. (d) Space telescope data. (e) MRI data. (f) Synthetic aperture radar data

the evolution of the system backwards in time) or the identification of physical parameters from the observations of the evolution of the system (parameter identification). While the study of concrete inverse problems involves the question how to enforce existing, uniqueness, continuous solution by additional information or assumptions, as we can see, most such inverse problems often do not fulfill Hadamard's postulates of well-posedness.

The approach of solving these ill-posed inverse problems is how to learn about the underlying process mechanism of a physical or biology system (such as vision systems), how to influence and design a system via its present state or parameters in order to steer it to a desired state. Therefore, we might say the inverse problems are concerned with underlying and determining causes for a desired or an observed effect and phenomenon, e.g., from an observed image to estimate, reconstruct, and recognize the desired images or objects.

Blind image restoration and image reconstruction including blur identification, deblurring and denoising can be considered as an ill-posed inverse problem [43]. For example, in real environments, blurring and noising occur naturally but deblurring and denoising need extra work on the system. Mathematically, image deblurring is intimately connected to back-ward diffusion processes, e.g., inverting the heat equation, which are notoriously unstable. As inverse problem solvers, deblurring and denoising models therefore crucially depend upon proper regularization which secures existence, stability and uniqueness of restoration.

Theoretical Perspective

From a theoretical perspective, regularization is the approximation of an ill-posed problem by a family of neighboring well-posed problems. The regularization theory [241] provides a convenient

way to solve ill-posed problems and to compute solutions that satisfy prescribed smoothness constraints. The studies of regularization theory can be found in inverse problems by Bertero [24], Engl [68], Groetsch [100], [101], Hansen[105], [239], early vision by Katsaggelos [127], Hellwich [111], Schnörr [221], and Weickert [259], and brain and cognitive sciences by Poggio et al. [91], [192], [193], [195] etc. This formalism has been recognized as a unified framework for studying several problems in computer vision and image processing. Regularization is especially crucial for vision problems, and present numerous challenges as well as opportunities for further statistical and mathematical modeling.

In theoretical statistics and statistical learning theory, regularization is usually interpreted either in Bayesian terms, or as some form of Stein-like shrinkage [64]. It covers important topics of traditional statistics, especially in discriminant analysis, regression analysis and the density estimation problem [248]. Other disciplines, notably image processing and computer vision, offer more pragmatic and interesting perspectives, and often furnish more aggressive attacks on computational aspects, e.g., numerical computing scheme based on physical laws. Recent developments in the statistical literature offer promising new approaches to the fine-tuning of regularization techniques, particularly in the selection of regularization parameters [94], [106], regularization operators and optimization [94]. Regularization has been extended and discussed in discrete graph space based regularization [300], kernel based regularization [222], [248], semi-supervised regularization [309] and so on.

In applied statistics, regularization often identified as "penalty-based methods" or "soft thresholding", is associated primarily with nonparametric regression and density estimation. In such cases, it is often referred to rather imprecisely as "smoothing". One of the primary objectives of stochastic routes to regularization would be to encourage a further diversification of smoothing objectives. Several techniques exist, such as variational regularization using a convex stabilizer. In reality, *a priori* knowledge often requires non-convex functionals, resulting in a no longer convex solution space. Therefore, stochastic methods or "converting non-convex to convex" strategy are needed to escape from local minima [189].

Another regularization method consists of choosing a discrete solution space with finite dimensions and imposing generic constraints. This may seem like a harsh restriction at first but this isn't really so. Indeed, our real world is highly structured and is constrained by physical laws to a number of basic patterns. In computer vision, people gradually built textons (the atom of visual perception) [308] and "bag-of-words" [228], [72], for recognition or semantic searching. The apparent complexity of our environment is produced from this limited vocabulary by compounding these basic forms in different combinations. If the intrinsic complexities of our environment were approximately the same as its apparent complexity, there would be no lawful relations and no intelligent prediction. It is the internal structuring of real environment that allows us to reason successfully using simplified descriptions [274]. Different techniques in engineering, statistics, or biology, have been transferred and described to determining these generic constraints [140]. When the *a priori* assumptions are violated in specific instances, the obtained solution may not correspond to the real world situation. Therefore, extracting and modeling of descriptive prior information from uncertainty become important.

A more fundamental problem that arises in inverse problems is the scale problem which includes related localization, and orientation problems in scale-space [146], [147], and the concepts of inverse scale spaces [217], [218]. In other words, which scale is the right resolution to operate? Scale-space methods are asymptotic formulations of the Tikhonov regularization [241]. These techniques consider the behavior of the result across a continuum of scales. From the

viewpoint of regularization theory, the concept of scale is related quite directly to the regularization parameter. It is tempting to conjecture that methods used to obtain the optimal value of regularization parameters may provide the optimal scale which is associated with the specific instance of certain problems.

Practical Perspective

A wide range of ill-posed problems concerned with recovering information from indirect and usually noisy measurements arise from image processing, stereology, computerized tomography (J. Radon), medical imaging, inverse scatting, inverse heat conduction problems, geophysics, geodesy, image deconvolution, and related vision problems.

First, regularization theory offers a unifying perspective on these diverse ill-posed inverse problems. Given an individual blurred image, one is interested in more information that can be extracted, restored from these blurred images or blurred regions or objects. However, because of the growing demands for solutions and the complexity and uncertainty of problems, the integration of statistic learning and regularization is a preferred approach for achieving a high-quality solution. This approach can be formulated as a convex optimization problem and is solved by numerical iterative schemes. The cost function is a combination of the error learning term and the stabilizing term which can be optimized in a convex regularization approach. The stabilizing term usually reflects physical constraints arising within the application for which the proposed solution is a model, and acts by limiting the energy of the solutions.

Second, we can obtain the best results with the blind deconvolution algorithms for most existing blurred images and signals. The reason is that the blind deconvolution algorithms do not use the measured point spread function (PSF) for the other algorithms, but approximate the PSF iteratively. This is due to the fact that measured PSFs itself contain noise and therefore the deconvolution is biased by the noisy PSF. Although the blind deconvolution does not use any information of the actual optical system, it yields better results, since the PSF is approximated and not influenced by noise.

Third, for non-uniform or non-stationary (e.g., partially-blurred) blurred image restoration, we need to restore the blurred regions or objects without influencing unblurred regions or objects in an image. The proposed regularization in discrete graph spaces is formulated in combinatorial optimization which allows to segment and identify blurred regions or objects. More important, the interesting analogy between regularization and spectral graph theory [51] brings crucial insights to the understanding of eigenvalues, eigenvectors and the Laplacian of graphs. Our proposed regularized spectral graph clustering approach on discrete graph spaces is a novel approach which can directly get global image understanding using sparse local patches in cluttered images. It is also a novel approach towards the perceptual image segmentation for various images. This approach also suggests that only incorporated segmentation work can become meaningful and more useful in practical environments.

Furthermore, such blurred images are mostly non-stationary and non-uniformly blurred. It means that we cannot directly to represent these real blur kernels using some simple parametric blur kernels. Therefore, based on previous work, we extend our previous double regularized Bayesian estimation to a more tractable variational Bayesian learning approach. This approach allows the true posterior to be approximated by a simpler approximate distribution for which the required inference are tractable. Moreover, natural image learning helps us find translation

and scale-invariant spatial prior distribution. In particular, the approach makes effective use of the natural image statistics through the whole variational learning scheme. Our experiments show that the results derived from the algorithm are superior to this type of blurred images. The scheme can be further extended to other types blurred image restoration in real environments. The approach can be used solved related kernel identification, pattern recognition and computer vision problems.

In summary, other important applications are that the integrated statistical learning and regularization approach can achieve accurate blur identification and image restoration in convex optimization. The optimization theory and methods can be reasonably applied to solve many kinds of vision problems, e.g., reconstruction, recognition and so on. These methods can also be easily extended to data mining, semantic data searching [139] and related model selection problems. The relation to other corresponding problems shows that the previously described applications are only some examples based on this proposed mathematic framework. Furthermore, there are much more vision, image processing, data mining and related problems that can be solved via the extension of our proposed methods.

1.2 Problem Statement

The goal of the present work is to contribute to statistical learning, especially Bayesian learning and regularization approaches for solving the ill-posed inverse problems in image processing, pattern recognition and computer vision.

1.2.1 Ill-posedness of Image Deblurring and Denoising

To recover a sharp image from its blurry and noisy observation is a problem known as image deblurring and denoising. The observation of blurring and noise is one way for blur understanding and deblurring and denosing. Through the observation, these underlying natural phenomena can help us design more robust and flexible deblurring and denoising models. Normally, deblurring and denoising can be taken into account and processed respectively. However, in most situations, deblurring and denoising must be processed cooperatively due to the complexity of blur and noises. Chan and Shen [43] provide a general and sound overview for the problem of deblurring. Here, we add some understandings on them.

Image Deblurring

1. **Deblurring is inverting lowpass filtering.** Blurring is one of the most important degradation processes for images and signals. For most real blurred images, power spectral densities in the frequency domain vary considerably from low frequency domain in the uniform smoothing region to medium and high frequency domain in the discontinuity and texture regions, and different blur in a given image has different magnitude and phase in the frequency domain. The high frequency discontinuities are often diminished by vanishing blur multipliers. As a consequence of deblurring, we need to multiply the approximate reciprocals of vanishing multipliers. However, these multipliers are conceivably unstable to noises and other high-frequency perturbations in the image data.

2. **Deblurring is Shannon information increasing and entropy decreasing.** The goal of deblurring is to reconstruct the detailed image features from a modified blurred image. Therefore, from an information theoretic and statistic mechanics points of view, deblurring is a process to increase Shannon information [225] and decrease entropy. Based on the second law of statistical mechanics [88], blur is natural and easily takes place but deblurring process never occurs naturally and extra efforts need to be contributed on the system.

3. **Deblurring is backward diffusion.** Following the PDE theory, an image blurred with a Gaussian kernel is equivalent to running the heat diffusion equation for some finite duration with the given image as the initial data. Thus, deblurring is the inverse process of heat diffusion. Moreover, image diffusion corresponds to the Brownian motions of initial ensemble of particles in the stochastic processing domain. The blurring process is a random spreading process and the deblurring process amounts to reversing an irreversible random spreading process, which is ill-posed.

4. **Deblurring is inverting compact operator.** A blurring process is typically a compact operator [43]. A compact operator maps any bounded set to a much better behaved set according to the associated Hilbert or Banach norms. Compact operators allow us to generalize classical results for operator operations in finite-dimensional normed spaces to infinite-dimensional normed spaces via approximation and a limiting process. Compactness plays a key role in functional analysis. Intuitively, a compact operator has to mix spatial information or introduce some coherent structures. These coherent structures are often realized essentially by dimensionality reduction using vanishing eigenvalues or single values. Therefore, to invert a compact operator is equivalent to de-correlating spatial coherence or reconstructing the formerly suppressed dimensions of features and information during the blurring process. For example, the equation $g = Hf + \eta$ has often either no solution or infinite solution of H and f with an observed image g. An unique meaningful solution has to be estimated in some proper way.

Influences of Blur Identification

Since most degraded images suffer from unknown disturbance, unknown blur information, and unknown noises in the real world, blind image deblurring becomes more difficult. Blur also influences the automation, robustness and efficiency of many visual systems in some respects. In visual surveillance systems, blurred frames or blurred objects in video sequences influence the efficiency of such systems. During the 3D reconstruction from uncalibrated video data, freely taken digital video sequences may have some kind of blur. Those blurred images can heavily influence the next processing step, e.g. feature based image matching.

Recent research connected with the blind image deconvolution (blur identification and deblurring) problem has shed light on the characteristics of the image blur or point-spread function (PSF) and especially its dimensions. There are a lot of assumptions for the process of blind image deconvolution, for example, the image background encompasses at least blur-invariant, uniform blurred, and so on. The true image can be restored up to a complex constant using the inverse PSF, given a blurred image free of noise. If the blurred image is contaminated with noise, which gives rise to artifacts, the technique would be rendered useless. This algorithm can be extended to mitigate noise considering the symmetric nature of most PSFs. In general, we

can accurately identify the blur, then we can restore it. However, there are several limitations that lead to unsuccessful cases. Some limitations in restoring blurred images are summarized in the following:

1. **The point spread function (PSF) of the blur, in general, varies spatially within an image.** It is the main limitation (it is also called non-stationary blur), since blur identification at every pixel uses the pixels within a neighborhood of that pixel. One assumption is that the blur kernel varies slowly in spatial coordinates. However, sometimes, it is even more difficult to justify this assumption. There is a trade off between having a big enough window for blur identification, and the validity of the assumption that the PSF is stationary within the window. The extent of the PSF should be represent the blur of the sampling region. Therefore, in some sense, this poses a limitation in the processing of spatially variant blurs.

2. **Incorporating nonlinear sensor characteristics into blur identification and image deconvolution procedures.** Normally, through varying image and PSF models, improved restorations can be obtained. It is particularly important for spatially variant PSFs that the models change accordingly. This is in contrast to the image model for which restorations are fairly insensitive. An important consideration for adaptive filtering is that the regions must contain sufficient data for the identification of model parameters. For example, as the PSF size increases, the amount of image data used for identification must also increase.

3. **From optimization point of view, one of the problems with the identification technique is the existence of local optima.** In some cases these suboptima correspond to minimum phase and non-minimum phase parameterizations of the PSF. One technique which will avoid minimum phase PSF identification is to assume PSF symmetry, whenever possible. In other cases, the image is restored with the identified parameters corresponding to the local optima and then making a choice by visual inspection of the restoration or by comparison of mean square errors and other measuring criteria.

4. **There are observation noises.** The presence of observation noise imposes a fundamental limitation on how much we can restore the resolution of the image before the filtered noise starts dominating the restored image. Also, the traditional film grain noise is usually signal-dependent, which causes theoretical difficulties. Those additive and multiplicative or impulsive noises also heavily influence the blur identification.

5. **There are ringing artifacts in restored images.** The ringing artifacts are visually objectionable. Moreover, they sometimes mask important image information. It is possible to suppress ringing artifacts to a certain extent. During the deconvolution and restoration process, periodic boundary condition easily generates ringing artifacts, while Neumann boundary condition do not have such artifacts. However, periodic boundary condition is suitable for large-size images restoration but needs adaptive filtering algorithms to eliminate the ringing effects.

1.2.2 Demands of Image Restoration and Segmentation

Image restoration has been investigated for several decades by now. Traditional deconvolution techniques are assumed to be linearly degraded by a convolution with a blurring kernel, which

is known a priori. The naive solution was to use inverse filtering, which was generalized to the optimal linear Wiener filter, to account for additive noise and zeros in the blurring kernel. More modern nonlinear deconvolution methods are used today based on statistical methods [71], [232], Tikhonov-regularization [241], or wavelet-based techniques[1], [233], among other methods.

Normally, the blur kernels are not known, some blind image deconvolution methods from Kattsaggelos [126], [18], Kundur [134], [161] try to achieve an adequate solution based on general assumptions with respect to the smoothness of images and the blurring kernels. Partial differential equations-based methods were also proposed achieving good restoration results [44], [208]. These image restoration methods can be classified into several categories depending on data sources, restoration targets and restoration methods.

1. **High fidelity image restoration.** Firstly, keeping high fidelity to the original data is based on the definition of image restoration. Restoration can be achieved by restoring all tiny and detailed discontinuities and structures of degraded images. The restored image can be gradually restored towards the original image. Restoration methods from spatial domain and frequency domain have different advantages in image restoration.

2. **Human visual perception image restoration.** Using the fact that human visual perception is adapted to the statistics of natural images and sequences, the classes of restoration models are not based on an image model but on a model of the human visual system. In particular, the non-linear model of early human visual processing is used to obtain locally adaptive image restoration without any a priori assumption on the image or noise.

3. **Simultaneous image identification, segmentation and restoration.** According to the target of simultaneous image restoration and segmentation, restoration then might not focus on the restoration of tiny structures of images but emphasize main discontinuities and structures of the restored images, e.g., partially-blurred image restoration.

In our work, to ensure our algorithms can be directly applied for different data sets such as tomography data, SAR data, etc., we do not improve the contrast or surface difference to improve the human visual perception results. For example, the contrast represents one of the key information in SAR data or tomography data, i.e., contrast in spatial domain is the amplitude information of such data in frequency domain. If the contrast is enhanced, the original data information will be modified or lost in such datasets. On the other hand, we keep the idea to ensure high fidelity of restored images so that the suggested algorithms can be applied for different data sources.

1.3 Proposed Approaches and Related Work

The goal of the present work is to contribute in the field of blur identification, image restoration and segmentation in computer vision. We are interested in improving state-of-the-art methods for ill-posed inverse problems. The underlying strategy is to integrate statistical learning and regularization in a convex optimization functional which is well-posed of minimization problems. Firstly, we propose a global nonparametric model selection with local parametric optimization in a Bayesian estimation based regularization approach. We focus on unsupervised Bayesian

model selection methods for sampling blurred regions and blur identification in a nonparametric density estimation approach. The identified blur kernel (without accurate parameters) can be an initial value in the adaptive weighted regularization. Subsequently, the locally parametric optimization can further improve the accuracy of the identified blur kernel.

On the other hand, we are also interested in high-fidelity image restoration. A data-driven image restoration method in the BV space is proposed and proved to be an "active" data-preserving image restoration approach. Furthermore, different from the traditional continuous regularization framework, we have implemented a regularization functional in discrete graph spaces. This method unifies regularization theory and spectral graph theory in a discrete regularization functional. This approach allows us to achieve partially-blurred image restoration without influencing unblurred regions or objects. By addressing statistic learning and regularization approaches, this thesis shall provide a systemic and Bayesian based variational energy optimization framework for the design of robust and high quality blur identification, image restoration and segmentation.

In order to specify our contributions in detail, a short introduction to Bayesian model selection for blur identification, data-driven variational image restoration in the BV space and discrete regularization in graph spaces is presented. Furthermore, some relevant work that is related to these fields of research are also presented.

1.3.1 Double Regularized Bayesian Estimation for Parametric Blur Identification

How to reliably and accurately identify blur kernels and their parameters in practical environments? Based on the theory of statistical learning, we classify such blur identification methods into global nonparametric estimation and local parametric optimization methods. Thereby, we integrate global nonparametric estimation and local parametric optimization for accurate blur identification.

Since statistic learning is a consequence of the ability to integrate information over time, the Bayesian estimation provides a basis for the design of learning algorithms. Bayesian estimation also provides a means of updating the distribution from the prior to the posterior in light of observed data. In theory, the posterior distribution captures all information inferred from the data about the parameters. This posterior is then used to make optimal decisions or predictions, or to select between models.

However, Bayesian approaches are often avoided by many statisticians, partly because there are problems for which a decision is made only once, and partly because there may be no reasonable way to determine the prior probabilities [63]. Neither of these difficulties seems to present some drawbacks in typical pattern recognition applications: For nearly all important pattern recognition problems we will have training data and we will use the recognizer more than once. For these reasons, the Bayesian approach will continue to be of great use in pattern recognition. The single important drawback of the Bayesian approach is the difficulty of determining and computing the conditional density functions. The multivariate Gaussian model may provide an adequate approximation to the true density, but there are some problems for which the densities are far from Gaussian. To simplify and decrease such difficulties in our work, two main ideas are used for the improvement of Bayesian estimation based blur identification,

1. Bayesian estimation expresses likelihood energy for approximate inferences that can be interpreted as a family of regularization functionals from Tikhonov [241], Geman and

Geman [85], [86], Osher [213], Mumford and Shah [173], Molina and Katsaggelos [169], Bishop et al. [28], Jordan et al. [121], Opper et al. [183], [181], Schölkopf et al. [222], Blake and Zisserman [33] and so on. We introduce Bayesian probability estimation to a convex regularization functional which computes the negative log-likelihood in an energy optimization manner. It therefore becomes possible to design a unified statistical learning and regularization system that can rely on prior knowledge and evidence. To ensure the global convergence, we formulate the regularization in a strictly convex functional. Following Bayesian paradigm, the true f, the PSF h and observed g in $g = hf + \eta$ on,

$$P(f, h|g) = \frac{p(g|f, h)P(f, h)}{p(g)} \propto p(g|f, h)P(f, h) \tag{1.1}$$

This formula utilizes prior information for getting a convergent posterior. Thereby, the search of prior knowledge $P(f, h)$ becomes crucial for the whole system.

2. Initially Inspired by Hellwich [111], Bishop [28], [27], Duda [63], Freeman [80], Geman and Geman [85], Szeliski [237], [238], Winkler [272], [273], Zhu et al. [303], [307] and so on, the prior knowledge should be descriptive information for measuring at the first step. Secondly, it may largely represent the uncertainty information. For the special case of blur identification, we design a blur kernel solution space based on characteristic properties of blur kernels and blurred images. Moreover, we employ the constraints of the restored image and the PSF as alternating priors for local parametric PSF adjustment. On the other hand, the use of prior information expresses an underlying idea in modeling a regularization approach with some physical constraints. Some physical constraints become generative information after statistical estimation, while some become nonnegative prior information (e.g., image and PSF are always positive). These constraints combined with the data information define a solution by trying to achieve smoothness and yet remain "faithful" to the data.

1.3.2 Data-Driven Variational Image Restoration in the BV Space

How to largely improve image deblurring and denoising in human visual perception? In the other words, how to represent an image in a mathematical model in the spatial domain and this model can further help us to reconstruct a high-fidelity image? A simple image including a white disk on a black background is not in any Sobolev space, but belongs to the BV space. The BV space is the space of functions for which the sum of the perimeters of the level sets is finite. Since the seminal work of Rudin, Osher and Fatemi (ROF) [213], the BV space based total variation (TV) functionals have been widely applied to image restoration, super-resolution, segmentation approaches and related early vision tasks, e.g., Mumford-Shah functional [173], modeling of oscillatory components [164], anisotropic diffusion [259], modeling of inpainting and super-resolution [42]. Closely related work are from Alvarez, Lion and Morel [7], [5], Demengel and Teman [59], Giusti [92], [93], [69], Vese [249], Auburt and Deriche et al. [15], [16], Chen et al.[48] and so on. However, through the literature study, we find that only little work is done on how to determine regularization parameters, and optimal diffusion operators for achieving optimal image restoration results. A Bayesian estimation based double variational regularization in the space of functions of Bounded Variation (BV) is proposed. The main idea is described in the following.

When an image f is discontinuous, the gradient of f has to be understood as a measure, and the space $BV(\Omega)$ of functions of bounded variation is well adapted for this purpose. The Osher-Rudin functional (TV) is strictly convex and is lower semicontinuous with respect to the weak-star topology of BV. Therefore, the minimum exists and is unique. The decomposition of the TV model heavily depends on the specific norm which is chosen on BV. However, the Osher-Rudin functional (TV functional) is a special example of a more general smoothing algorithm [164]. We relax the TV functional to a more general convex functional in the space $BV(\Omega)$ where$|Df| \rightarrow \phi(|Df|)$, and the formulation of the problem is

$$\inf_{f \in BV(\Omega)} \mathcal{J}(f_{(g,h)}) = \frac{1}{2} \int_{\Omega} (g - hf)^2 dA + \lambda \int_{\Omega} \phi(|Df(x,y)|) dA \tag{1.2}$$

This equation is studied by Vese et al. [249] for image deblurring and denoising. Furthermore, we study a more general variant exponent, linear-growth variational, convex functionals in the $BV(\Omega)$ space by Chen, Levine and Rao [48] and [49],

$$\inf_{f \in BV(\Omega)} \mathcal{J}(f_{(g,h)}) = \frac{1}{2} \int_{\Omega} (g - hf)^2 dA + \lambda \int_{\Omega} \phi(x, Df(x,y)) dA \tag{1.3}$$

where $\phi(|Df(x,y)|) \rightarrow \phi(x, Df(x,y))$. For the definition of a convex function of measures, we refer to the works of Goffman-Serrin [93] Demengel-Temam [59], and Aubert [15]. For $f \in BV(\Omega)$, we have,

$$\int_{\Omega} \phi(x, Df(x,y)) dA = \int_{\Omega} \phi(x, \nabla f(x,y)) dA + \int_{\Omega} |D^s f(x,y)| dA \tag{1.4}$$

The main importance and benefit of Eq. 1.3 is that we can study and inference a new variant exponent, linear growth functional in the BV space for image denoising [48].

Since the degradation of images includes not only random noises but also multiplicative, spatial degradations, i.e., blur, we extend this equation for simultaneous image deblurring and denoising. We construct a Bayesian estimation based double variational regularization with respect to the estimation of PSFs and images. The proposed functional in strictly convexity is shown in the following,

$$\inf_{f \in BV(\Omega)} \mathcal{J}_{\varepsilon}(\hat{f}, \hat{h}) = \tfrac{1}{2} \int_{\Omega} (g - \hat{h} * \hat{f})^2 dA + \lambda \int_{\Omega} \phi_{\varepsilon}(x, D\hat{f}) dA + \beta \int_{\Omega} (\nabla \hat{h}) dA \tag{1.5}$$

where $\hat{f}$ and $\hat{h}$ are the iteratively restored image and PSF. ε is a small constant to avoid the zero denominator during discrete numerical approximation. To achieve perceptual image restoration, we also use techniques from the theory of time-dependent minimal surfaces and the hyperbolic conservation laws for the numerical approximation. The proposed approach has several important effects: firstly, it shows a theoretically and experimentally sound way of how local diffusion operators are changed automatically in the BV space. Secondly, the self-adjusting regularization parameters also control the diffusion operators simultaneously for image restoration. Furthermore, the time of stopping the process is optimally determined by measuring the signal-to-noise ratio. Finally, this process is relatively simple and can be easily extended for other regularization or energy optimization approaches.

1.3.3 Variational Bayesian Learning and Discrete Regularization for Nonuniform Blurred Image Segmentation and Restoration

How to restore Nonuniform blurred (e.g., partially-blurred) images without influencing unblurred regions and objects? Furthermore, how to automatically identify and perceptually segment blurred regions or objects and restore them respectively? In the real world, CCD and CMOS images, tomograph data, remote sensing or medical data are often entirely-blurred or partially-blurred in a stationary or non-stationary way. Blind image restoration (BIR) of partially-blurred images is to restore blurred regions without influencing unblurred regions for achieving better visual perception based on the Gestalt theory [270].

However, we can not directly apply normal image restoration methods to restore partially-blurred images. The restoration of partially-blurred images generates an interesting question. From the mathematical viewpoint the question is, how to get a global convergence of multi-levels of local distributions. These multi-levels of local distributions include local pixel gray level distributions, randomly distributed local blurry regions and unblurred regions or objects. Therefore, it becomes a challenging partial convergence problem [248]. A novel mathematical model needs to be constructed for the solution. The main strategy is summarized in the following,

1. To motivate the algorithm, different characteristic properties [80], [153] (gradient, frequency, entropy, etc.) [67], [191] between blurred and unblurred regions or objects endowed with pairwise relationships can be naturally considered as a graph. We treat blind image restoration of partially-blurred images as a combinatorial optimization problem [130], [85], [80] based on regularization theory [241], and spectral clustering theory in discrete graph spaces [51], [80], [226], [282] and call it discrete regularization [300]. Some connections between some of these interpretations are also observed in [300], [282], [142], [51] based on differential geometry and transductive inferences [248], [222], [300]. More important, this integration brings crucial insights to the understanding of these theories, underlying relationships and their potential roles.

2. The main objective of the standard regularization techniques is to obtain a reasonable reconstruction which is resistant to noise in inverse problems. Based on these inherent characteristic properties, discrete regularization is about converting high-level targets (e.g., identification and segmentation of blurred and unblurred regions or objects), guiding low-level image processing (e.g., similarity measure) and learning the optimal segmentation based on multi-levels of local distributions (e.g., blur and unblurred regions, different distributions of color, texture, and gray values). Conceptually, the discrete regularization paradigm also reveals the roles of some well-known optimization algorithms. Algorithms such as graph-cuts [132], variational regularization [184], [173], [267] can be viewed as either discrete regularization [24] with energy in binary discrete spaces or in continuous bounded variation spaces. Compared to Markov random fields based stochastic optimization approaches [85], [80], this paradigm in the discrete graph space is optimized in a deterministic way.

3. Through large observations and experiments, we classify natural blurred images into three main blurred groups so that we can design an efficient methods. Natural image statistics [227], [209], [73], [109] has some properties to represent images. As a result, we obtain an approach, which can compute and use the translation and scale-invariant marginal probability distribution of image gradients as *a priori* through the Bayesian learning scheme.

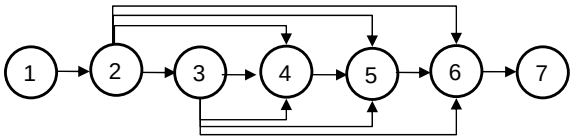


Figure 1.2: Diagram of chapters

In a sense, the distribution can be shared by most similar type of blurred images and therefore requires relatively few training images. Moreover, we approximate the Bayesian ensemble learning [114], [27], [166], [167], [73] into a variational manner in graphical models [121], and closely related with mean field theory [181], variational free energy [183]. The variational methods make the Bayesian ensemble learning more tractable, practical and efficient. Finally, inspired by the multi-scale [197], [73] and multigrid methods [37], the blur kernel is identified and interpolated from low-resolution to high resolution. Therefore, we can avoid local minima and achieve high accuracy blur kernels. Experiments show that the suggested method is more robust and can restore large nonuniform blurred images.

1.4 Organization and Contributions

Before our discussion of main contribution can begin in earnest, certain fundamental concepts and results must be introduced. Chapter 2 is devoted to that task. The chapter consists of material that is standard, elementary functional analysis, background and is essential for further pursuit of our objectives.

Chapter 3 and 4 are devoted to the statistical model selection and regularization for blur identification. Chapter 3 develops the general statistical model selection methods, illustrating it with some applications for nonparametric blur identification and blurred image selection from individual images and large video sequences. Chapter 4 deals solely with the suggested method of Bayesian estimation based double regularization to parametric blur identification including global nonparametric blur identification and local parametric optimization of blur kernels. During the iterative double regularization, the estimated PSF is prior knowledge for the next iterative estimation of the image, and vice visa.

Chapter 5 is devoted to improve the fidelity and quality of restored images including deblurring and denoising. It is in this chapter that we introduce several general linear growth functional for image processing in the space of functions of bounded variation. The concepts of the bounded variation space and variational regularization are pursued in this chapter through the introduction of a linear growth variation functional, a variant exponent L^p linear growth functional and our suggested Bayesian estimation based double variational regularization functional.

Chapter 6 considers a more general blur problem in the real world, e.g., partially-blurred images including stationary and nonstationary blur kernels. There are a lot of existing segmentation approaches which cannot achieve satisfactory results on the identification and partition of blurred regions or objects. Inspired by spectral clustering image segmentation concepts and the underlying mathematic connections with regularization theory, we investigate the convex regularization in discrete graph space and have good performances. The restoration of non-stationary blurred images are solved using a proposed variational Bayesian ensemble learning approach with natural image statistics prior. The results outperform most state-of-art methods. The techniques

in this chapter is an outgrowth of the principles of variational Bayesian learning, regularization and convex optimization based graph theoretic concepts. In the course of the development and compare to existing methods, the perceptual image restoration problem is firstly treated into two simultaneous problems, i.e., perceptual image segmentation and image restoration based gestalt law.

Finally, Chapter 7 contains a summary of the suggested techniques for the solution of this ill-posed inverse problems in the field of statistical learning, pattern recognition and computer vision. Some publications are explained with respect to previous chapters, but many new ideas and plans are proposed for the future work. The structure of the whole thesis is presented in Fig. 1.2.

The main contributions of this thesis are summarized in the following.

1. **Bayesian estimation based global nonparametric model selection and local parametric optimization for blur identification.** Global nonparametric estimation and local parametric optimization is an ongoing research topic. Through the case of blur identification, we study this statistic strategy in an alternative way. We investigate the systematic design of convex and non-convex regularization by integrating statistical learning and a variety of regularization models. This approach combines global nonparametric estimation techniques and local parametric optimization techniques for improving the accuracy of blur identification. In the context, we also introduce several new ideas that improve the quality of blur identification with respect to noise, mixed blur and noise, in individual images or large video data. Moreover, we present a systematic framework for blur identification methods based on the integration of statistical learning and convex regularization. This system proves to be useful in several respects: Firstly, statistical learning provides accurate initial values for the iterative optimization approach which largely improves the results. Secondly, the prior learning terms in the regularization can be considered as a convex penalty term for keeping the energy functional in a strictly convex functional.

2. **Adaptive data-driven variational image denoising and deblurring in the BV space.** A novel method is proposed for determining the optimal parameters and operators to achieve optimal high-fidelity image restoration. The selection of regularization parameters is self-adjustable following the spatially local variance. Simultaneously, the linear and non-linear smoothing operators are continuously changed following the strength of discontinuities. The time of stopping the process is optimally determined based on the improvement of signal-to-noise ratio. The numeric implementation of these algorithms are based on the hyperbolic conservation laws which can largely improve the visual perception results. These criteria are used to adjust regularization parameters for balancing the global energy minimization to achieve perceptually high-fidelity image restoration.

3. **Unified variational Bayesian learning and regularized spectral graph clustering on discrete graph spaces for nonuniform blurred (e.g., partially-blurred) image identification, segmentation and restoration.** Different from the traditional regularization approaches in continuous spaces (e.g., Hilbert space, BV space), we have designed a discrete regularization approach based on the integration of spectral graph theory and regularization theory. This approach unifies spectral clustering and spectral eigenvalues analysis in a regularized spectral graph approach. Moreover, we extend a family of discrete regularization operators in Riemannian manifold for the smoothness of optimization.

Consequently, the restoration of identified and segmented regions and objects are solved in a variational Bayesian ensemble learning framework. Natural image statistics can be scale-invariant prior through the Bayesian learning. High-quality perceptual image segmentation and restoration can be achieved for such nonuniform and nonstationary blurred images.

This thesis presents a unified solution for solving some of the most challenging problems in image processing, pattern recognition and computer vision. It also present a state-of-the-art architecture for the integration of statistical learning and functional optimization. All the approaches are formulated in a well-defined sense from one underlying mathematic principle, namely Bayesian learning theory and regularization theory. Thereby, these approaches are refined from the well understood principles of PDEs in the continuous Hilbert space, BV space and discrete graph spaces in Bayesian learning frameworks. These approaches can be easily extended to other related pattern recognition, and computer vision tasks.

2 Regularization for Image Deblurring and Denoising

Mathematic optimization: "What is new?" is an interesting and broadening eternal question, but one which, if pursued exclusively, results only in an endless parade of trivia and fashion, the silt of tomorrow. I would like, instead to be concerned with the question. "What is best?", a question which cuts deeply rather than broadly, a question whose answers tend to move the silt downstream. –"Zen and the Art of Motorcycle Maintenance", Robert M. Pirsig, (1974)

2.1 Image and Blur Modeling

2.1.1 A Mathematical Model for Image Formation

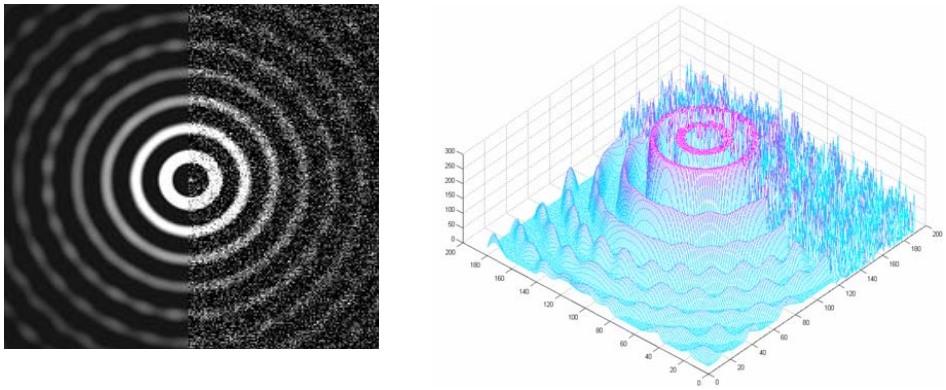
Let us denote $f(x', y')$ the irradiance function of the object under observation in the object plane with coordinates x' and y', and $g(x, y)$ denotes the observed irradiance function in the image plane with coordinates x and y. A very general image degradation model is

$$g(x,y) = \Phi\left\{\iint h(x,x',y,y') f\left(x',y'\right) dx'dy'\right\} \odot \eta(x,y) \tag{2.1}$$

where $\Phi\{\cdot\}$ represents a nonlinear function, $h(x, x', y, y')$ is the response of the blurring system to a two-dimensional impulse at the (x, y) spatial location, and is normally called point spread function, $\eta(x, y)$ denotes the corruptive noise process and is usually random and highly oscillated [164], and $\odot$ represents a point-wise operation (additive or multiplicative). The function $\Phi\{\cdot\}$ usually defines a pointwise (memoryless) operation which is used to model the response of the image sensor. For example, $\Phi\{\cdot\}$ can be the *Hurter-Drifield* curve [243] used for describing the recording medium in traditional photographic films. Andrew and Hunt [12] and others have proposed restoration techniques with the nonlinearity taken into account. However, a general conclusion reached by previous researchers. That is, there is no significant improvement of the restoration results by taking the nonlinearity into account. Therefore, in most of the work of image restoration, $\Phi\{\cdot\}$ is ignored.

While photoelectronic systems (e.g., CCD, CMOS) sense, acquire, and process the signal from the detector's photoelectronic surface for image recording, the noise stems from the random fluctuations in the number of photons and photoelectrons on the photoactive surface of the detector and the random thermal noise sources in the circuits. A stochastic model for the data distribution D_{ij} recorded by the ijth pixel of a CCD array is given by

$$D_{ij} \propto \text{Poisson}(\eta_{p(ij)}) + \text{Normal}(0, \sigma^2) \tag{2.2}$$



(a) (b)

Figure 2.1: (a) Original image and half side of additive Gaussian noise. (b) Related surface.

The distribution is a combination of Poisson noise $\eta_{p(ij)}$ and Gaussian noise with variance σ^2. The Poisson distribution models the photon count, while the additive Gaussian term accounts for background noise in the recording electronics. Although the first process generates signal-dependent noise, both photoelectronic and thermal noise are usually modeled by a zero mean additive white Gaussian (AWG) process, shown in Fig. 2.1. Therefore, due to these two basic simplifications, the degradation takes the form

$$g(x,y) = S\left\{\iint h(x,x',y,y')f\left(x',y'\right)dx'dy'\right\} + \eta(x,y) \tag{2.3}$$

However, this degradation model has found limited use, due primarily to high computation requirement of four variables in the blur kernel $h(x,x',y,y')$. In most practical situations, the blur can be modeled in a linear space-invariant (LSI)way with two variables $h(x,x',y,y') = h(x-x',y-y')$. Thus, the yielding degradation model applies:

$$g(x,y) = \iint h(x-x',y-y')f\left(x',y'\right)dx'dy' + \eta(x,y) \tag{2.4}$$

The equation is called the superposition or Fredholm integral of the first kind. This expression is of fundamental importance in linear system theory. The solution of image restoration becomes available using the extensive tools of linear system theory. As we know, linear inverse problems frequently lead to integral equations of the first kind, which is the reason of such equations playing an important role in the study of inverse problems. On the other hand, many basic inverse problems are inherently nonlinear even if the corresponding direct problem is linear. Along the lines of solving the inverse problem, nonlinear methods play an important role for nonlinear inverse problems. The theory of regularization methods [241] is well-developed for linear inverse problems and at least emerging for nonlinear problems [68].

Based on this image formation equation in Eq. 2.4, we assume that both images are square of size $N \times N$, that is, $0 \leq x \leq (N-1)$ and $0 \leq y \leq (N-1)$. By stacking or lexicographically ordering the $N \times N$ arrays of $f(x,y)$, $g(x,y)$ and $\eta(x,y)$. The previous equation becomes

$$g = Hf + \eta \tag{2.5}$$

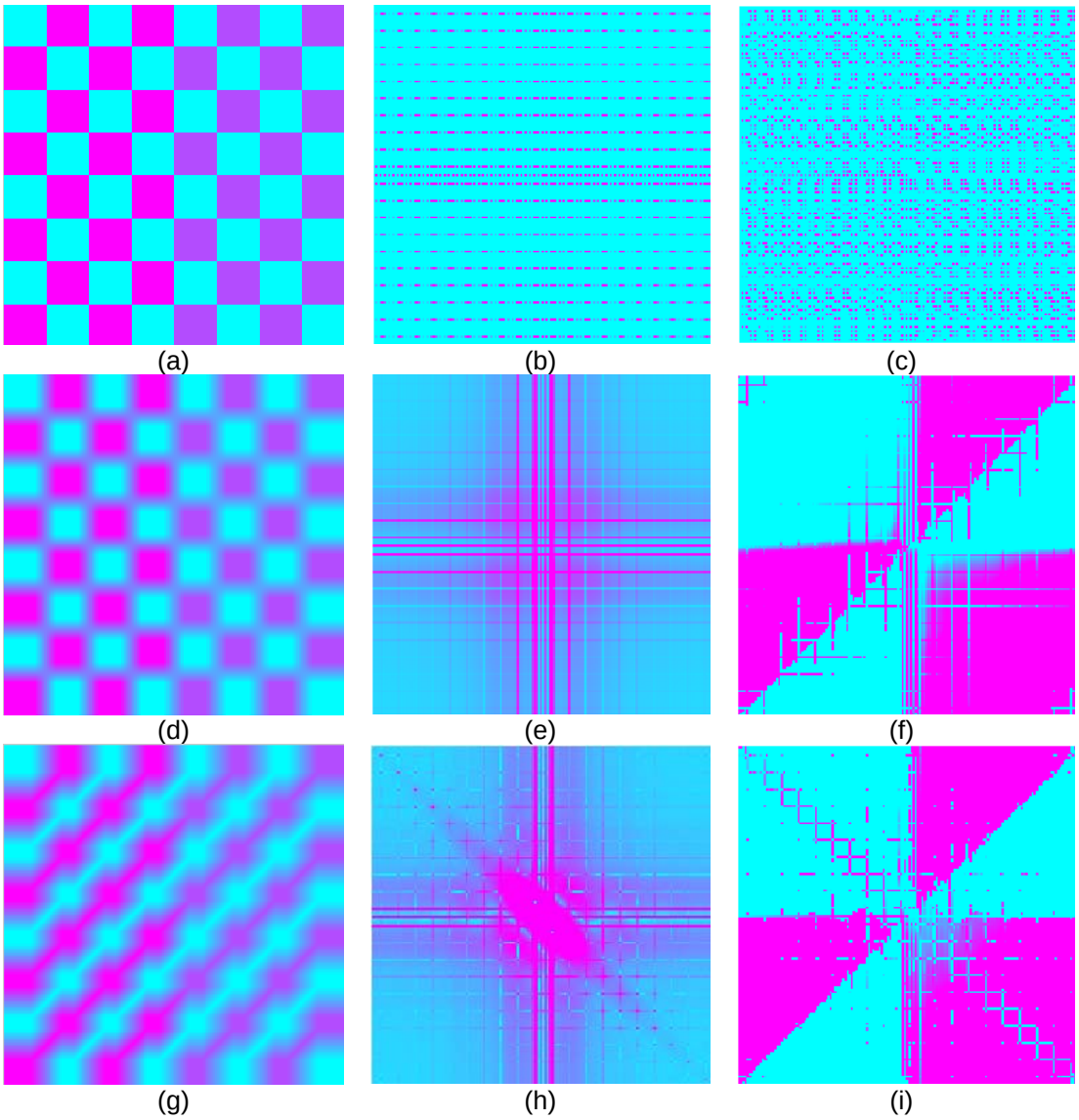


Figure 2.2: Different blurred images with different FFT magnitude and phase. (a)(d)(g) Original synthetic image, Gaussian blurred image and motion blurred image. (b)(e)(h) 2D FFT log magnitude spectrum. (c)(f)(i) 2D FFT phase.

where g, f and n are $N^2 \times 1$ vectors (color images are $N^2 \times 3$). The PSF H is $N^2 \times N^2$ matrix. For the space invariant blur case, H is a block Toeplitz matrix. Such matrix can be approximated by block circulant matrices. Block circulant matrices are easily diagonalized since their eigenvalues are the 2D discrete Fourier transform (DFT) values of the defining 2D sequences, and their eigenvectors are defined in terms of Fourier kernels. Thus, the equation can also be written in the discrete frequency domain.

$$G(k,l) = H(k,l)F(k,l) + \eta'(k,l) \tag{2.6}$$

where $G(k,l)$, $H(k,l)$, $F(k,l)$ and $\eta'(k,l)$ represent the 2D DFTs, for $0 \leq k \leq (N-1)$ and $0 \leq l \leq (N-1)$. The $H(k,l)$ are the "unstacked" eigenvalues of the matrix H, as was already mentioned. We can arrive at Eq. 2.6 by taking the 2D DFT of both sides of Eq. 2.4, under

the assumption that it represents circulant convolution, shown in Fig. 2.2. The 2D arrays can always be appropriately padded with zeros (for example), so that the result of circular convolution equals that of linear convolution. It is noted that although the degradation model is linear space invariant (LSI), the restoration filter may be nonlinear or space-variant or both.

The discrete convolution product defines a linear operator. A discrete analogue of the continuous Fourier transform can be used to efficiently compute regularized solutions. In the practical environments some representative restoration algorithms with signal dependent and multiplicative noise are mostly simulated for MRI, and CT images, and additive Gaussian noise is simulated for CCD, CMOS images, mixed noise of Poisson and Gaussian noise [85]. In the field of signal processing, artificial intelligence, and pattern recognition, the independent component analysis (ICA) has the similar basic formula and mechanism as the induced image formation form in Eq. 2.6. In this thesis, we focus on this induced image formation form in Eq. 2.6 for blind image restoration and segmentation problems.

2.1.2 Nonparametric and Parametric Image Models

Blind image restoration includes two parts such as blur identification and image restoration (deblurring and denoising). To restore images, we need firstly to specify one image model for applying algorithms. The basic distinction between image models is that between deterministic and stochastic models. According to Andrews and Hunts [12] and Katsaggelos [127], [19], deterministic image models can be divided into parametric (an image is represented in terms of primitives) or nonparametric. With a stochastic model an image is considered to be a sample function of an array of random variables called a random field. Stochastic models can also be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along; for example, Markov random field, also known as undirected graphical models, in which the links do not carry arrows and have no directional significance. The undirected graph is suited to expressing soft constraints between random variables, whereas directed graphs are useful for expressing causal relationships between random variables.

A possible division of stochastic models is parametric and nonparametric. A parametric stochastic image model is to assume that the image field is described by a 2D Gaussian probability density function (PDF) with two parameters, the covariance matrix and the mean vector. Maximum likelihood estimation is commonly used for such an image model. If the mean and covariance value are not suitable as parameters for a PDF (unknown PDF) and still used in modeling an image, the non-parametric stochastic model may become to use. Both models can be defined as covariance models [119] whatever the covariance is, a parameter of a PDF or not. Covariance models can be divided into stationary or homogeneous and nonstationary or inhomogeneous [127]. The detailed description are presented in the following,

1. A *stationary* model is defined as one having a constant or stationary mean and a stationary covariance. In most cases, the covariance matrix can be approximated in block Toeplitz in image restoration. Most existing discrete approximation models are extended from this model for solving and simulating related boundary conditions for the stationary model in image restoration.

2. *Nonstationary* image models are classified into three types of model by assuming the nonstationary mean and/or nonstationary covariance. The first model is a Gaussian image

model with nonstationary or space-variant mean and stationary covariance which was proposed by Hunt and Cannon [12]. The second model is that the local mean and the local variance are used for image estimation and restoration. The third model is a generalized model including two inhomogeneous image models, e.g., partially-blurred images. Practically, a freely taken image is an inhomogeneous random field so that the homogeneous model can not fit it well. The sampling and mean techniques then become more important for blur identification.

2.2 Convex Regularization

Regularization is the approximation of an ill-posed problem by a family of neighboring well-posed problems. Convex regularization is keep the regularization function in convexity, especially strictly convexity.

2.2.1 Ill-Posed Inverse Problems and Regularization Approaches

Ill-posed Inverse Problems

The concept of a well-posed problem was introduced by J. Hadamard (1923), in an attempt to clarify what types of boundary conditions are most natural for various types for differential equations. As a result of his investigation, a problem characterized by the equation $Ax = y$, where $x \in \mathcal{H}_1$, $y \in \mathcal{H}_2$ (both $\mathcal{H}_1$ and $\mathcal{H}_2$ denote Hilbert spaces) and A is a bounded linear operator, is defined to well-posed provided the following condition are satisfied:

1. for every element $y \in \mathcal{H}_2$ there exists a solution in the space $\mathcal{H}_1$;
2. the solution is unique;
3. the problem is stable on the space $(\mathcal{H}_1, \mathcal{H}_2)$, which means that the solution depends continuously on data.

Otherwise the problem is ill-posed. Later, the concept of well posedness in the least-squares sense has been introduced by Nashed [174]. It is according to which $Ax = y$ is well-posed if for each $y \in \mathcal{H}_2$ there exists a unique least-squares solution (of minimal norm) which depends continuously of the data. For years, ill-posed problems have been considered as mere mathematical anomalies. Indeed, it was believed that physical situations only lead to well-posed problems. However, this attitude was erroneous and many ill-posed problems arise in practical situations. A detailed list of the ill-posed problems arising in mathematical physics is provided in the monograph by Lavrentiev [138].

If the image formation process is modeled in a continuous infinite dimensional space, the distortion operator H becomes an integral operator and $g = Hf + \eta$ becomes a Fredholm integral equation of the first kind in Eq. 2.4. Then the solution is always an ill-posed problem. This means that the unique least-squares solution of minimal norm of $g = Hf + \eta$ does not depend continuously on the data or a bounded perturbation (noise) occurs in the data. It results in an unbounded perturbation in the solution. This solution of the generalized inverse of blur kernel H could be unbounded [174], [127]. The integral operator H has a countably infinite number of

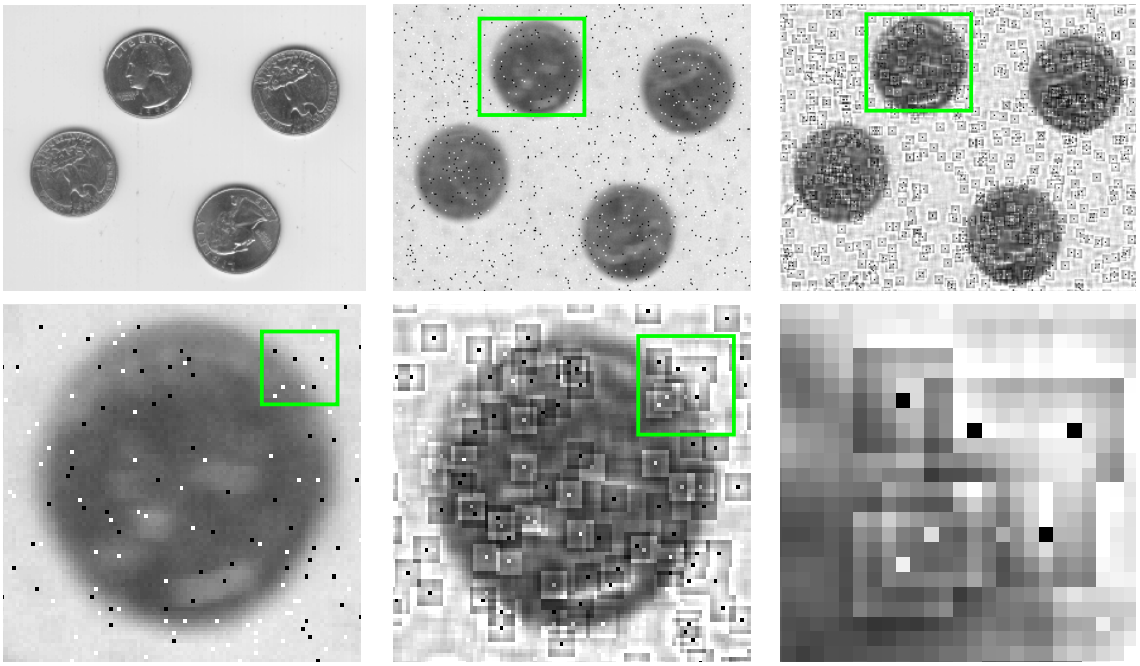


Figure 2.3: $\frac{a|b|c}{d|e|f}$ Noise is amplified during the deconvolution. (a) Original image. (b) Blurred image with salt-pepper noise (impulsive noise). (c) Deconvolved image using Richard-Lucy filter. (d)(e)(f) Zoom in images

singular values. Since the finite dimensional discrete problem of image restoration results from the discretization of an ill-posed continuous problem, the matrix H has a cluster of small singular values. Clearly, the finer the discretization (the larger the size of matrix H) the closer the limit of the singular values is approximated. Therefore, although the finite dimensional inverse problem is well-posed in the least-squares sense, the ill-posedness of the continuous problem translates into an ill-conditioned matrix H. The detailed proof we refer [138].

In quantifying the conditioning of a matrix the condition number of $\mathcal{N}(H)$ can be used, defined according to the inequality [271]

$$\frac{\|e\|}{\|f\|} \leq \|H\|\|H^{inv}\|\frac{\eta}{Hf} = \mathcal{N}(H)\frac{\eta}{\|Hf\|} \tag{2.7}$$

where H^{inv} denotes the generalized inverse of H, f is the solution of ideal noiseless image, and e denotes the error in the solution when the noisy input image g is available. If the value of $\mathcal{N}(H)$ is small, a small relative change in g cannot produce a very large relative change in f. If $\mathcal{N}(H)$ has a large value, a small perturbation in the image may result in large (although bounded) perturbation in the solution, and the system is said to be ill-conditioned. By using the L^2 norm for vectors and matrices, $\mathcal{N}(H)$ takes the simplified form,

$$\mathcal{N}(H) = \|H\|_2 \cdot \|H^{inv}\| = \frac{\mu_1}{\mu_r} \tag{2.8}$$

where $\mu_1, ..., \mu_n$ are the singular values of H, r is the rank of H, and it was assumed that $\mu_1 \geq \mu_2 \geq ... \geq \mu_r \geq \mu_{r+1} = ... = \mu_n = 0$. Since the largest singular value of H is different from

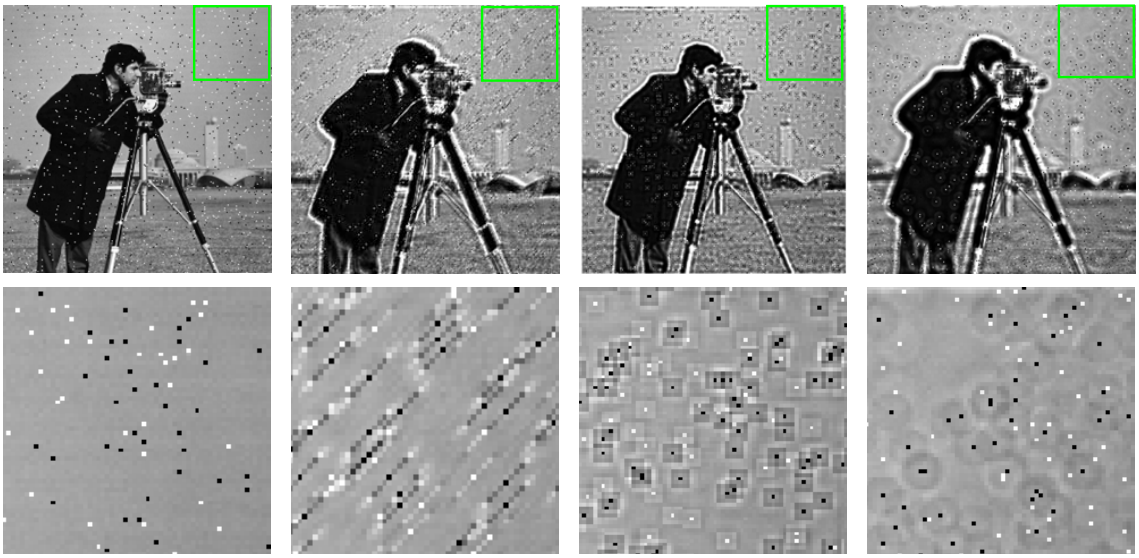


Figure 2.4: $\frac{\text{a|b|c|d}}{\text{e|f|g|h}}$ Noise is amplified with different blur deconvolution using Richard-Lucy filter. (a)(e) Salt-pepper noise. (b)(f) Motion blur deconvolution. (c)(g) Gaussian blur deconvolution. (d)(h) Pill-box blur deconvolution.

zero due to the assumption of lossless imaging, $\mathcal{N}(H)$ is an increasing function of the image dimensions.

The problem of noise amplification can be further explained by using a spectral approach. That is, the minimum norm least-squares solution of $g = Hf + \eta$ can be written as

$$\hat{f} = \sum_{i=1}^{r} \frac{(v_i, Hf)}{w_i} + \sum_{i=1}^{r} \frac{(v_i, \eta)}{w_i} v_j \tag{2.9}$$

where v_i and v_j are respectively the eigenvectors of $HH^\top$ and $H^\top H$, and (v_i, v_j) denotes the inner product of the vectors v_i and v_j. Clearly, since H is an ill-conditioned matrix some of its singular values will be very close to zero, so that some of the weights w_i^{-1} are very large numbers. If the ith inner product (v_i, η) is not zero (as is true when it is broadband), the noise of the second term is amplified. Similar observations can be made by using the spectral decomposition of an operator in infinite dimensional spaces. If matrix H is block circulant, the singular values w_i are equal to $|H(x, y)|$ in Eq. 2.6, where the $|\cdot|$ denotes complex magnitude. Different deconvolution methods have different amplification of noise. For example, inverse filter and Wiener filter are very sensitive to noise. Richard-Lucy methods is relatively robust for noise but the noise can be still amplified. In Fig. 2.3, impulsive salt-pepper noise distributes randomly in individual pixels. It is strongly amplified during the deconvolution. Fig. 2.4 shows the deconvolution using different blur kernels. Therefore, denoising is also very important in image restoration.

Regularization Approaches

"Regularization of ill-posed problems" is a phrase used for various approaches to circumvent lack of continuous dependence. Roughly speaking, a regularization method entails an analysis of an ill-posed problem via an analysis of an associated well-posed problem, whose solution yields

meaningful answers and approximations to the given ill-posed problem. According to A. N. Tikhonov [241], the regularization method consists of finding regularizing operators that operate on the data, and determining the regularization parameters from supplementary information pertaining to the problem. The regularization operator depends continuously on the data and results in the true solution when the regularization parameters go to zero, or equivalently when the noise goes to zero. On the other hand, in the 1970s, Vapnik et al. [248] generalized the theory of the regularization method for solving the so-called stochastic ill-posed problems. They define stochastic ill-posed problems as problems of solving operator equations in the case when approximations of the function on the right-hand side converge in probability to an unknown function and / or when the approximations to the operator converge in probability to an unknown operator. In particular, the regularization methods have been extended for solving the learning problems: estimating densities, conditional densities, and kernel based classifiers.

Numerous methods have been proposed for treating and regularizing various types of ill-posed problems. The various approaches to regularization involve essentially one or more of the following intuitive ideas in different research streams,

1. change of the concept of a solution [126], [219], ;
2. additional information for the restriction to a compact set [105];
3. projection for the change of the space and/or topologies [126];
4. shift the spectrum for the modification of the operator[251], [290], [288], [186], [221];
5. well-posed stochastic extension, convergence with respect to Lèvy-Prokhorov metric on the collection of probability measures on a given metric space: Banks-Bihari, Engl-Wakolbinger, Engl-Hofinger-Kindermann [68], [43].

The various approaches to regularization overlap in many aspects, especially in theoretical progresses and practically possible solutions for ill-posed inverse problems. Most existing image restoration methods have a common estimation structure in spite of their apparent variety. The common structure is expressed by regularization theory. Such a statement can be also made for most early vision approaches [25], [194], kernel based regularization approaches [222], multilayer network learning approaches [193], [70] and so on. In most of these approaches, the underlying idea of regularization is to combine the prior information with the data information and defines a solution by trying to achieve smoothness and yet remain fidelity to the data. In other words, a regularized solution is a solution between the "ultra-rough" least-squares solution and an "ultra-smooth" solution based on *a priori* knowledge.

From optimization point of view, the solution of regularization is to put the objective or cost (energy) functions into an optimization problem which makes the best possible choice of objective functions from a set of candidate choices. The objective or cost function might be a measure of the overall risk or variance. For example, two cases are presented,

1. In the case of image restoration, the solution of optimization corresponds to a choice that has minimum cost among all choices that meet the firm requirements, i.e., input a degraded image and output an restored image with high-fidelity to the original image.

2. In the case of blur identification, the task is to find a model, from a family of potential models, that best fits some observed data and prior information. Here the variables are the descriptive parameters in the model, the constraints or knowledge of prior information or limits on the parameters (such as nonnegativity of images in the physical world).

The objective function can be a measure of misfit or prediction error between the observed data and the values predicted by the model, or a statistical measure of the unlikeliness or implausibility of the parameter values. Thereby, the optimization problem is to find the model parameter values that are consistent with the prior information, and give the smallest misfit or prediction error with the observed data (or, in a statistic framework).

From these analysis, restoring the image f can be seen as a minimization problem [16]. The general minimization model incorporates the strengths of the various types of diffusion arising from,

$$\mathcal{J}(f) = \frac{1}{2}\int_{\Omega} |g - hf|^2 dxdy + \lambda S(f) \tag{2.10}$$

where $S(f) = \int_{\Omega} |\nabla f|^p dxdy$, for $1 \leq p \leq 2$, $\lambda \geq 0$, and Ω is an open bounded subset of $\mathbb{R}^n$ (we consider n =2, or 3 dimensions), Here it denotes the support of image. The first term in Eq. (2.10) measures the fidelity to the data. The second term is a penalty smoothing term. The positive regularization parameter λ controls the trade-off between the fidelity to the observation and smoothness of the restored image. When a magnitude of a gradient is $p = 2$, the Eq. (2.10) becomes a L^2 norm Tikhonov solution [165], [241]. The L^2 norm regularization has very strong isotropic (Laplace) smoothing properties but penalizes strongly the gradients corresponding to the discontinuities and edges. In order to handle discontinuities, the issue of non-directional versus directional operator has been debated firstly by Marr and Hildreth [157], [158]. Later, some of the pioneer work in this direction was done by Rudin, Osher, and Fatemi [213], [212], who proposed to use the L^1 ($p = 1$) norm of the gradient of f in Eq.(2.10) and called the total variational (TV) regularization. The TV method with L^1 norm encourages smoothing in the direction tangential to the edges and weakly penalize in the direction orthogonal to the edges in the space of a bounded total variation [17], [44], [258], [260].

To preserve the textures, edges and small scale details, more elegant constraints are proposed and explored by researchers like the forward-backward diffusion sharpen operator and some more detailed optimization [245]. Perona and Malik [191] replaced the classical isotropic diffusion ($p = 2$) with the values of $1 < p < 2$ in general nonlinear diffusion which is effective in reconstructing piecewise smooth regions between the isotropic, anisotropic nonlinear and TV-based smoothing [259], [266], [29], [173]. To further improve the fidelity of image restoration, different integration models of L^1 and L^2 norms have been explored by Chambolle, Chan as well as discontinuity-preserving and fidelity enhancement by [39], [48], [205]. Recently, Yves Meyer (2001)[164] presented an mathematical analysis of the Rudin-Osher-Fatemi model (1992) [213] in the bounded variation (BV) space of functions. The Fourier vs. wavelet series is expansions of BV functions. He also introduced a new space which is called G space to model oscillating patterns and widely used for image structure, texture and homogenous layer decomposition.

Information theory have also been extended to regularization theory. For example, maximal entropy regularization can use an entropy measure term instead of normal L^p term, e.g., $S(f) = \int_{\Omega} f \ln(f/m)$, where m is some positive function reflecting a priori information about f.

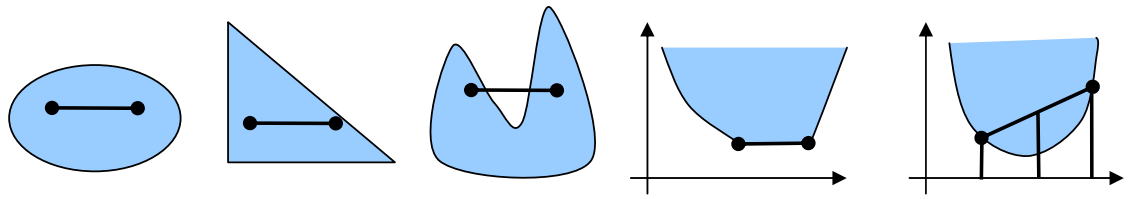


Figure 2.5: *a|b|c|d|e.* Convex sets and convex functions. (a)(b) Convex sets. (c) Nonconvex set. (d) Convex function. (e) Strictly convex function.

Integration and combination of information, statistical learning and variational regularization have been also investigated by researchers and still be an interesting research point.

2.2.2 Convex Optimization

Convex functionals (shown in Fig. 2.5) play a special role in the theory of optimization because most of the theory of local extrema for general nonlinear functionals can be strengthened to become global when applied to convex functionals. Conversely, results derived for minimization of convex functionals often have analogs as local properties for more general problems. The study of convex functionals leads then not only to an aspect of optimization important in its own right but also to increased insight for a large portion of optimization theory. The principle idea of regularization based approaches [16], [213], [29], [173], [221] is based on the optimization process so that the behavior can be analyzed by the convexity and non-convexity. Furthermore, the consistency of local and global convergence can be preserved based on the convexity.

Nonlinearity does not mean that a problem is difficult, but non-convexity does in general. Even though an ill-posed problem is non-convex, the sound approaches are still relying on convex optimization approaches as basic components. For example, to recover the image f given an observed blurred noisy image g by minimizing an energy function is an ill-posed inverse problem. Convexity is crucial for ensuring an existing, unique and stable convergent solution for such tasks. Therefore, the study of criteria of convexity and convex functional is important for understanding and designing a strictly convex functional.

Local Minima and Global Minima

Perhaps the first question that arises in the minimization problem is whether a solution exists, and then comes to the solution of uniqueness and stability. To study these concepts, we distinguish two kind of solution points: local minimum points, and global minimum points.

Definition 2.2.2.1 *A point $x^* \in \Omega$ is a relative minimum point or a local minimum point of a function f over Ω if there exist an $\varepsilon > 0$ such that $f(x^*) \leq f(x)$ for all $x \in \Omega$ within a distance ε of x^* (that is, $x \in \Omega$ and $|x - x^*| < \varepsilon$). On the graph curve of a function, its local minima will look like the bottoms of valleys.*

Definition 2.2.2.2 *A point $x^* \in \Omega$ is said to be a global minimum point of f over Ω if $f(x^*) \leq f(x)$ for all $x \in \Omega$. If $f(x^*) < f(x)$ for all $x \in \Omega$, $x \neq x^*$, then x^* is said to be a strict global minimum point of f over the whole set Ω. Any global minimum is also a local minimum; however, a local minimum need not also be a global minimum.*

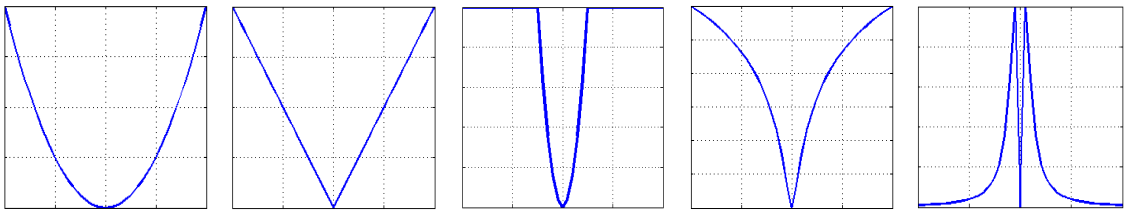


Figure 2.6: $a|b|c|d|e$. Function curves. (a). Tikhonov. (b) Total variation. (c) Huber function. (d) Log-quadratic. (e) Saturated-quadratic

Table 2.1: Convex and Nonconvex Functions

Function	Formula	Convexity
Quadratic function:	$\phi_1(x_i, x_j) = (x_i, x_j)^2$	convex.
Total variation:	$\phi_2(x_i, x_j) = \|(x_i, x_j)\|$	convex.
Huber function:	$\phi_3(x_i, x_j) = \begin{cases} \frac{1}{2\delta}(x_i - x_j)^2, & \text{if } \|x_i - x_j\| \le \delta \\ \|x_i - x_j\| - \frac{\delta}{2}, & \text{otherwise} \end{cases}$	mixed.
Log-quadratic:	$\phi_4(x_i, x_j) = \ln[1 + \frac{(x_i - x_j)^2}{\delta^2}]$	nonconvex.
Saturated-quadratic:	$\phi_5(x_i, x_j) = \frac{(x_i - x_j)^2}{\delta^2 + (x_i - x_j)^2}$	nonconvex.

The main result that can be used to address this issue is the theorem of Weierstrass, which states that if f is continuous and Ω is compact, a solution exists. This is a valuable result that should be kept in mind throughout our development. In the practical reality, searching for the minimum point by a convergent stepwise procedure based on differential calculus, comparison of the values of nearby points is all that is possible and attention using relative minimum points. Global conditions and global solution can only be found if the problem possesses certain convexity properties that essentially guarantee that any relative minimum is a global minimum. Thus, in formulating and attacking the problem $\arg\min f(x)$, subject to $x \in \Omega$ is usually considered as a searching for the relative minimum point.

Convex Sets and Convex Functions

A convex set is the set of basic solutions for convex programming [206], shown in Fig. 2.5. It means that x_1 and x_2 are feasible solutions, their linear combinations is $\lambda x_1 + (1 - \lambda)x_2$, $\forall \lambda \in [0, 1]$, must be feasible solutions. For convex programming to be applicable the cost must be a strictly convex functional over the convex set of feasible solutions. A functional $F : X \to [-\infty, \infty]$ is strictly convex if, for any two feasible solutions x_1 and x_2 such that $F(x_1) < \infty$, and $F(x_1) < \infty$, the inequality

$$F((1 - \lambda)x_1 + \lambda x_2) < (1 - \lambda)F(x_1) + \lambda F(x_2), \qquad \forall \lambda \in (0, 1) \tag{2.11}$$

always holds. This definition of a convex functional requires that Eq. (2.11) be valid over the set of feasible solutions. The result is also known as Jensen's *inequality* and it can be applied to information theory and machine learning. The generic convex optimization problem is to minimize the convex function $F(x)$ over a convex set. Convexity is a sufficient condition for all local minima to be global minima. There are three main properties about the convex

optimization, e.g., a convex function is continuous, a convex function has a single minimum on a convex domain, and the sum of convex functions is convex.

Following this definition, we study several functions that are commonly used as penalty terms in regularization approaches for image restoration [36], [22], [219], [267], [176]. The potential functions, $\phi(\cdot)$ are described in Table. 2.1. These five functions are representatives of three major categories, strictly convex ϕ_1 and ϕ_2, hybrid convex ϕ_3 and nonconvex ϕ_4 and ϕ_5, shown in Fig. 2.6. The convex quadratic function ϕ_1 in the regularization penalizes the differences of neighboring pixels at an increasing rate, which tends to force the image to be smooth everywhere. The total variation function ϕ_2 [213] is a L^1-norm cost function, which behaves in an absolute error in convex manner. Many convex functions have also been proposed recently such as robust anisotropic diffusion [30], [29], half-quadratic [84], linear programming [47], second-order cone programming [110] and low-dimensional flat Euclidean embedding in semi-definite programming [268].

The Huber function [115] is a semi-convex hybrid between quadratic and L^1 functions. It is quadratic for small values and becomes linear for larger values. Thus, it has the outlier stability of L^1. Therefore, the priors do not differentiate substantially between slow monotonic changes and abrupt changes. As a consequence, it does not penalize the presence of edges or boundaries in the image.

Non-convex functions have saturating properties that actually decrease the rate of penalty applied to intensity differences beyond a threshold. Consequently, the positivity of the presence of edges can be preserved in the image restoration. However, the non-convex functions present difficulties in computing global estimates. Non-convex optimization algorithms can also achieve good results with some constraints or in some special processing discipline. Recently, some non-convex optimization algorithms have also been developed. For example, binary spectral graph clustering and semidefinite relaxation [130] have been investigated for perceptual grouping and segmentation.

Multiple Model Criteria for Global Convergence

One of the key questions of image restoration and segmentation is how to optimize the proposed cost or energy function in global convergence. According to "http://www-fp.mcs.anl.gov/otc/Guide/OptWeb/index.html", we describe an optimization tree shown in Fig. 2.7. This tree introduces the different subfields of optimization and includes outlines of the major algorithms in each area. Through the literature study, several model criteria of optimization can be summarized in the following aspects:

1. **Direct local minimization to global convergence.** A local minimum energy can be substituted for the global minimum $\mathcal{J}$ using a plausible initial guess. Such algorithms are simple to implement but are relatively sensitive to the pertinence of the initialization. For example, the original ICM algorithm [26] uses the maximum likelihood algorithm based on the well-posed assumption where the noise should be very weak.

2. **Stochastic simulated annealing to global convergence.** Optimization using simulated annealing (SA) is based on the distribution $p_t(x|y) = \exp[-\mathcal{J}(x)/t]$, where t denotes temperature. $t \to 0$ decreases toward zero for objects x different from the global minima $\hat{x}$. $p_t(x|y)$ is processed to construct a Markov chain which converges to the set of the global

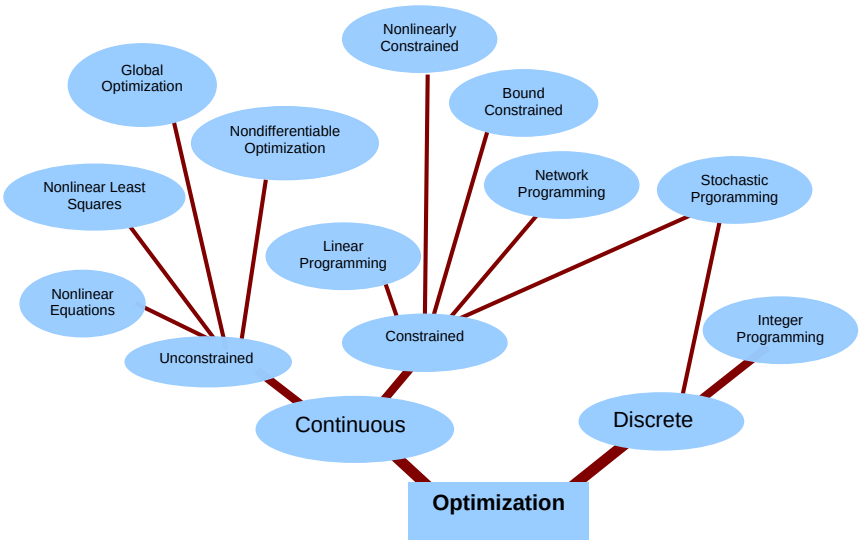


Figure 2.7: Optimization tree. Three main optimization criteria can be considered in this optimization tree, e.g., continuous versus discrete, global versus local, and convex versus non-convex.

minima of $\mathcal{J}$. In this process, the temperature decreases slowly from an initial high temperature toward zero. The Markov chain can be constructed based on stochastic gradient maximization of $p_t(x|y)$ [82], [273], metropolis dynamical sampling of $p_t(x|y)$ [83], [273], Gibbs dynamical sampling of $p_t(x|y)$ [111], [85], [273]. This type of algorithms is widely used in image and signal processing.

3. **Deterministic relaxation to global convergence.** A class of approximate (relaxed) energies is constructed by reducing the nonconvexity of $\mathcal{J}$. Thus, the nonconvexity is "converted" into convexity and it reaches a relaxed energy to achieve a global minima, e.g., mean field annealing (MFA) [81], [229].

4. **GNC relaxation to global convergence.** The graduated non-convexity (GNC) algorithm proposed by Blake and Zisserman [33] constructs an approximating convex function free of spurious local minima, while stochastic methods avoid local minima by using random motions to jump out of them. The underlying principle of the general GNC alogrithm is also "convert" the nonconvexity to convexity because the algorithm approximates the global minima by minima of suitable approximating functions. Blake and Zisserman [31] has made a detailed comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. Piecewise continuous reconstruction of real-valued data can be formulated in optimization problems. They also point out that the deterministic algorithm (GNC) outstrips stochastic (simulated annealing) algorithms both in computation efficiency and in problem-solving power.

5. **The graph cuts algorithm to global convergence.** The graph cuts algorithm was first used in combinational optimization by Greig et al. [98] and recently was intensively studied for computer vision tasks [132]. The algorithm is based on linear programming and is used for binary optimization with the min-cut/max-flow algorithm. Each variable in this algorithm has one of two possible values. The cost function in the graph cuts algorithm need *not* to be convex but the cost function must be regular.

We can observe that most of these modeling criteria are based on the convexity for achieving global convergence. Following Fig. 2.7, global convergence via stochastic optimization can be computed in discrete spaces (using discrete simulated annealing, mean field theory, multiscale optimization) [127]. Global convergence via deterministic regularization approaches can be computed in continuous or discrete spaces (using continuous simulated annealing, conjugate gradient, gradient descent, Gauss-Seidel algorithm) [4], [234]. Therefore, there is an underlying relationship between discrete optimization and continuous optimization. Stochastic programming is the bridge between the two fields. Moreover, stochastic optimization and deterministic optimization approaches might be unified given certain conditions.

2.2.3 Stochastic Optimization and Regularization

The concept of an MRF is essentially due to Dobrushin (1968) [62] and is one way of extending Markvian dependence from 1-dimensional to general settings. It can also be considered as one kind of regularization of conditional probabilities and conditions. When *a priori* knowledge of statistical properties of the signal and of the noise is available, a probabilistic version of regularization methods is possible. Several authors have stressed the stochastic interpretation of spline approximation in which the smoothness properties of spline correspond to suitable prior probabilities [252]. Bertero, Poggio and Torre [25] have discussed a Bayesian approach which has the advantage of showing the connection between Markov Random Field models and standard regularization. In particular, they show how standard regularization can be regarded as a special case of MRF models and is itself equivalent to Wiener filtering for image restoration, and multilayer networks for learning [193]. These techniques, though computationally expensive, represent a powerful extension of the methods.

Markov random field (MRF) and compound Gaussian Markov random field (CGMRF) [26], [123] are widely used for image restoration and segmentation in computer vision since the milestone work of S. Geman and D. Geman [85]. Winkler [273], [272] considers various natural prior models in discrete Markov random field in a general Bayesian framework for image analysis (mainly on "inverse optics"). Dynamic Monto Carlo methods, stochastic relaxation algorithms and spectral graph methods are investigated and integrated in continuous time and space. Therefore, a bridge between discrete Markov chains and diffusion process is constructed, especially to indicate how different discrete processes can be embedded in the continuous setting to obtain comparison of their ergodic behaviour in stochastic and deterministic optimization manner. Hellwich [111] has developed an unsupervised Bayesian estimation in the MRF with stochastic simulated annealing for edge extraction and objects detection in synthetic aperture radar (SAR) data with stronger multiplicative noises. Because of the robustness with respect to noise and unsupervised properties, this approach can be directly extended to current computer vision problems such as unsupervised labeling, searching of natural prior knowledge, unsupervised Bayesian estimation based image restoration, segmentation and objects recognition. Molina, Katsaggelos et al. [169], have developed a multichannel image restoration algorithm using compound Gauss Markov Random Field (CGMRF) model. It is an extension of the classical simulated annealing and iterative conditional models. Figueiredo et al. [78] have proposed to interpret discontinuities (in fact their locations) as deterministic unknown parameters of the CGMRF, which is assumed to model the intensities. This strategy allows inferring the discontinuity locations directly from the image with no further assumptions. To solve the problem of number of unknown parameters (edges), they propose a new unsupervised discontinuity-preserving image restoration criterion using the minimum description length (MDL) criterion. Li [143] illustrates how to convert a

specific vision problem involving uncertainties and constraints into essentially an optimization problem under the MRF setting, the related problem of parameter estimation and function optimization. Recently, some learning-based methods in MRF have also been investigated, e.g., a Markov random field based filter learning method [209] for image denoising and segmentation [304].

An outstanding problem at present in the area of early vision is the detection and localization of discontinuities. Different methods, such as Markov Random Fields, seem capable of performing approximation and reconstruction while detecting and preserving discontinuities. There are promising approaches to the problem of integrating different visual modules such as stereo, motion, color, and texture that rely on coupled Markov Random Field models and their capability to detect and represent discontinuities. Because of the equivalence between regularization and generalized splines, it is impossible to deal directly with discontinuities in the framework of the classical regularization theory.

Moreover, the flexibility of partial differential equations based variational regularization can achieve similar or better results in that the smoothing term in the variational regularization can be extended to linear or nonlinear, isotropic or anisotropic, flow-driven, data-driven or knowledge-driven image diffusion and smoothing. Partial differential equation (PDE) based variational regularization approaches [258], [259], [257], [267], [265], [6], [7], [221], [219], [173], [184], [185], [186], [16] have been intensively developed for image processing since the 1990s. The partial differential equations, which belong to one of the most important parts of mathematical analysis, are closely related to the wave equation and the heat equation in the physical and the mechanical world. PDEs have been also extended into biology, finance analysis and so on. Once the existence and the uniqueness have been proven, we can directly refer to analogue images in the continuous setting using the well established PDE theory.

PDEs-based models are very important for our work due to their efficiency and robustness. In detail, how to understand and integrate PDE-based diffusion methods in a flexible manner is still an attractive task. Consequently, this chapter presents some state-of-the art work, novel modification and detailed experiments for deblurring and denoising from different point of view, e.g., image processing, energy optimization, computational physics and human visual perception. The term definitions and classification are following the definition from Aubert and Kornprobst [16] and Weickert [259][1].

2.3 PDE-Based Image Diffusion Filters in Scale Spaces

The study of PDE-based nonlinear isotropic and anisotropic diffusion filters is very important. First, linear and nonlinear diffusion operators can smooth and enhance images independently. Second, these operates can be included in penalty terms in variational regularization approaches. These smoothing terms can help "smoothing or enhancing" the convexity of energy functionals which can achieve the global convergence. Finally, data-driven image diffusion can be achieved by integrating or modifying these operators.

According to the image degradation model that has been defined previously, we have the following form $g = Hf + \eta$. A restored image $\hat{f}$ can be considered as a version of the observed image g at a special scale. Precisely, we consider an ideal image f being embedded in an evolution

[1] Here we refer to the definitions of nonlinear isotropic and anisotropic diffusion

Table 2.2: PDE-based Approaches

PDE Classification	Main Explanation
Smoothing PDE:	Heat equation and related diffusion filters.
Smoothing-enhancing PDE:	Perona-Malik model, Weickert's approach, etc.
Enhancing PDE:	Osher-Rudin shock filters, Alvarez-Mazorra filter.

process. We denote it by $f(t,\cdot)$. At time $t=0$, $f(0,\cdot)=f_0(\cdot)$ is the input image. Based on the transformation of the Hamilton-Jacobi equations and the theory of viscosity solutions, the restored image with an evolution process can be formulated in a generic form,

$$\begin{cases} \frac{\partial f}{\partial t}(t,x) + F(x, f(t,x), \nabla f(t,x), \nabla^2 f(t,x)) = 0 \text{ in } (0,T)\times\Omega \\ \frac{\partial f}{\partial N}(t,x) = 0 \text{ on } (0,T)\times\Omega, \text{ (Neumann boundary condition)} \\ f(0,x) = f_0(x), \text{ (initial condition)} \end{cases} \tag{2.12}$$

where $F(x, f(t,x), \nabla f(t,x), \nabla^2 f(t,x))$ with $(t \geq 0,\, x \in \Omega)$ is a second-order differential operator. $f(t,x)$ is a restored image of the initial observed image $f_0(x)$. ∇f and $\nabla^2 f$ are the gradient and Hessian matrix of f with respect to the space variable x, t is a scale variable. This form is a very generic form, different modification of this generic operator F can get different diffusion operators for the target of smoothing and discontinuity-preserving. The generic PDE-based diffusion filters are classified according to Aubert and Kornprobst [16] and Weickert [259] in Table. 2.2: (1) Smoothing or forward-parabolic PDEs, used mainly in pure restoration [7]. (2) Smoothing-enhancing or backward-parabolic PDEs concerning restoration-enhancement process [191]. (3) Hyperbolic PDEs for enhancing blurred images, focusing on shock filters [185] and shock filter combining anisotropic diffusion [6].

2.3.1 From Linear to Nonlinear Smoothing PDEs

The most popular and classical smoothing PDE in image restoration is the parabolic *linear* heat equation [7]:

$$\frac{\partial f}{\partial t}(t,x,y) = \Delta f(t,x,y)) \tag{2.13}$$

where Δ is the Laplace operator $\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2$. After some evolution time t, $f(t,x,y)$ is an unique solution in this equation. The main property of this solution is equivalent to a Gaussian convolution of the observed image g with the Gaussian kernel G_σ and the standard deviation $\sigma = 2\sqrt{t}$ according to [264]. Thus we have $f(t,x,y) = (G_\sigma * g)(x,y)$ with $G_\sigma = \frac{1}{2\pi^2}\exp\left[-(x^2+y^2)/(2\sigma^2)\right]$. Because of its over-smoothing property in linear homogeneous diffusion, some *nonlinear properties* for preserving discontinuities are introduced in this model. From the theoretical point of view, the equivalence of Gaussian convolution and linear diffusion is useful to replace linear homogeneous diffusion by nonlinear inhomogeneous diffusion in a similar formalism.

$$\frac{\partial f}{\partial t} = \mathrm{div}(c(|\nabla f|^2)\nabla f) \tag{2.14}$$

Table 2.3: Smoothing-Enhancing Nonlinear Diffusion Filters

Nonlinear Diffusion	Functionals
Scalar-valued diffusion:	$\frac{\partial f}{\partial t} = \mathrm{div}(c(\lvert\nabla f\rvert^2)\nabla f)$, initial value $f(x,y,0) = I(x,y)$
Scalar-valued diffusion:	$\frac{\partial f}{\partial t} = \mathrm{div}(c(\lvert\nabla G_\sigma * f\rvert^2)\nabla f)$, initial value $f(x,y,0) = I(x,y)$
Vector-valued diffusion:	$\frac{\partial f_i}{\partial t} = \mathrm{div}\left(c\left(\sum_{m=1}^{M} \lvert\nabla f_m\rvert^2\right)\nabla f_i\right)$, $i = 1,...,M$
Matrix-valued diffusion:	$\frac{\partial f_{ij}}{\partial t} = \mathrm{div}\left(c\left(\sum_{m=1}^{M}\sum_{n=1}^{N} \lvert\nabla f_{mn}\rvert^2\right)\nabla f_{ij}\right)$, $i = 1,...,M; j = 1,...,N$
Anisotropic diffusion:	$\frac{\partial f_{ij}}{\partial t} = \mathrm{div}\left(D\left(\sum_{m=1}^{M}\sum_{n=1}^{N} \nabla f_{mn}\nabla f_{mn}^\top\right)\nabla f_{ij}\right)$, $i = 1,...,M; j = 1,...,N$

where the function c is fixed for keeping the equation remains parabolic and ∇ is the gradient operator $(\partial_x + \partial_y)^\top$ In order to preserve the discontinuities,

$$c(|\nabla f|^2) \approx \frac{1}{\sqrt{|\nabla f|^2}} \tag{2.15}$$

where $|\nabla f|^2 \to +\infty$ is assumed. To study this equation, a framework of nonlinear semigroup theory is well-adapted. The basic idea is to show that the divergence operator *div* in this equation is maximal monotone. A convenient way to demonstrate this is to identify the divergence operator with the sub-differential of a convex lower semi-continuous functional in a established theory of existence, uniqueness of a solution.

Alvarez-Guichard-Lions-Morel have introduced a very original notion of scale-space via PDEs [7] in 1992. Given some axioms and invariance properties for an "image-oriented" operator T_t, the idea is to try to identify this operator. The model can be established that $f(t,x) = (T_t f_0)(x)$ is the unique viscosity solution of

$$\frac{\partial f}{\partial t} = F(\nabla f, \nabla^2 f) \tag{2.16}$$

In this equation, if T_t satisfy some natural assumptions, it can be solved through a PDE depending only on the first and second derivatives of f. F can then be solved introducing more assumption. One of Weickert's diffusion methods [259] can be considered as a tensor-based version of Eq. 2.14 where the scalar coefficient c controls the diffusion which is replaced by a function of the diffusion tensor,

$$\nabla f \nabla f^\top = \begin{pmatrix} f_x^2 & f_x f_y \\ f_x f_y & f_y^2 \end{pmatrix} \tag{2.17}$$

2.3.2 Nonlinear Smoothing-Enhancing PDEs

Perona-Malik filter

An important improvement of the classical linear analysis, with a more accurate multi-scale edge detection, was proposed by Perona and Malik [191]. The main idea of Perona and Malik is to

introduce a part of the edge detection step in the filtering itself, allowing an interaction between scales into the algorithm. They proposed to replace the linear heat equation by a nonlinear equation. The nonlinear diffusion equations Eq. 2.14 can thus behave locally as inverse heat equations due to the choices of c. Perona and Malik suggested two diffusion coefficients of c,

$$c(|\nabla f|^2) = e^{-|\nabla f|^2/k^2} \qquad \text{and} \qquad c(|\nabla f|^2) = \frac{1}{1+|\nabla f|^2/k^2} \tag{2.18}$$

where k is a contrast parameter to be tuned for a particular application. These two diffusion coefficient c are scalar-valued, decreasing functions with isotropic but non-homogeneous [230] nonlinear diffusion [259]. They also have similar basic properties such as positive coefficient, non-convexity and the ability of local enhancement for distinguished gradients. Detailed analysis and comparison of two diffusion coefficients and k are shown in Fig. 2.8. The P-M process removes noise while keeping edges and discontinuities. Some isolated noise points still remain, while some detailed textural information is lost during the process.

Catte filter

However, some drawbacks and limitations of the original model can drive the diffusion process to undesirable results [191] as mentioned in this original paper. For example, "staircasing" effects can easily happen around smooth edges. The ill-posedness of the diffusion may be alleviated through a smoothing operation to the variable in the diffusion coefficient $c(s) = c(|\nabla f|^2)$ in a regularized framework. This idea was introduced by Catte et al. [38] and the P-M model Eq. 2.14 is extended to in the following,

$$\frac{\partial f}{\partial t} = \operatorname{div}(c(|\nabla G_\sigma * f|^2)\nabla f) \tag{2.19}$$

with Neumann boundary condition. $\nabla G_\sigma * f$ denotes a convolution of the image at time t with a Gaussian kernel of standard deviation σ, which is to be given *a priori*. This formulation has solved a theoretical problem associated with Perona-Malik process. However, the selection of σ is critical to the Catte diffusion in the sense that the diffusion process would be ill-posed for too small scale, while the image features would be smeared for too large scale σ. One possible solution is to use a large scale initially to suppress the noise and then to reduce the scale so the image features are not further smeared [283]. Thus, the optimal selection of scale is still an open question. In Fig. 2.9, we show the role of the standard deviation σ. The isolated noise points are "cleaned" using the right σ. However, some detailed textural information is still weakly lost.

Different smoothing-enhancing nonlinear filters

Through the literature study, we list some other state-of-the-art nonlinear diffusion filters in Table. 2.3. These nonlinear smoothing-enhancing diffusion filters have been developed for achieving different purposes in image restoration:

1. Based on the pioneer work of *scalar-valued diffusion* technique from Perona and Malik [191], we can directly smooth color images in each channel independently. However, this

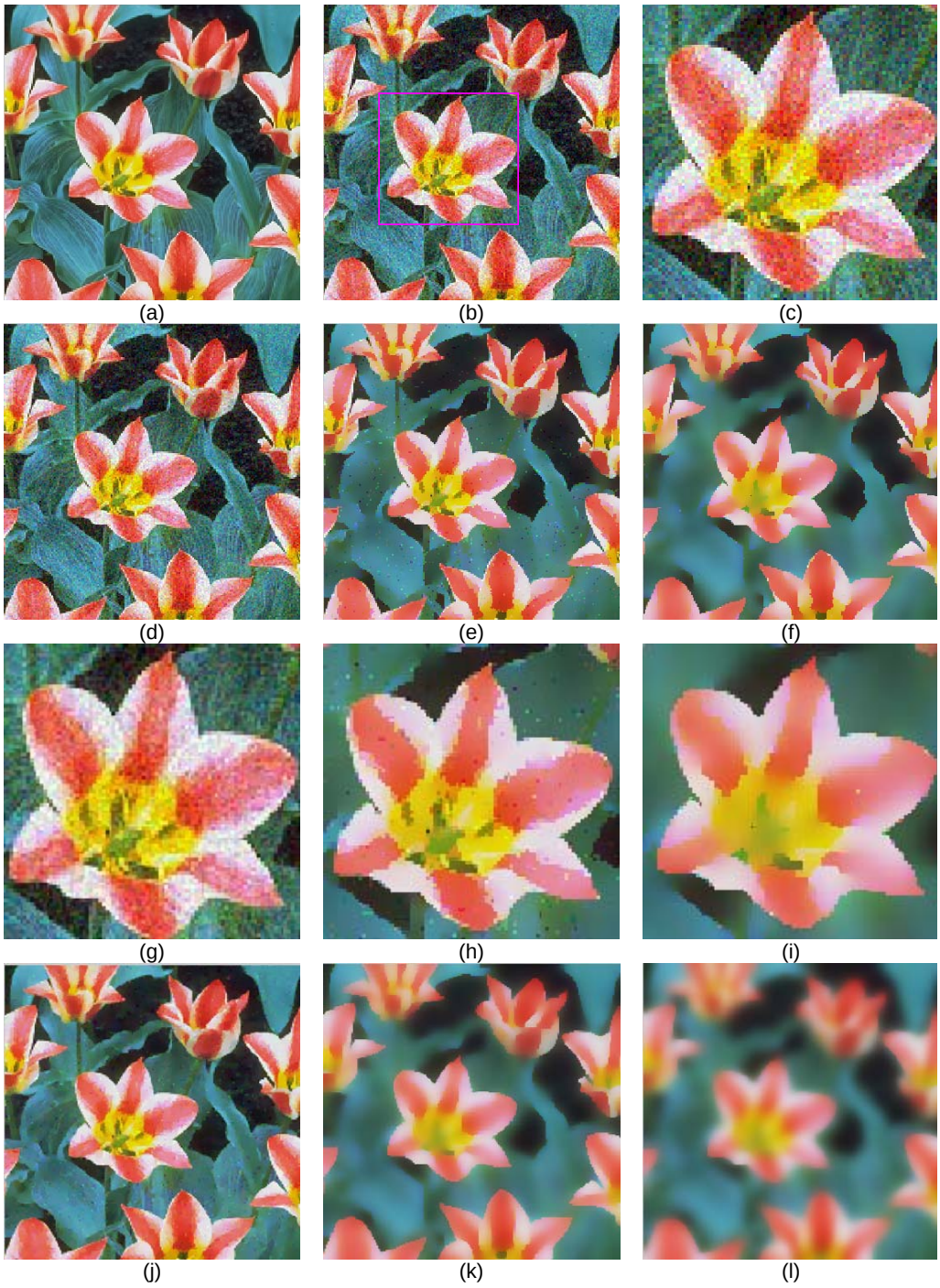


Figure 2.8: Perona-Malik (P-M) scalar-valued image diffusion filter. (a) Original tulip image. (b)(c) Input color image with independent Gaussian noise in each RGB color channel, sigma =20. (g)(h)(i) zoom in (d)(e)(f). (d)(e)(f) are processed on each channel using P-M(I): $c(s^2) = \exp(\frac{-s^2}{2k^2})$ with respect to k =5, k= 20, k=35. (j)(k)(l)are processed on each channel using P-M(II): $c(s^2) = \frac{1}{1+s^2/k^2}$ with respect to k=5, k=20, k=35. Comparing (d and j), (e and k), (f and l), we can note that P-M(II) is relatively stronger than P-M(I) for image diffusion. However, these two filters have the same properties. Isolated noise points have disappeared in the whole image, some textural information is completely lost. For such methods it can be shown that small scales are smoothed faster than large ones, so if the method is stopped at a suitable final time, we may expect that noise is smoothed while large-scale features are preserved to some extent.

(a) (b) (c)

Figure 2.9: Comparison of different σ value to suppress isolated noise points in Catte diffusion. Perona-Malik $c(s^2) = \exp(\frac{-s^2}{2k^2})$ with k= 20 based Catte diffusion $\frac{\partial f}{\partial t} = \text{div}(c(|\nabla G_\sigma * f|^2)\nabla f)$. We input a similar noise color image for testing. (a) $\sigma = 0.1$. (b) $\sigma = 0.2$. (c) $\sigma = 0.3$. While $\sigma = 0.3$, isolated noise points have disappeared in the whole image, but some detailed textural information have lost.

procedure is not an optimal method in that the diffusion ignores the information from its neighbor channels. To find an optimal approach, *vector-valued diffusion* filters have been proposed for solving this problem [246]. The main idea is to smooth all vector channels f_i using a joint gradient information of all channels which can achieve better optimized results.

2. Following the vector-valued diffusion scheme, one can introduce nonlinear diffusion for matrix-valued data which is useful for tomography data processing. Different from the *matrix-valued diffusion* scheme [244], a coupling between all matrix channels is crucial for preserving matrix properties [262], orientation estimation in the matrix fields and different goals in smoothing.

3. Although these diffusion schemes enhance and restore images in an edge-preserving manner, the diffusion strength is performed equally in all directions. An ideal diffusion manner is to perform the smoothing along edges without smoothing across edges. Following this idea, Weickert et al. [259] developed an *anisotropic diffusion filter* which replaces the scalar-valued diffusivity c by a matrix-valued diffusion tensor D. The matrix tensor D controls smoothing along edges with forward diffusion and enhancing edges by backward diffusion in perpendicular direction simultaneously. The anisotropic diffusion method is useful in many area. Further study is referred to [259], [256], [258], [261], [267].

4. The *structure tensor* [79] is listed here because it is closely related to those diffusion filters and have gained significant importance in the field of scientific visualization and image processing [263]. It is also named second moment matrix which includes the estimation of orientation and the local analysis of image structure. Structure tensor is a fundamental concept for corner detection [79] [207], passive navigation [103], image segmentation [155] as well as surface reconstruction. The linear structure tensor is based on Gaussian convolution, while a nonlinear structure tensor is based on different nonlinear diffusions. Therefore, the underlying mechanism defining structure tensors can be implemented by means of different optimization approaches in scale space based on the demands of the goal. Currently, many elegant methods have been developed in the tensor field referring to the book by Weickert and Hagen [263].

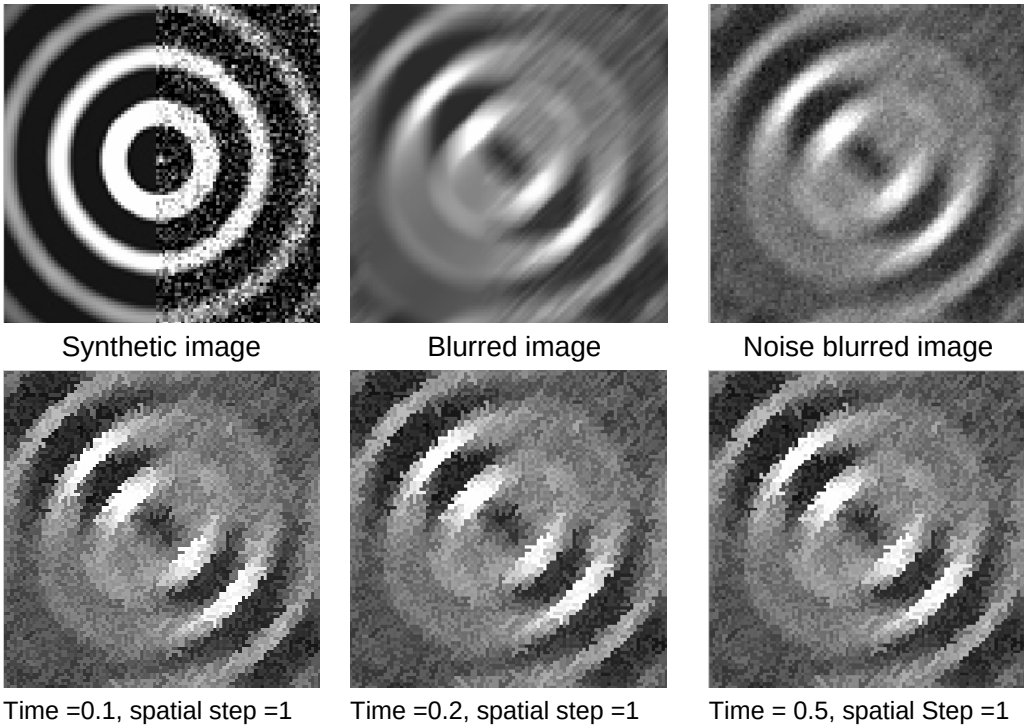


Figure 2.10: Shock filter diffusion and sharpening. From these experiments, we can summarize the main properties of the shock filter. Firstly, the filter is local extrema remain unchanged in time. No erroneous local extrema are created. Secondly, the steady state (weak) solution is piecewise constant (with discontinuities at the inflection points of f_0. Thirdly, the process can be approximated to deconvolution. Finally, the shocks amplify at inflection point (second derivative zero-crossings).

2.3.3 Enhancing and Sharpening PDEs

Shock filter

The deblurring or enhancement is essentially devoted to the shock filter model in a hyperbolic equation proposed by Osher and Rudin[185] (1990). This filter can serve as a stable deblurring algorithm approximating deconvolution. In 1-D case, the shock filter model is following,

$$f_t(t,x) = -\ \text{sign}(f_{xx}(t,x))|f_x(t,x)| \tag{2.20}$$

where $\text{sign}(f_{xx}(t,x)) = 1$ if $(f_{xx}(t,x)) > 0$, $\text{sign}(f_{xx}(t,x)) = -1$ if $(f_{xx}(t,x)) < 0$, $\text{sign}(0) = 0$. To better understand the action of this shock filter, the equation can be written in a simpler way. The initial conditions $f(x,0) = f_0(x)$ and Neumann boundary conditions ($\frac{\partial f}{\partial N} = 0$ where N is the direction perpendicular to the boundary) are used.

In the 2D case, the shock filter is commonly generalized to the following equation,

$$f_t(t,x) = -\ \text{sign}\ (f_{\delta\delta}(t,x,y))|\nabla f| \tag{2.21}$$

where δ is the direction of the gradient. The discretization of the 1D process is approximated by the following discrete scheme,

$$f_i^{(n+1)} = f_i^{(n)} - \Delta t|Df_i^n|\ \text{sign}\ (D^2 f_i^n) \tag{2.22}$$

where $Df_i^n = m(\Delta_+ f_i^n, \Delta_- f_i^n)/h$ and $D^2 f_i^n = (\Delta_+ \Delta_- f_i^n)/h^2$. $m(x, y)$ is the minmod function that is developed based on the work of Osher, Rudin and Sethian [184], [185].

$$m(x,y) = \begin{cases} (\text{sign}) \min(|x|, |y|) & \text{if } xy > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2.23}$$

where $\Delta_+ = (f_{i+1} - f_i)$ and $\Delta_- = -(f_{i+1} - f_i)$. As mentioned in the original paper, any noise in the blurred signal will also be enhanced. Thus, this filter has some sensitive properties to noise. From experimental results, we can find that any white noise added to the signal is amplified and disrupts the diffusion process. Similar to the previous work, the common way is to convolve the signals in a second derivative with a lowpass filter, for example, convolve with a Gaussian kernel in 1D case,

$$f_t(t,x) = -\text{ sign}\left[G_\sigma * f_{xx}(t,x)\right] |f_x(t,x)| \tag{2.24}$$

where G_σ is a Gaussian kernel with standard deviation σ. We can note that the filter has the same scale control problem as in the Catte's filter [38]. For the large scale Gaussian, most noise and its generated inflection points are diminished with a cost of lower accuracy at the first step. Secondly, the width of Gaussian σ is normally larger than the length of signal, the boundary conditions can strongly affect to the solution. Thirdly, convolving with a Gaussian kernel is equivalent to a change of sign at each pixel, the diffusion flow may go in opposite direction at each side. The improved idea is to smooth the noise parts, whereas edges are enhanced.

Regularized shock filters

Alvarez and Mazorra (1994) [6] were the first to couple the shock and the diffusion process for enhancing edges and smoothing noise simultaneously. The idea is to add a penalty enhancing term based on Eq. 2.24. The equation becomes,

$$f_t = -\text{ sign}\left[G_\sigma * f_{TT}\right] |\nabla f| + \lambda f_{\xi\xi} \tag{2.25}$$

where λ is a positive constant and ξ is the direction perpendicular to the gradient ∇f. Roughly speaking, the shock filter in one dimension develops shocks in the position of the zero crossing of $G_\sigma * f_{TT}$, and it produces an enhancement of the edges in this way. In two-dimensional space (i.e. image), the parabolic-hyperbolic equation diffuses the initial image $f(x, y, 0)$ in the directions parallel to the edges (noise elimination) and develops a shock in the perpendicular direction of the edge (edge enhancement and deconvolution). The boundary conditions are also

Table 2.4: Enhancing and Sharpening PDEs Diffusion Filters

Enhancing Filters,	Formula of Enhancing Filters
Alvarez and Mazorra :	$f_t = -\text{ sign}\left[G_\sigma * f_{\delta\delta}\right] \lvert\nabla f\rvert + \lambda f_{\xi\xi}$
Kronprobst et al. :	$f_t = \alpha_f (f - f_0) + \alpha_r (h_\tau f_{\delta\delta} + f_{\xi\xi}) - \alpha_e (1 - h_\tau) \text{ sign}\left(G_\sigma * f_{\delta\delta}\right) \lvert\nabla f\rvert$
Coulon et al.:	$f_t = \text{div}(\lambda \nabla f) - (1 - \lambda)^\alpha \text{ sign}\left(G_\sigma * f_{\delta\delta}\right) \lvert\nabla f\rvert$
Gilboa et al:	$f_t = -\frac{2}{\pi} \arctan(a f_m(\frac{f}{\theta})) \lvert\nabla f\rvert + \lambda f_{\xi\xi}$

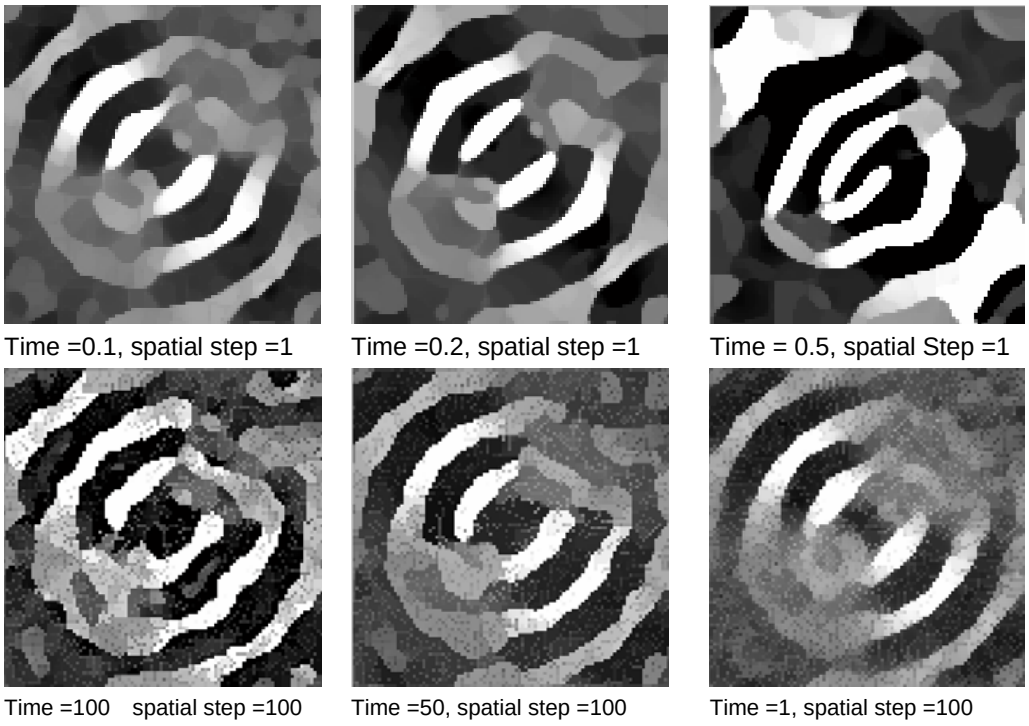


Figure 2.11: Alvarez-Mazorra filter for denoising and deblurring. The noisy blurred image is the same image that is used in the shock filter. From these experiments, we summarize several properties. First, we note that the noise is diminished. while time=0.1, spatial step =1, the restoration result is best. Second, this filter approximates deconvolution for deblurring. The discontinuities are enhanced while the spatial step is bigger, the individual noise points can not be diminished. Third, the results in different time scales and spatial step scales show a "balance" between time scales and spatial step scales. Good "balance" can achieve better restoration results. All evolution results are for 100 iterations.

imposed in a natural way to minimize the boundary influence., e.g. $\partial f/\partial N = 0$. Experiments of the Alvarez-Mazorra filter are shown in Fig. 2.11.

Some related filters have also been developed based on the combination of the smoothing term and the enhancing term using regularization strategies, shown in Table. 2.4.

1. Kornprobst [133] proposed an advanced scheme. The fidelity $\alpha_f(f - f_0)$ keeps the fidelity of the original image. In the equation, $h_\tau = h_\tau(|G_{\bar{\sigma}} * \nabla f|) = 1$ if $|G_{\bar{\sigma}} * \nabla f| < \tau$ and 0 otherwise.

2. Coulon and Arridge [52] developed this functional that was originally used for classification in a probabilistic framework. The functional is adapted for image denoising, where $\lambda = \exp(\frac{|G_\sigma * \nabla f|^2}{k})$.

3. One of the most complex filters is developed by Gilboa et al. [89]. A complex diffusion term is proposed to the shock filter equation and achieved to smooth out noise and indicate inflection points simultaneously. The main principle is to smooth the images by using second derivative scaled by time and control the process in a regularization framework which is somehow to find a balance between smoothing and sharpening in the restoration.

2.3.4 Inverse Scale Space Methods

Since the noise in images is usually expected to be a small scale feature, particular attention has been paid to methods separating scales, in particular those smoothing small scale features faster than large scale ones, so-called *scale space methods* [191], [276]. *Inverse scale space methods* have been introduced in [217], which are based on a different paradigm. Instead of starting with the noisy image and gradually smoothing it, inverse scale space methods start with the image $f(x, 0) = 0$ and approach the noisy image g (which will be normalized to have mean zero) as time increases, with large scales converging faster than small ones. Thus, if the method is stopped at a suitable time, large scale features may already be incorporated into the reconstruction, while small scale features (including the "noise") are still missing.

The inverse scale space method can also be related to regularization theory, in particular iterated Tikhonov regularization [217], [99] with the same regularization functionals as for diffusion filters. The construction of inverse scale space methods in [217] worked well for quadratic regularization functionals, which led to an interesting, but linear evolution equation, but did not yield convincing results for other important functionals, in particular for the total variation functional [213].

Through the literature, some main properties and utilization of inverse scale space have been discussed and investigated. For example, Bregman distance is stronger than L^2 for the regularization, Bregman distance is used in regularization for denoising and relaxed inverse scale space methods. Recently, Xu and Osher [276] introduced a different version of constructing inverse scale space methods as the limit of an iterative regularization to image restoration. With this approach, they have implemented nonlinear inverse scale space methods for the total variation functional and, in contrast to diffusion filters, a rigorously justified and simple stopping criterion is obtained for the methods. This approach has obtained encouraging restoration results. Moreover, inverse scale space is applied to wavelet based image denoising according to Xu [276], Didas and Weickert [61] in single scale and multiple scale cases.

In other words, inverse scale space can be considered as scale space interpreted regularization for inverse problems. It integrates several advantages from scale space and regularization, e.g., accurate stopping time for continuous evolution, faster computation after some relaxation, reduction of complexity using forward Euler time integration. It seems clear that the appealing aspect of these characteristic properties is due its scale space interpretation.

2.4 Boundary Conditions

In mathematics, a boundary value problem consists of a differential equation and the initial or boundary values required to solve the equation. The solution to the differential equation will not only satisfy the differential equation everywhere inside the boundary but will also satisfy the boundary conditions themselves. Boundary value problems may be posed for ordinary differential equations as well as partial differential equations. To be useful in applications, a boundary value problem should be well posed. This means that given the input to the problem there exists a unique solution, which depends continuously on the input. Much theoretical work in the field of partial differential equations is devoted to proving that boundary value problems arising from scientific and engineering applications are in fact well posed.

The difficulties caused by boundary conditions in computing would be hard to overemphasize.

Boundary conditions can easily make the difference between a successful and an unsuccessful computation, or between a fast one and a slow one. Yet in many important cases, there is little agreement about what the proper treatment of the boundary should be. Images are shown in a finite region, but points near the boundary of a boundary image are likely to have been affected by information outside the field of view. Since this information is not available, for computational purposes, we need to make some assumptions about the boundary conditions.

Normally, deblurring and denoising use same boundary conditions. The only difference is that the deblurring needs to consider the PSF and the image deconvolution. For deblurring problem, we begin with the one-dimensional deblurring problem. Consider the original signal

$$\tilde{f} = (\ldots, f_{-m+1}, \ldots, f_0, f_1, \ldots, f_n, f_{n+1}, \ldots, f_{n+m}, \ldots)^\top \tag{2.26}$$

and the blurring function given by

$$\mathrm{h} = (\ldots, 0, 0, h_{-m}, h_{-m+1}, \ldots, h_0, \ldots, h_{m-1}, h_m, 0, 0, \ldots)^\top \tag{2.27}$$

The blurred signal is the convolution of h and $\tilde{f}$, i.e., the ith entry g_i of the blurred signal is given by

$$g_i = \sum_{j=-\infty}^{\infty} h_{i-j} f_j \tag{2.28}$$

The deblurring problem is to recover the vector $\mathrm{f} = (f_1, \ldots, f_n)^\top$ given the blurring function h and a blurred signal $\mathrm{g} = (g_1, \ldots, g_n)^\top$ of finite length. From Eq. 2.28, we have

$$\begin{pmatrix} h_m & \cdots & h_0 & \cdots & h_{-m} & & & & & \\ & h_m & & h_0 & & h_{-m} & & & 0 & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & 0 & & & h_m & & h_0 & & h_{-m} & \\ & & & & & h_m & \cdots & h_0 & \cdots & h_{-m} \end{pmatrix} \begin{pmatrix} f_{-m+1} \\ f_{-m+2} \\ \vdots \\ f_0 \\ f_1 \\ \vdots \\ f_n \\ f_{n+1} \\ \vdots \\ f_{n+m-1} \\ f_{n+m} \end{pmatrix} = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix} \tag{2.29}$$

Thus the blurred signal g is determined not by f only, but also by

$$(f_{-m+1}, \ldots, f_0, f_1, \ldots, f_n, f_{n+1}, \ldots, f_{n+m})^\top$$

The linear system Eq. 2.29 is underdetermined. To overcome this, we make certain assumptions (called boundary conditions) on the unknown data and $f_{-m+1}, \ldots, f_0$ and $f_{n+1}, \ldots, f_{n+m}$ so

as to reduce the number of unknowns. We can rewrite this equation in the following form for discussing the boundary conditions,

$$T_l \mathrm{f}_l + T\mathrm{f} + T_r \mathrm{f}_r = \mathrm{g} \tag{2.30}$$

where

$$T_l = \begin{pmatrix} h_m & \cdots & h_1 \\ & \ddots & \ddots \\ & & h_m \\ 0 & & \end{pmatrix}, \quad \mathrm{f}_l = \begin{pmatrix} f_{-m+1} \\ f_{-m+2} \\ \vdots \\ f_{-1} \\ f_0 \end{pmatrix} \tag{2.31}$$

$$T = \begin{pmatrix} h_0 & \cdots & h_{-m} & & 0 \\ \vdots & \ddots & \ddots & \ddots & \\ h_m & \ddots & \ddots & \ddots & h_{-m} \\ & \ddots & \ddots & \ddots & \vdots \\ 0 & & h_m & \cdots & h_0 \end{pmatrix}, \quad \mathrm{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix} \tag{2.32}$$

$$T_r = \begin{pmatrix} & & 0 \\ & & \\ h_{-m} & & \\ \vdots & \ddots & \\ h_{-1} & \cdots & h_{-m} \end{pmatrix}, \quad \text{and} \quad \mathrm{f}_r = \begin{pmatrix} f_{n+1} \\ f_{n+2} \\ \vdots \\ f_{n+m-1} \\ f_{n+m} \end{pmatrix} \tag{2.33}$$

2.4.1 Dirichlet Boundary Conditions

The Dirichlet (zero) boundary condition assumes that the signal outside the domain of the observed vector g is zero, i.e., $\mathrm{f_l} = \mathrm{f_r} = 0$ are the zero vectors. The matrix system in Eq. 2.29 becomes

$$T\mathrm{f} = \mathrm{g} \tag{2.34}$$

where T is a Toeplitz matrix. There are many iterative or direct Toeplitz solvers that can solve the Toeplitz system with cost ranging from $O(n \log n)$ to $O(n^2)$. Zero boundary conditions imply a black boundary, so that the pixel outside the borders of the image X are all zero. i.e. A is a block Toeplitz matrix with Toeplitz blocks ($BTTB$) in the 2D case. Matrix vector multiplication are done by embedding A into a larger $BCCB$ matrix, padding outside the borders of the image with an appropriate number of zeros, and then using FFTs. The amount of padding depends on the extent of the PSFs. For a large size image, overlap-add and overlap-save memory methods are used to partition the image domain into regions based on the size of PSF. In linear algebra terms, the approach is equivalent to exploiting sparsity (bandedless) of the matrix h (PSF).

2.4.2 Periodic Boundary Conditions

For practical applications, especially in the two-dimensional case, where we need to solve the system efficiently, one usually resorts to the periodic boundary condition. This amounts to setting

$$f_j = f_{n-j}, \quad \text{for all } j \tag{2.35}$$

for the Eq. 2.29. The matrix system in Eq. 2.30 becomes

$$B\mathrm{f} = [(0|\, T_l) + T + (T_r|\, 0)]\,\mathrm{f} = \mathrm{g} \tag{2.36}$$

where $(0|\, T_l)$ and $(T_r|\, 0)$ are n-by-n Toeplitz matrices obtained by augmenting $(n-m)$ zero columns to T_l and T_r, respectively.

The most important advantage of using the periodic boundary condition is that B so obtained is a circulant matrix. Hence B can be diagonalized by the discrete Fourier matrix and the Eq. 2.36 can be solved by using three FFTs (one for finding the eigenvalues of the matrix B and two for solving the system. Thus the total cost is of $O(n \log n)$ operations.

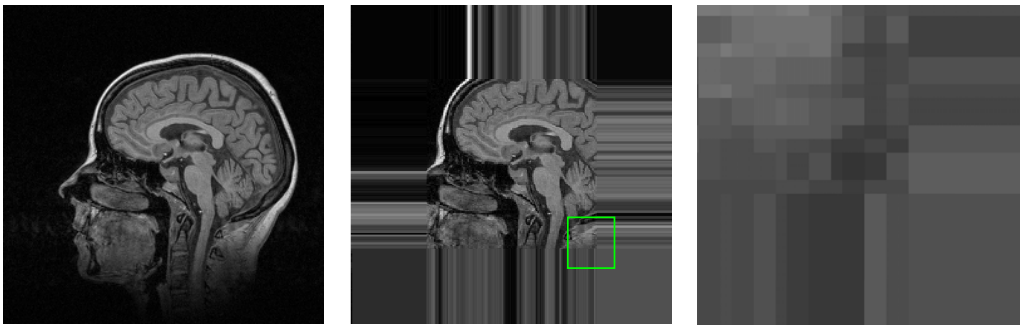
In the two-dimensional case, the blurring matrix is a block-circulant-circulant-block (BCCB) matrix and can be diagonalized by the two-dimensional FFTs (which are tensor-products of one-dimensional FFTs) in $O\left(n^2 \log n\right)$ operations. Periodic boundary conditions for a rectangular domain lead instead to a wrap-around of image information between opposite boundaries. The periodic boundary condition satisfies $f(x+h) = f(x)$ for all x in R^n, h in Z^n. The natural extension of an image adopted to extend f by reflection across the boundary of the rectangle. Thus, $f(-x, y) = f(x, y)$, if $-1 \leq x \leq 0$, and $0 \leq y \leq 1$, etc. are defined. We can easily find that f is assumed to be periodic on $(2Z)^n$.

Moreover, the periodic boundary conditions also introduce discontinuities which entail ringing artifacts or some false discontinuities and edges in the boundaries of the restored image frequently. To mitigate these artifacts as well as the undesired wrap-around of image information in the deblurring with periodic boundary conditions, the image can be extended continuously to a larger image with equal gray-values at opposing boundaries. Periodic boundary conditions will not introduce the false discontinuities or edges any more. The wrap-around influences the amended parts of the image. Periodic extension of this larger image is equivalent to reflecting extension of the original image. Fortunately, the periodic boundary conditions are compatible with any shift-invariant blur, without imposing symmetry constraints on the blur kernel. The periodic boundary conditions can be utilized directly or in a modified way.

2.4.3 Neumann Boundary Conditions

The Neumann boundary condition. For the Neumann boundary condition, we assume that the data outside f are a reflection of the data inside f. More precisely, we set

$$\begin{cases} f_0 = f_1 \\ \vdots \quad \vdots \quad \vdots \\ f_{-m+1} = f_m \end{cases} \text{and} \begin{cases} f_{n+1} = f_n \\ \vdots \quad \vdots \quad \vdots \\ f_{n+m} = f_{n-m+1} \end{cases} \tag{2.37}$$



(a) (b) (c)

Figure 2.12: Homogeneous Neumann boundary condition. (a) An original MRI head image. (b)(c) Homogeneous Neumann boundary condition can be implemented by mirroring many boundary pixels in four directions. Eq. 2.37 shows that the standard Neumann boundary condition is implemented by mirroring one boundary pixel in four directions.

Thus the original equation becomes

$$A\mathrm{f} = [(0|\, T_l)\, J + T + (T_r|\, 0)\, J]\, \mathrm{f} = \mathrm{g} \tag{2.38}$$

where J is the n-by-n reversal matrix. We remark that the coefficient matrix A in Eq. 2.38 is neither Toeplitz nor circulant. It is a Toeplitz-plus-Hankel matrix. Although these matrices have more complicated structures, the matrix A can always be diagonalized by the discrete cosine transform matrix provided that the blurring function h is symmetric, i.e., $h_j = h_{-j}$ for all j. It follows that Eq. 2.38 can be solved by using three FCTs in $O(n \log n)$ operations . This approach is computationally attractive as FCT requires only real operations and is about twice as fast as the FFT [160]. Thus solving a problem with the Neumann boundary condition is twice as fast as solving a problem with the periodic boundary condition. Ng et al. [175] proposed to establish similar results in the two-dimensional case for deblurring, where the blurring matrices will be block Toeplitz-plus-Hankel matrices with Toeplitz-plus-Hankel blocks (BTHTHB).

Neumann boundary conditions can be written in this PDE form ($\frac{\partial f}{\partial n}(x,t)$ on ∂R). As discussed in [7], the choice of Neumann boundary conditions is a natural choice in image diffusion. It corresponds to the reflection of the image across the boundary and has the advantage of not imposing any value on the boundary and not creating edges on it shown in Fig.2.12. The Neumann boundary conditions work well because these diffusion-based image processing methods in the PDE guarantee conservation properties as well as a continuous extension of the image at its boundary. Indeed, the Neumann condition corresponds to the reflection of the image across the boundary with the advantages of not imposing any value on the boundary and not creating "edges" on it. If we assume that the boundary of the image is an arbitrary cutoff of a large scene in view, the Neumann boundary condition is a therefore a natural method.

There is also a boundary conditions that is not often used. Reflexive boundary conditions imply that the scene outside the image boundaries is a mirror image of the scene inside the image boundaries. i.e. A is a sum of a $BTTB$ matrix and a block Hankel matrix wit Hankel blocks ($BHHB$). This case is similar to zero boundary conditions, except that the values that are padded around the outside of the image are obtained by reflecting the pixel values from the inside of the image boundaries. However, the utilization of reflecting boundary conditions for image deconvolution with space-invariant kernels is bound to fail if the kernel is not symmetric w.r.t. the image boundary directions. The reason is that the reflected parts of the image would be blurred with a reflected kernel, violating the model assumptions.

Based on the variational regularization, we model the unsupervised blur identification, image restoration and segmentation in a convex problem in that the convex problem can be solved reliably and efficiently. Several basic convex components can help to get convex optimization so that we can still achieve the results in a reliable and robust manner. When there are many feasible solutions that are consistent with both known prior information and the measured blurred image. A restored image can be defined as a continuous, strictly convex functional that assigns a cost to each feasible solution and the selects the one which minimizes the cost. There are two factors affecting the complexity of the algorithms. One is the cost functional itself and the other is the set of constraints for the restored image. Different strategy and choices may influence the restoration results at different degree. In addition, the prior information about the image and noise must be expressed in the form of equality of inequality constraints.

3 Bayesian Model Selection and Nonparametric Blur Identification

"A knowledge of statistics is like a knowledge of foreign languages or of algebra; it may prove of use at any time under any circumstances." – A. L. Bowley

In the previous chapter, we recall and discuss regularization for ill-posed inverse problems, PDE based image diffusions and energy functionals, and convex optimization for the construction of an image deblurring and denoising framework. In this chapter, firstly, we will introduce the important concepts in information theory, and related model selection methods as a necessary preparation of further discussion. Most of algorithms in statistical learning, including feature extraction and classification, are essentially process of information collection, transmission and utilization. The introduction of information theory into statistical learning opens a new perspective for us to explore the nature of these statistical learning topics. Secondly, we will introduce a nonparametric model selection based method for blur identification and analyze the experimental results for a large image or video sequence. The integration of nonparametric model selection techniques and locally parametric optimization techniques for blur identification are presented in the next chapter.

3.1 Introduction

In pattern classification, three main estimation techniques are intensively investigated, e.g., parametric estimation techniques, nonparametric estimation techniques, and semi-parametric estimation techniques.

Parametric estimation techniques are based on the assumption that the data set has a predefined distribution [63], [128], [188], which describes the data set in a compact way. For example, parametric density estimation assumes the data is drawn from a density in a parametric class, e.g., the class of Gaussian densities. The estimation problem can thus be assumed to finding the parameters of the Gaussian that fits the data set. However, in most pattern recognition and model selection, this assumption is suspect. The common parametric forms rarely fit the densities actually encountered in practice.

The Gaussian mixture modeling technique with isotropic covariance and anisotropic covariance is known as a semi-parametric density estimation technique. It is also to be placed in between two extremes such as parametric and non-parametric density estimation. To apply mixture models, we firstly need to think some basic questions, the number of components and which classes of component densities should be used. Therefore, the questions become a model selection problem.

The nonparametric estimation techniques that can be used with arbitrary distributions without any assumptions, i.e., the forms of underlying densities are unknown. There are several types of nonparametric methods of interest. One consists of procedures for estimating the density

functions from sample patterns. If these estimates are satisfactory, they can be substituted for the true densities for designing the classifier. Another one consists of procedures for directly estimating the *a posteriori* probability. This is closely related to nonparametric design procedures such as nearest-neighbor rule, which bypass probability estimation and go directly to decision functions. In this chapter, we focus on some non-parametric techniques to classify blur kernels and blurred images in large image or video sequences.

3.2 Bayesian Learning of Finite Mixture Models

The statistic learning approach to modeling data constructs models by starting with a flexible model specified by a set of parameters. The principle idea is that if we can explain our observation well, then we should also be confident that we can predict future observations well. We might also hope that the particular setting of the best-fit parameters provides us with some understanding of the underlying processes. The procedure of fitting model parameters to observed data is termed learning a model. The above idea can be formalized using the concept of probability and the rules of Bayesian inference.

3.2.1 Finite Mixture Models

Mixture models have been intensively investigated, e.g., McLachlan et al. [162], [163], Jain et al. [120], Titterington et al. [242], Figueiredo and Jain [77]. Mixture models are able to describe and represent arbitrary complex probability density functions (pdf's). This fact makes them an excellent choice for representing complex class-conditional pdf's [77]. For example, likelihood functions in Bayesian learning, or priors for Bayesian parameter estimation. Finite mixture models can also be used to perform feature selection. The general optimization method is to use *maximum likelihood* (ML) or *maximum a posteriori* (MAP) to estimate of the mixture parameters.

Let us denote the data set by y, which may be made up of several idependent and identically distributed variables indexed by n: $\mathcal{Y} = \{\mathbf{y}^{(1)}, ..., \mathbf{y}^{(n)}, ..., \mathbf{y}^{(N)}\}$. For example, $\mathcal{Y}$ could be a random sampling space of images for which the variables might be measurements of the type of descriptive features foundation(feature corners, textures, entropy, and some features in frequency domain). Generally each variable can be discrete or real-valued in the physical world. Statistic learning approaches define a generative model of the data through a set of parameters $\theta = \{\theta_1, ..., \theta_K\}$ which define a probability distribution over data, $p(\mathcal{Y}|\theta)$.

The generative models introduce hidden (latent) variables to account for the generating process of given data sets. One way to learning the model then includes finding the parameters θ^* such that

$$\theta^*_{ML} = \arg\max_{\theta} p(\mathcal{Y}|\theta) \tag{3.1}$$

The process is normally called maximum likelihood learning as the parameters θ^*_{ML} are set to maximize the likelihood of θ which is probability of the observed data under the model. While the maximum a posteriori (MAP) learning aims at the maximization of the posterior probability $p(\theta|\mathcal{Y}) \propto p(\mathcal{Y}|\theta)p(\theta)$ and yields,

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{Y}) = \arg\max_{\theta} p(\mathcal{Y}|\theta)P(\theta) \tag{3.2}$$

where $P(\theta)$ denotes the prior probability of different parameter values. We can note that the MAP estimator is a ML estimator with a prior. The difference between ML and MAP is a point of interest. In the case of infinite training samples, $N \rightarrow \infty$, and a selected prior distribution does not effect the outcome. ML and MAP deliver identical results. However, since the number of data samples is limited in the practical environments, there are some difference between these two methods. First, from the computational point of view, ML methods are often preferable since they are based on first and second order derivatives where the MAP approaches can result in a time-consuming high-dimensional integration. For interpretability reasons, ML estimates are calculated from one single model, while the MAP results in a weighted average of two models that can also have different functional forms. The difficulty in MAP estimation is the choice of a suitable prior distribution. To achieve this goal, a good strategy is not to take a fixed static prior but rather to derive it from the underlying data. This strategy is extended and further developed in our work. Good prior knowledge is "descriptive model" or "enhance" information which are built on features and statistics extracted from the signal, and use complex potential functions to characterize the given data [307].

3.2.2 Bayesian Parameter Estimation

The generative model may also include latent or hidden variables in a set of n labels $x = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\}$ associated with n samples. These variables are unobserved yet interactive through the parameters to generate the data. The probability of the data can then be formulated by summing over the possible settings of the hidden states:

$$p(\mathcal{Y}|\theta) = \sum_{x} p(x|\theta)p(\mathcal{Y}|x,\theta), \tag{3.3}$$

where the summation is called complete-data likelihood (single quantity is called incomplete-data likelihood) and is often replaced by an integral for those real-valued hidden variables. For a particular parameter setting, the posterior distribution over the hidden variables can be written using Bayesian rule,

$$P(x|\mathcal{Y},\theta) = \frac{p(\mathcal{Y}|x,\theta)P(x|\theta)}{p(\mathcal{Y}|\theta)} \tag{3.4}$$

For the case of blur identification, we might have some hidden variables for the estimation of blur, and this can be inferred based on the observation of descriptive measurements (e.g., the support size of blur kernels, the diminishing of discontinuities and mean square error of the restored image), through the generative model [307] with parameter θ. The term $P(x|\theta)$ is a prior probability of the hidden variables based on the modeling of prior knowledge which can reflect the distribution of parameters of the real blur kernel. Note that the probability of Eq. 3.3 is a denominator in Eq. 3.4. Since the hidden variables are unknown definition, finding θ^*_{ML} becomes more difficult. The model is learnt by alternating between estimating the posterior distribution over hidden variables for a particular setting of the parameters and then re-estimating the best-fit parameters given that distribution over the hidden variables. This method is the well-known expectation-maximization (EM) algorithm[60].

Given parameters are unknown quantities and we treat them as random variables. It is the Bayesian approach to uncertainty, i.e., Bayesian approach treat all uncertain quantities as random variables based on the laws of probability to manipulate those uncertain quantities. The

proper Bayesian approach integrates over the possible settings of all uncertain quantities rather than optimize them as in Eq. 3.1. Therefore, the marginal likelihood is the resulting quantity from integrating both hidden variables and the parameters.

$$P(\mathcal{Y}) = \int P(\theta) \sum_{x} P(x|\theta) p(\mathcal{Y}|x,\theta) d\theta \tag{3.5}$$

where $P(\theta)$ is a prior over the parameters of the model. The marginal likelihood is a key quantity for choosing different models in a Bayesian model selection task. Model selection is a necessary step in understanding and representing the observed data. However, the marginal likelihood $P(\mathcal{Y})$ is an intractable quantity to compute the models of interest. Traditionally, the marginal likelihood has been approximated either using analytical methods, e.g., the Laplace approximation, information measurements, variational free energy [57], or via sampling-based approach such as Markov chain Monte Carlo [273]. The techniques emphasize the same underlying principle of maximum a posteriori (MAP) from different approximations of interest.

3.2.3 Parameter Estimation Using the EM Algorithm

The general choice for achieving ML or MAP estimates of the mixture parameters is the classical expectation-maximization (EM) algorithm [60], [77]. For the case of single component density, the maximum likelihood method is easy to find the convergent parameters based on a wide range of component densities in a closed-form. However, for the case of Gaussian mixtures, the estimation becomes more intractable, since the log-likelihood as a function of the parameters may have many local maximum.

The EM algorithm [60] estimates the parameters at the local maximum of the log-likelihood function gives some initial parameter values. Some advantages of the EM algorithm over other methods are (a) no parameters needed for the iterative optimization process, (b) its simplicity and robustness. The shortage is similar to most other deterministic methods, i.e., the solution is based on initial parameter values. Such sensitivity can be avoided or partially solved by either (a) running the performance using different initial values and finding the best one. or (b) using some measure criteria to find the best fit number of components and models in deterministic methods (e.g., BIC, MDL [203], MML [77]) or in stochastic and resamping methods (e.g., Markov chain Monte Carlo (MCMC) based model selection criteria [273], resampling based schemes [162] or cross-validation approaches [236]).

The EM algorithm is based on the interpretation of $\mathcal{Y}$ as *incomplete* data. EM is an iterative procedure to find local maxima of $\log p(\mathcal{Y}|\theta)$ or $[\log p(\mathcal{Y}|\theta) + \log p(\theta)]$. In the case of finite mixtures, the missing part is a set of n labels $\mathcal{Z} = \{\mathbf{z}^{(1)}, ..., \mathbf{z}^{(n)}\}$ associated with the n samples, indicating which component produced each sample. Each label is a binary vector $\mathbf{z}^{(i)} = [z_1^{(i)}, ..., z_k^{(i)}]$, where $z_m^{(i)} = 1$ and $z_p^{(i)} = 0$, for $p \neq m$, means that sample $\mathbf{y}^{(i)}$ was produced by the m-th component. The complete log-likelihood (i.e., the one from which we could estimate θ if the complete data $\mathcal{X} = \{\mathcal{Y}, \mathcal{Z}\}$ was used in [77] is

$$\log p(\mathcal{Y}, \mathcal{Z} \,|\theta\,) = \sum_{i=1}^{n} \sum_{m=1}^{k} z_m^{(i)} \log \left[\alpha_m p\left(\mathbf{y}^{(i)} \,|\theta_m\right)\right] \tag{3.6}$$

The EM algorithm produces a sequence of estimates $\{\hat{\theta}(t),\ t = 0, 1, 2, 3, ...\}$ by alternatingly applying two steps (until some convergence criterion is met):

E-Step: Computes the conditional expectation of the complete log-likelihood in Eq. 3.6, given $\mathcal{Y}$ and the current estimate $\hat{\theta}(t)$. Since $\log p(\mathcal{Y}, \mathcal{Z}|\theta)$ is linear with respect to the missing $\mathcal{Z}$, we simply have to compute the conditional expectation $\mathcal{W} \equiv E[\mathcal{Z}|\mathcal{Y}, \hat{\theta}(t)]$, and plug it into $\log p(\mathcal{Y}, \mathcal{Z}|\theta)$. The result is the so-called Q-function:

$$Q(\theta, \hat{\theta}(t)) \equiv E[\log p(\mathcal{Y}, \mathcal{Z}|\theta)|\mathcal{Y}, \hat{\theta}(t)] = \log p(\mathcal{Y}, \mathcal{W}|\theta) \tag{3.7}$$

Since the elements of $\mathcal{Z}$ are binary, their conditional expectations (i.e., the elements of $\mathcal{W}$) are given by

$$w_m^{(i)} \equiv E\left[z_m^{(i)} \left| \mathbf{y}, \hat{\theta}\left(t\right)\right.\right] = \Pr\left[z_m^{(i)} = 1 \left| \mathbf{y}, \hat{\theta}\left(t\right)\right.\right] = \frac{\hat{\alpha}_m\left(t\right) p\left(\mathbf{y}^{(i)} \left| \hat{\theta}_m\left(t\right)\right.\right)}{\sum_{j=1}^{k} \hat{\alpha}_j\left(t\right) p\left(\mathbf{y}^{(i)} \left| \hat{\theta}_m\left(t\right)\right.\right)}, \tag{3.8}$$

where the last equality is simply Bayesian law (notice that α_m is the α *priori* probability that $z_m^{(i)} = 1$, while $w_m^{(i)} = 1$ is the a posteriori probability that $z_m^{(i)} = 1$, after observing $\mathbf{y}^{(i)}$ for any (i).

M-Step: Updates the parameter estimates according to

$$\hat{\theta}\left(t+1\right) = \arg\max_{\theta} \left\{Q\left(\theta, \hat{\theta}\left(t\right)\right) + \log p\left(\theta\right)\right\} \tag{3.9}$$

in the case of MAP estimation, or

$$\hat{\theta}\left(t+1\right) = \arg\max_{\theta} Q\left(\theta, \hat{\theta}\left(t\right)\right). \tag{3.10}$$

3.3 Measure Criteria for Model Selection

It is old adage that more descriptions of an object, or more proofs of a statement, are better than one. Sometimes, this becomes true. For example, in object boundary detection (which is conceptually different from edge detection), different information of texture, color, intensity, brightness, and shape are combined to find the reasonable object boundary based on human recognition concepts [159]. However, this is certainly not true if a description is redundant in the sense that some parts already give complete description. It is such a redundancy that we must first eliminate [204]. Model selection criteria is the reasonable rule to reduce such redundancy and keep optimal information. Moreover, model selection methods can help identify useful models, in the sense of predictive accuracy or generalization. Models should be evaluated based on generalization ability, not on goodness of fit. The main principle of model selection is not select the best-fitting model but shall select the best-predicting model. There are also some other non-statistical but very important selection criteria such as plausibility, interpretability, explanatory adequacy, falsifiability.

For statistic learning based model selection, a model is defined as a collection of probability distributions, indexed by model parameters $M = \{f(\mathcal{Y}|\theta)|\theta \in \Omega\}$ forming a Riemannian manifold, embedded in the space of probability distribution. Akaike information criterion (AIC) [3] derived as asymptotic approximation of Kullback-Liebler information distance between the model of interest and the truth within a set of proposed models.

In the following, we review some of the existing methods for approximating marginal likelihoods. Then we present our proposed methods for statistic model selection based nonparametric blur identification. These methods are analytic approximations such as the Laplace method [124], the Bayesian Information Criterion (BIC) [223], the Minimum Description Length (MDL) [203] and the Minimum Message Length (MML) [253]. All these methods make use of the MAP estimate is usually straightforward procedure.

3.3.1 Entropy and Information Measure

The original information theory is developed based on data compression and data transmission. With the evolution of statistics, researchers found that the information theory handles a lot of underlying disciplines of the physical world. However, this broad realm of information theory with numerous topics all root in two basic concepts: entropy and mutual information, which are functions of probability distribution that underlie the process of communication.

According to Mackay [154], the terminology of *entropy* measures the uncertainty of a *random variable*, which can be considered as the information embedded in the variable. In mathematics, the entropy of a discrete random variable X, denoted by $S(X)$, is defined as the expected value of negative-logarithm of probabilities, shown in Table. 3.1. According to the definition, we can derive the nonnegativity property of entropy $S(X) \geq 0$. The reason is $\log p(x) \leq 0$ for any discrete variables, $\forall x \in \mathcal{X}, 0 \leq p(x) \leq 1$, where $\mathcal{X}$ is the set of all possible values for X. Therefore, $S(X) = E_p(-\log p(x)) \geq 0$.

The definition of entropy can be extended to the case of multiple variables, e.g., *joint entropy* and *conditional entropy*. The joint entropy of a pair of discrete random variables X and Y, denoted by $S(X, Y)$. The conditional entropy of a random variable Y given another variable X, denoted by $S(Y|X)$, is to measure the uncertainty of Y when X is known. The definitions are shown in Table. 3.1.

Table 3.1: List of Definitions of Entropy and Mutual Information for Discrete Random Variables

Types of Entropy	Definition
Entropy:	$S(X) = E_p(-\log p(x)) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$, (single variable)
Joint entropy:	$S(X,Y) = E_p(-\log p(x,y)) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$
Conditional entropy:	$S(Y\|X) = \sum_{x\in\mathcal{X}} p(x)S(Y\|X=x) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y\|x)$
Relative entropy :	$S_{KL}(p(x),q(x)) = E_p \log\frac{p(x)}{q(x)} = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$
Mutual information:	$M(Y;X) = S(X) - S(Y\|X) = S(Y) - S(X\|Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x)p(y)}{p(x,y)}$

The relative entropy, also called *Kullback-Leibler (KL) divergence* is a measure of the distance between two distributions. The divergence is a function of two probability mass functions $p(x)$ and $q(x)$ that potentially characterize the same random variable X, shown in Table. 3.1. The relative entropy is nonnegativity. The proof can be based on the convexity of the logarithm function. For any given probability mass functions $p(x)$ and $q(x)$, we have $S_{KL}(p(x), q(x)) \geq 0$ and the equality establishes if and only if $p(x) = q(x)$. The convention is used that $0 \log \frac{0}{q(x)} = 0$ and $p(x) \log \frac{p(x)}{0} \to \infty$. The KL-divergence is nothing else but Shannon's measure of uncertainty for a random variable X, if $q(x)$ is a uniform probability distribution. Thus, Shannon's entropy can be interpreted as the amount of information in a model $q(x)$ of X compared to the maximum incertitude model - the uniform distribution. The uniform distribution is the one with maximum entropy.

The concept *mutual information* is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other. Consider two random variables X and Y is given by the uncertainty reduction for Y when X is known. Likewise, the information about X contained Y is given by the uncertainty reduction for X when Y is known. Mutual information is defined by $M(Y;X) = S(X) - S(Y|X) = S(Y) - S(X|Y)$ and these two definitions are equivalent. The definition of mutual information is shown in Table. 3.1.

While combining the two basic concepts of entropy and mutual information, we can also make an extension to *continuous random variables* for the definition of differential entropy which replaces the notation of "sum" by "integral".

3.3.2 Laplace's Method

We infers Bayesian information criteria (BIC) from Laplace approximation, and we can easily understand some related information measure criteria such as AIC, MDL and MML. Based on the Bayesian rule, the posterior over parameters θ of a model is

$$P(\theta|\mathcal{Y}, m) = \frac{p(\mathcal{Y}|\theta, m)P(\theta|m)}{p(\mathcal{Y}|m)} \tag{3.11}$$

The logarithm of the numerator is defined in the following,

$$t(\theta) = \log[p(\mathcal{Y}|\theta, m)P(\theta|m)] = \log P(\theta|m) + \sum_{i=1}^{n} \log p(\mathbf{y}^{(i)}|\theta, m) \tag{3.12}$$

the Laplace approximation [124] makes a local Gaussian approximation around a MAP parameter estimate $\hat{\theta}$ in Eq. 3.2. The validity of this approximation is based on the large data limit and regularity constraints. The $t(\theta)$ is expanded to second order as a Taylor series at this point,

$$t(\theta) = t(\hat{\theta}) + (\theta - \hat{\theta})^\top \frac{\partial t(\theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} + \frac{1}{2!}(\theta - \hat{\theta})^\top \frac{\partial^2 t(\theta)}{\partial \theta \partial \theta^\top}\bigg|_{\theta=\hat{\theta}} + ... \tag{3.13}$$

$$\approx t(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^\top H(\hat{\theta})(\theta - \hat{\theta}) \tag{3.14}$$

where $H(\hat{\theta})$ is the Hessian of the log posterior evaluated at $\hat{\theta}$, it is a matrix of the second derivatives of Eq. 3.12,

$$H(\hat{\theta}) = \left.\frac{\partial^2 \log p(\theta|y,m)}{\partial\theta\partial\theta^\top}\right|_{\theta=\hat{\theta}} = \left.\frac{\partial^2 t(\theta)}{\partial\theta\partial\theta^\top}\right|_{\theta=\hat{\theta}} \tag{3.15}$$

where the linear term has vanished as the gradient of the posterior $\frac{\partial t(\theta)}{\partial\theta}$ at $\hat{\theta}$ is zero because it is the MAP setting or a local maximum. Substituting Eq. 3.14 into the log marginal likelihood in Eq. 3.5 and integrating yields,

$$\begin{aligned}
\log p(\mathcal{Y}|m) &= \log \int d\theta P(\theta|m)p(y|\theta,m) = \log \int d\theta \exp[t(\theta)] && (3.16)\\
&\approx t(\hat{\theta}) + \frac{1}{2}\log\left|2\pi H^{-1}\right| && (3.17)\\
&= \log P(\hat{\theta}|m) + \log p(\mathcal{Y}|\hat{\theta},m) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log|H| && (3.18)
\end{aligned}$$

where d is the dimensionality of the parameter space. $|H|$ denotes the determinant value of H. Thus, the Eq. 3.18 can be written,

$$p(\mathcal{Y}|m)_{Laplace} = P(\hat{\theta}|m)p(\mathcal{Y}|\hat{\theta},m)\left|2\pi H^{-1}\right|^{1/2} \tag{3.19}$$

where the Laplace approximation to the marginal likelihood consists of a term for the data likelihood at the MAP setting, a penalty term from the prior, and a volume term calculated from the local curvature. However, this approximation has several shortcomings in that the second derivatives of approximation are intractable to compute.

Bayesian model selection method preference for simpler models is a spin-off and built-in Ockham's razor. Bayesian model selection methods include Bayesian factor and Bayesian information criteria (BIC) are also used for model selection and parameter estimation [255]. BIC is considered as an approximation of Bayesian factor [201]. It is based on a large sample approximation of the marginal likelihood yielding the easily-computable BIC.

3.3.3 BIC and MDL

AIC derived as asymptotic approximation of Kullback-Liebler information distance between the model of interest and the truth. AIC and BIC have similar prediction mechanism. They estimate a generalized model which has the ability to fit all "future" data samples from the same underlying process, not just the current data sample.

Because of the intrinsic simplicity, Akaike information criterion AIC, shown in Table. 3.2, and MDL [203] are widely applied to estimating two terms: a maximization of the likelihood data term and a penalty term of the complexity of the model. However, within a Bayesian framework, model selection appears more complex as it involves the evaluation of Bayes factors [124]. These Bayes factors require the computation of high-dimensional integrals with no closed-form analytical expression. These computational problems have restricted the use of Bayesian model selection, except for the cases for which asymptotic expansions of the Bayes factor are valid [13], [23].

The Bayesian Information Criterion (BIC) [223] like AIC, is applicable in settings where the fitting is carried out by maximization of a log-likelihood. The BIC can be obtained from the Laplace approximation by retaining only those terms that grow with n. From Eq. 3.18, we have

$$\log p(\mathcal{Y}|m)_{Laplace} = \log P(\hat{\theta}|m) + \log p(\mathcal{Y}|\hat{\theta}, m) + \frac{d}{2}\log 2\pi - \frac{1}{2}\log |H| \tag{3.20}$$

where each term dependence on n has been annotated. Here we use the "big-O" notation to see the probability distribution of each term. Retaining $\mathcal{O}(n)$ and $\mathcal{O}(\log n)$ terms yields,

$$\log p(\mathcal{Y}|m)_{Laplace} = \log p(\mathcal{Y}|\hat{\theta}, m) - \frac{1}{2}\log |H| \tag{3.21}$$

From Eq. 3.12 and Eq. 3.15, we know that the Hessian scale linearly with n, we have,

$$\lim_{n\to\infty} \frac{1}{2}\log |H| = \frac{1}{2}\log |nH_0| = \frac{d}{2}\log n + \frac{1}{2}|H_0| \tag{3.22}$$

and then assuming that the prior is non-zero at $\hat{\theta}$, thus the Eq. 3.21 in the limit of large n becomes the BIC score,

$$\log p(y|m)_{BIC} = \log p(y|\hat{\theta}, m) - \frac{d}{2}\log |n| \tag{3.23}$$

There are two main advantages of BIC. Firstly, it does not depend on the prior $p(\theta|m)$. Secondly, it does not take into account the local geometry of the parameter space. Therefore, it is invariant to update parameters of the model. In practice, the utilized dimension of the model d is equal to the number of well-determined parameters when any potential parameter degeneracies have been removed.

The *Minimum Description Length* (MDL) principle [203] informally states that the best model is the one which minimizes the sum of two terms: first, the length of the model, and second, the length of the data when encoded using the model as a predictor for the data. In other words, the MDL criterion is utilized for resolving the tradeoff between model complexity (each retained coefficient increases the number of model parameters) and goodness-of-fit (each truncated coefficient decreases the fit between the received - i.e., noisy - signal and its reconstruction). We seek the data representation that results in the shortest encoding of both observations and constraints.

On the other hand, the BIC is in fact exactly minus the minimum description length (MDL) penalty used in Rissanen [203], [204] shown in Table. 3.2. The minimum description length method (MDL) [203] is an algorithmic coding theory and regularities (redundancy) can be used to compress the data. The main principle of MDL is to achieve the best model that provides the shortest description length of the data in bits by "compressing" the data as tightly as possible. It suggests a means of evaluating this representation system such as the representation of the data item using the model, and the mismatch between this representation and the actual data. Recently, the minimum message length (MML) framework of Wallace and Freeman [253], Lanterman [137], Figueiredo and Jain [77] has been intensively applied for unsupervised model selection techniques [77] which is closely related to Bayesian integration over parameters.

3.4 Nonparametric Model Selection

We follow a Bayesian framework whereby the unknown blur kernels, including blur type, blur parameters, and the noise variance are regarded as random quantities with general open prior knowledge from a designed blur solution space. Several previous works on Bayesian parameter estimation and model selection for such ill-posed inverse problems have been addressed in a series of papers.

As described Roweis and Ghahramani [211], factor analysis, principle analysis (PCA, mixture of Gaussian clusters, vector quantization (VQ), Kalman filter models and hidden Markov models can be unified as variations of unsupervised learning a single basic generative model.

3.4.1 Gaussian Mixture Model

Gaussian densities are probably the most commonly used densities to model continuous valued data. The first reason for this popularity is that maximum likelihood parameter estimation can be done in closed form and only requires computation of the data mean and covariance. The second reason is that of all densities with a particular variance, the Gaussian density has the largest entropy and therefore is the most "vague" density in this sense. This last property motivates the use of the Gaussian as a default density when there are no reasons to assume that some other parametric density is more appropriate to model the data at hand.

A Gaussian density in a D-dimensional space, characterized by its mean $\mu \in \mathbb{R}^D$ and $D \times D$ covariance matrix $\sum$, is defined as

$$\mathcal{N}(x;\theta) = (2\pi)^{-D/2} |\sum|^{-1/2} \exp[-\frac{1}{2}(x-\mu)^\top \sum^{-1} (x-\mu)] \tag{3.24}$$

where θ denotes the parameters μ and $\sum$ and $|\sum|$ denotes the determinant of $\sum$. In order to be a proper density, it is necessary and sufficient that the covariance matrix be positive definite. Throughout this thesis we implicitly assume that the likelihood is bounded, e.g. by restricting the parameter space such that the determinant of the covariance matrices is bounded, and hence the maximum likelihood estimator is known to exist [148]. Alternatively, the imperfect precision of each measurement can be taken into account by treating each data point as a Gaussian density centered on the data point and with small but non-zero variance. We then maximize the expected log-likelihood, which is bounded by construction.

3.4.2 K-Means Clustering as a Hard Gaussian Mixture Model

As a consequence, the Gaussian mixture model is often referred to as a "soft" clustering method, while K-means is "hard". The reason is that the number of clusters should be given manually,

Table 3.2: List of information theoretic model selection techniques

Criteria	Explanation
Akaike Information Criteria (AIC):	$AIC = -2\log f(y\|\hat{\theta}) + 2d$
Bayes Information Criteria (BIC):	$BIC = -2\log f(y\|\hat{\theta}) + d\log n$
Minimum Description Length (MDL):	$MDL = -2\log f(y\|\hat{\theta}) + 2d\log n$

while optimal Gaussian mixture model classifier can automatically find the number of clusters. Similarly, when Gaussian mixture models are used to represent the feature density in each class, it generates smooth posterior probabilities $\hat{P}(x) = \{\hat{p}_1(x), ..., \hat{p}_k(x)\}$ for classifying x. Although it is often interpreted as a soft classification, the classification rule is to achieve $\arg\max_k \hat{P}_k(x)$.

From the experiments, the K-means classifiers are memory-based, and require no model to be fit. Given a query point x_0, we find the k training points $x_{(r)}, r = 1, ..., k$ closest in distance to x_0, and then classify using majority vote among the k neighbors. For simplicity, we can assume that the features using Euclidean distance in feature space,

$$d(i) = \|x_{(i)} - x_{(0)}\|$$

Since the distance measures appropriate for qualitative and ordinal features, and how to combine them for mixed test data. Despite its simplicity, K-means classifier has been successful in a large number of classification problems. It is often successful where each class has many possible prototypes, and the decision boundary is very irregular.

Furthermore, the Gaussian mixture model can be thought of as a prototype method as similar as the spirit to K-means and Learning vector quantization (LVQ). The K-means clustering algorithm is a deterministic method that does not depend on initial parameter values and employs the K-means algorithm as a local search procedure. Instead of randomly selecting initial values for all cluster centers as is the case with most global clustering algorithms, the proposed technique proceeds in an incremental way attempting to optimally add one new cluster center at each stage.

3.4.3 From K-Means Clustering to Vector Quantization

The K-means clustering algorithm represents a key tool in the apparently unrelated area of image and signal compression, particularly in *vector quantization* (VQ) [87], [145], [187]. A K-means clustering (known as Lloyd's algorithm) runs in this space.

In data compression, vector quantization (VQ) is a quantization technique often used in lossy data compression in which the basic idea is to code or replace with a key, values from a multidimensional vector space into values from a discrete subspace of lower dimension. The lower-space vector requires less storage space and the data is thus compressed. The transformation into the subspace is usually achieved through projection, or by using a codebook. In some cases, a codebook implementation can be also used to entropy code the discrete value in the same step by generating a prefix coded variable-length encoded value as its output. In cryptography, a codebook is a document used for implementing a code. A codebook contains a lookup table for encoding and decoding; each word or phrase has (one or more) strings which replace it. To decipher messages written in code, corresponding copies of the codebook must be available at either end. The distribution and physical security of codebooks presents a special difficulty in the use of codes, compared to the secret information used in ciphers, the key, which is typically much shorter.

Design of VQ

In the earlier days, the design of a vector quantizer (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration. In 1980, Linde, Buzo, and Gray

Figure 3.1: (a) Distinct image. (b) Blurred image. (c) Band-pass filtered blurred image (bandpass filter is used for the selection to structure at different spatial scales).

(LBG)[145] proposed a VQ design algorithm based on a training sequence. The use of a training sequence bypasses the need for multi-dimensional integration. A VQ that is designed using this algorithm are referred to in the literature as an LBG-VQ.

Why we use VQ to work at all? The first reason is that it is difficult to find some descriptive information on blurred images in the spatial domain. However, VQ can find the difference between blurred images via different encoding error. The second reason is that for typical everyday images like photographs, many of the block look the same. In this case there are many almost pure white blocks, and similarly pure gray blocks of various shades. These require only one block each to represent them, and then multiple pointers to that block. It is one kind of lossy compression, but the VQ-based codebook is a good measurable "feature space" for vision work.

The k-dimensional, N-level vector quantizer is defined as a mapping from a k-dimensional Euclidean space R^k into a certain finite set $C = C_1, C_2, ..., C_N$. The subset C is called a codebook and its elements C_i are called codewords. The VQ encoding is to search and assign one codeword to the input test vector with minimum distortion. Given an image with $N_w \times N_h = M$ block, each block has k ($k = w \times h$ dimensions; for color image, $3k$ dimensions) pixels. For each codeword $C_t = c_{t1}, c_{t2}, ..., c_{tN}$, and the testing vector $X = (x_1, x_2, ..., x_M)$, the squared Euclidean distortion can be expressed as: $D(X, C_t) = sum(||x_i - c_{ti}||)^2, i = 1, 2, ..., k$. From this equation, we know that encoding each input vector requires N distortion computations and $N - 1$ comparisons. Therefore, the computational complexity of encoding each input vector includes KN multiplication, $(2k - 1)N$ additions and $N - 1$ comparisons.

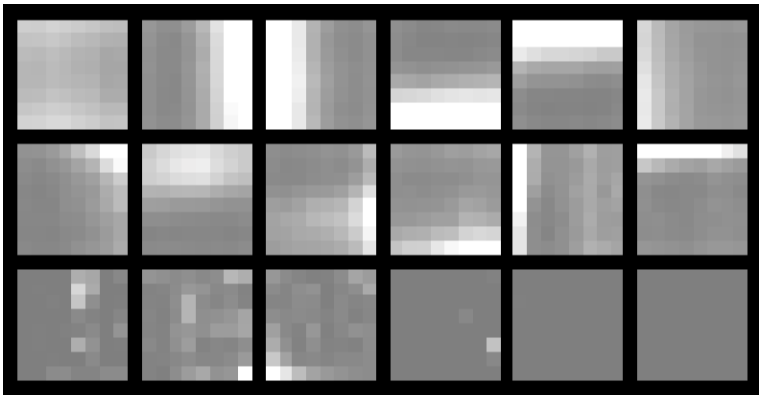


Figure 3.2: A VQ-codebook with 64 pixels per block, 18 block, SNR=24.30dB, representative vectors consists of various edges of different directions, amplitudes, and frequency.

In order to perform the closest vector search efficiently for building a larger size codebook. We adopted the approach for fast codeword search algorithm developed by Ra et al. [200] which is about 15 times quicker than "full search" algorithm for each codebook size, e.g., size $=(128, 256, 512, \cdots)$. This equal-average nearest neighbor search (ENNS) algorithm uses the mean value of an input vector to reject impossible codewords. It also reduces a great deal off-line computational time compared with other fast search algorithms with only N additional memory. In the proposed algorithm, the band passed pixel value in the codebook is treated as the label. VQ is subsequently applied to all vectors with the same label based on the LBG [145] algorithm. The VQ can be generated in a hierarchical way. The ENNS algorithm adapted as a kernel for VQ encoding by the proposed algorithm is briefly described,

1. Let $X = (x_1, x_2, ..., x_k)$ be a k-dimensional vector, the sum of k-dimensional vector X as $S_X = sum(x_i), i = 1, 2, , k$.
2. Assuming the current distortion D_{min}, the main sprit of ENNS can be stated as: If $(S_X - S_{C_j})^2 \geq k \cdot D_{min}$, then $D(X, C_j) \geq D_{min}$. This means C_j will not be the nearest neighbor to X, if $(S_X - S_{C_j})^2 \geq k \cdot D_{min}$ satisfied.
3. The sum of each codeword is calculated and these values are sorted in ascending order. The squared Euclidean distortion D_{min} between the input vector and this tentative matching codeword is calculated. Then the codewords C_j for which $S_X \geq S_{C_j} + (k \cdot D_{min})^{\frac{1}{2}}$ or $S_X \leq S_{C_j} - (k \cdot D_{min})^{\frac{1}{2}}$ are eliminated.
4. The search is performed up and down, left and right directions iteratively till the nearest codeword is found.

To apply VQ method for blur identification, blurred images are vector quantized in terms of the enhancement of blur representation. There are many potential features which can be used to represent the largest blur in an image. We use local non-flat region features to train the codebook so that a lot of redundancy in homogenous image regions can be avoided. Fig. 3.2 shows results of a blurred frame with representative vectors.

3.5 Experimental Results

3.5.1 Vector Quantization for Nonparametric Blur Identification

Main Steps

To identify the blurred frames in a real-life video sequence, we need to find an efficient method to classify and group different blurred images in a given video sequence. The vector quantization, codebook and its related encoding error are a basis of blur identification and blur degraded frame selection system as illustrated in Fig. 3.3. This approach combines blur identification and searching blur degraded images in a large video data in nonparametric vector quantizer codebooks. The vector quantization (VQ) based codebook method satisfies such demand and the detailed implementation has been presented by Zheng and Hellwich [295].

The method can be described in the following.

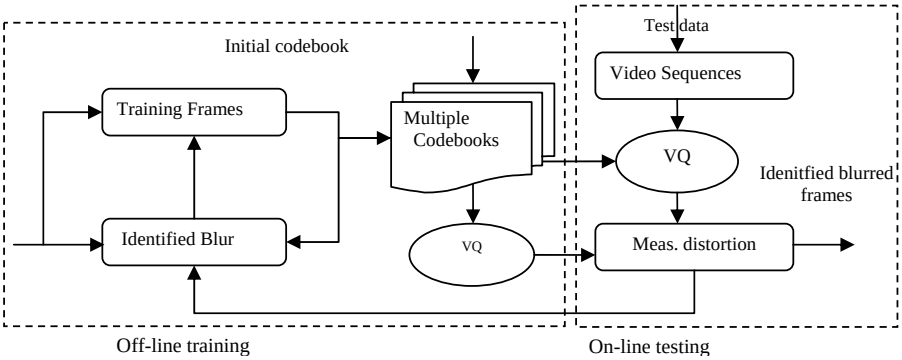


Figure 3.3: Diagram of blur identification and find blurred images in large video data

1. **Off-line training.** To apply VQ method for blur identification, blurred images are vector quantized in terms of the enhancement of blur representation. There are many potential features which can be used to represent the largest blur in an image. We use local non-flat region features to train the codebook so that a lot of redundancy in homogenous image regions can be avoided. In a consequence, blur is identified from a few dominant candidate blur functions in a set of training images. Each of the training sets with their related blur functions is used to train the codebook-based on LBG algorithm [87], [145], [187]. These trained codebooks can thus be used to measure the similarity of other blurred images. Fig. 3.2 shows results of a blurred frame with representative vectors.

2. **On-line testing (measuring).** After the off-line training period, on-line blur identification can be processed. Fast VQ encoding method speeds up the on-line blur identification in video sequence. Each frame will be checked by a trained codebook via VQ encoding approach. The distortion between the trained codebook and testing frames are measured by mean square error (MSE). The values of different distortion are used to classify the video frames into different blur clusters. VQ encoding of different frames get different mean square error distortions based on the similarity measurement of statistical intensity value. The testing frame with minimum distortion is identical blur in the frame which generated this codebook.

Experimental Results

In the first experiment, we have tested simulated images to demonstrate the accuracy of VQ-based blur identification and classification of blur degraded images. In Fig. 3.4, three groups of images with motion blur, Gaussian blur and mixed types of blur are tested in three different signal-to-noise ratios (SNR). The minimum VQ encoding distortion (MSE) of the testing image is identical with the trained codebook. The up-right diagram shows the motion blur identification where the codebook has a blur angle of 20 degree. The second curve diagram shows the Gaussian blur identification, codebook has a variance 1.5. The third curve diagram shows blur identification of mixed blur types, Gaussian variance $= \{1.5, 2.5, 3.5, 4.5, 5.5\}$ and motion blur with different blur angle. The codebook is generated by the image with index 3. The experiment also demonstrates that the approach is robust with respect to correlated noises.

The second experiment has been performed on real-life video sequences in Fig. 3.5. Firstly, one

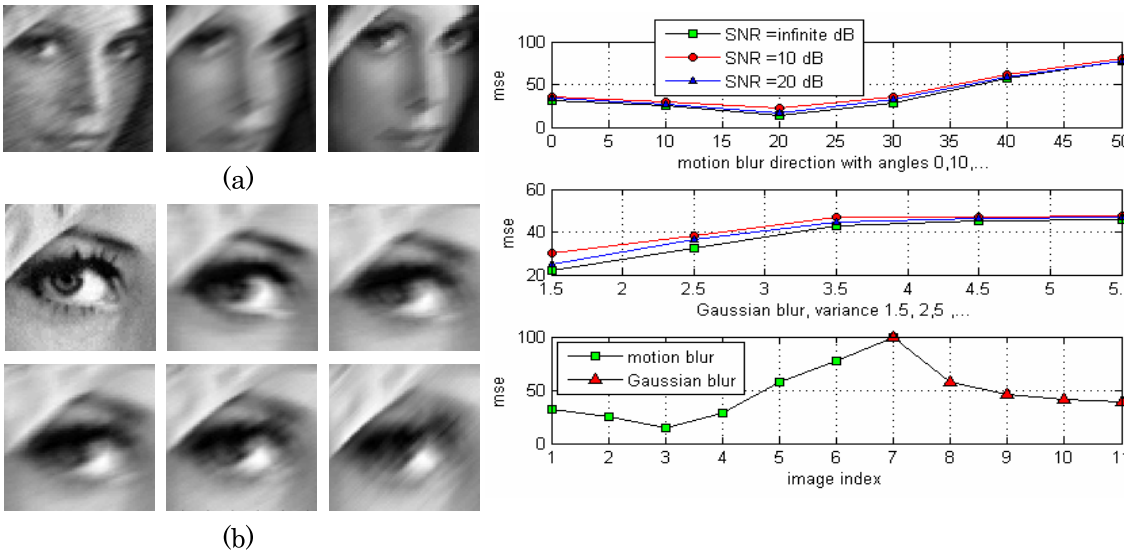


Figure 3.4: (a) Three images with 10dB, 20dB and ∞dB respectively. (b) An unblurred image with five blurred images. Right diagrams: The minimum MSE is blur identified.

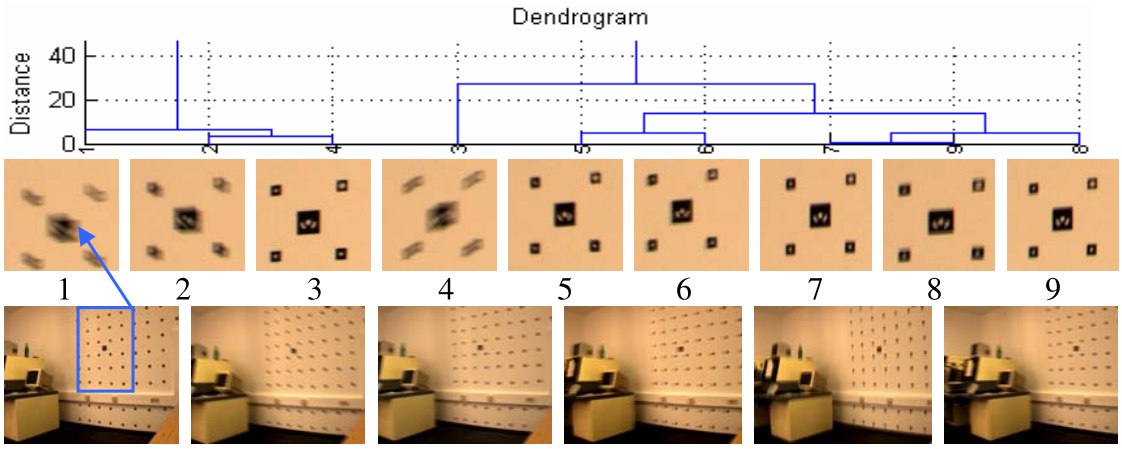
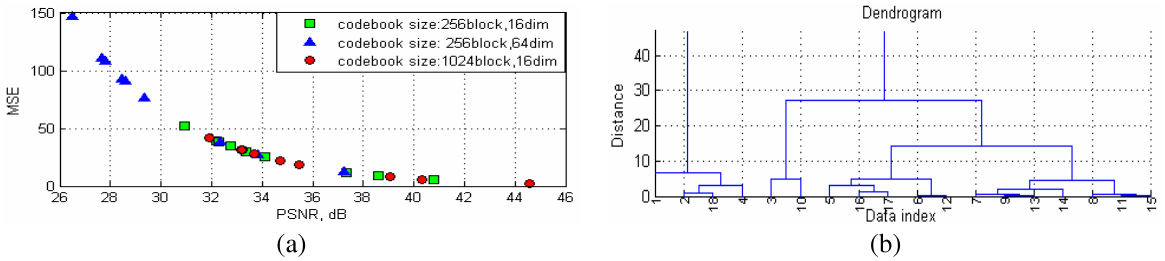


Figure 3.5: Blur Identification of frames in dendrogram (taken by "ptgrey" video camera, 15f/s). The abscissa is an index for 9 frames (index 012-020 from 201 frames), the ordinate denotes the encoding distortion values).

blurred frame is blur identified based on Bayesian MAP estimation in the off-line period. VQ-based codebook of this blur identified frame is used to check other unknown frames in the on-line period. We present the checking results in a clustering tree to demonstrate its efficiency. The results are visualized by a dendrogram clustering method based on the VQ encoding distortion. From the dendrogram, we can easily find the frames with different blur status are classified into different sub-tree. The blur frames are classified into two main classes. The first main class of images with index of $\{1, 2, 4\}$ are relatively stronger blurred. The second class of images with index of $\{3, 5, 6, 7, 8, 9\}$ are relatively weak-blurred or without blur.

The images with index of $\{2, 4\}$, $\{5, 6\}$ and $\{7, 9\}$ have most similar blur status. The PSF of images with index of 2 and 4 can be easily predicted in the cluster of $\{1, 2, 4\}$ because the image



(a) (b)

Figure 3.6: (a) PSNR-MSE distribution of different size of codebooks. (b) The dendragram of 18 frames (index: 012-029)

with index of 1 is trained as a codebook. If we continue identify more blurred frames precisely, we can continue the on-line process and add more codebooks. The Bayesian MAP estimation for more video frames uses the prior knowledge from the blur identified codebook and classified datasets. Higher accuracy PSF estimation follows the direction to a child on the sub-tree.

In Fig. 3.6(a), the influence of blur identification is also evaluated by checking the size of codebooks. For this case, codebooks with 256 blocks and 64 dimensions per block get encoding distortion in a large range. Large encoding distortions cause distinct classification. The size of codebook is selected based on such criteria. The PSNR-MSE diagram is drawn by measuring the relationship between the image degradation and VQ encoding MSE. The image degradation is quantified by peak-to-noise ratio (PSNR):

$$PSNR = 10\log_{10}(255^2/MSE)(dB) \tag{3.25}$$

In Fig. 3.6(b), we perform the algorithm on more images. 18 frames with continuous indices are classified. The dendrogram in Fig. 3.6(b) has a similar sub-tree structure to the dendrogram in Fig. 3.5. The blurred frames are classified and added in each sub-tree.

Compared to the existing methods, the approach can efficiently find out blurred images in different groups for given video sequences. Mechanisms with both off-line and on-line phases make the on-line performance in real-time. The approach is confirmed more practical in different video acquisition environments.

3.6 Conclusion

As we know, finding efficient descriptive features of test data is crucial for classification, categorization or recognition tasks. However, it is very hard to find descriptive information directly from blurred images. An indirect way is to find generative features for blur identification and classify blurred images. Vector quantization can be considered as a nonparametric classification method to measure the similarity and difference between images. One of the most useful advantages of this approach is its real-time performance and its robust with respect to noise. For example, when large data sets (e.g., large image or video sequences) are available, nonparametric blur identification techniques become crucial in that these methods can classify the blurred and unblurred images efficiently without thinking about the detailed parameters of blur kernels.

However, VQ can not be directly used as a blind image restoration method for a single blurred image. Therefore, more specific blur identification and image restoration method should be

addressed. Furthermore, such nonparametric methods have some difficulties to estimate bandwidth of probability densities or accurate local parameters of blur kernels, we need to combine parametric methods to solve such difficulties.

In the next chapter, we focus on the blind image restoration for a single blurred image. A new method of blind deconvolution using Bayesian MAP estimation in alternating minimization procedure is adopted and extended for blind image restoration. It also integrates the parametric information of the blur structures progressively throughout restoration.

4 Double Regularized Bayesian Estimation for Parametric Blur Identification

From where we stand, the rain seems random. If we could stand somewhere else, we would see the order in it. - T. Hillerman (1990) Coyote Waits., Harper-Collins

In this chapter, we propose a new method which combines global nonparametric model selection methods and local parametric optimization for parametric blur identification. The integration is processed in a weighted double regularized Bayesian learning approach. A proposed prior solution space includes dominant blur point spread functions as prior candidates for Bayesian MAP estimation. The double cost functions are adjusted in an alternating minimization approach which successfully computes the convergence for a number of parameters. The discussions of choosing regularization parameters for both image and blur functions are also presented. The algorithm is robust in that it can handle images that are formed in variational environments with different types of blur. Numerical tests show that the proposed algorithm works effectively and efficiently in practical applications.

4.1 Introduction

In two decades, there has been considerable interest in the regularization theory for blind image deconvolution (BID). As we know, the regularization method is originally proposed by Tikhonov [241], Miller [165] et al. which replaces an ill-posed problem by a well-posed problem with an acceptable approximation to the solution. Later, Katsaggelos et al. [126] have introduced an iterative regularization algorithm for image restoration based on a set theoretic approach. This algorithm using a deterministic framework introduces *a priori* knowledge in the form of convex sets, and decouples the nonlinear observation model into double linear observation models that are easy to solve. A projection-based method with conjugate-gradient minimization for BID has been proposed and extended by [136], [278], [280]. These methods have demonstrated how the parametric models in image restoration methods are used [134], [279] in some respects. However, these results are observed in underutilization of prior information. The ill-posed image restoration problem needs more effective and descriptive prior information or constraints to yield a unique solution to the corresponding optimization problem. Even if a unique solution exists, a proper initialization value is still intractable, e.g., the cost function is non-convex.

The Bayesian estimation provides a structured way to include prior knowledge concerning the quantities to be estimated [63], [85]. The Bayesian approach is, in fact, the framework in which the most recent restoration methods have been introduced. When blur is present, different approaches have been proposed to find a maximum a posterior (MAP) estimate. Besag [26] has introduced the statistical analysis of dirty pictures. Geman and Geman [85] combine image restoration and segmentation simultaneously in discrete and stochastic Bayesian estimation. Hellwich [111] has developed an unsupervised edge and object extraction method for noisy syn-

thetic aperture radar (SAR) data in Markov random field based Bayesian estimation. Opper et al. has developed a Bayesian estimation based free energy functional for approximate inference [182], [183]. Blake et al. [33] propose the use of gradually non-convexity method, which can be extended to the blurring problem. Molina and Ripley [170] propose the use of a log-scale for the image model. Green [97] and Bouman et al. [35] use convex potentials in order to ensure uniqueness of the solution. Recently, an appreciable extension of the range of hyperparameter estimation methods is used in Bayesian estimation. Molina et al. [169] use a hierarchical Bayesian paradigm resulting from the set theoretic regularization for estimating hyper-parameters. They also report that the accuracy of the obtained statistic estimates for the PSF and the image could vary significantly, depending on the initialization. To obtain accurate restorations in the Bayesian approach, accurate prior knowledge of PSF or image must be available.

In the Bayesian estimation, some main properties can be mainly focused to improve the performance. Firstly, prior knowledge can be achieved based on physical constraints and implementation. However, the Bayesian estimation could be sensitive to wrong priors, but we can learn priors too. Secondly, it is an ideal and simple approach to model selection using some measure criteria. Finally, the conception of Bayesian estimation is simple but often computation is hard. Therefore, we interpret Bayesian estimation as a regularization based optimization functional.

In this chapter, a space-adaptive regularization method is integrated into a Bayesian learning approach for parametric blur identification. A newly introduced solution space of PSF priors supports accurate parametric PSF in the form of Bayesian MAP estimation. An integrated quadratic cost function subject to convex constraints is minimized in an alternating minimization within a specified range. These positivity constraints and strictly convex property ensure that the alternating minimization procedure converges globally. Regularization parameters and weight matrices are estimated with the help of some parameter estimation techniques as well as comparison of these methods.

4.2 Bayesian Estimation Based Double Regularization

We use Bayesian MAP estimation to utilize prior information for getting a convergent posterior. Following Bayesian paradigm, the true $f(x)$, the PSF $h(x)$ and observed $g(x)$ in $g = hf + \eta$ on,

$$P(f,h|g) = \frac{p(g|f,h)P(f,h)}{p(g)} \propto p(g|f,h)P(f,h) \tag{4.1}$$

Applying the Bayesian paradigm to the blind deconvolution problem, we try to get convergence values from Eq. (4.1) with respect to $f(x)$ and $h(x)$. This Bayesian MAP approach can also be seen as a regularization approach which combines optimization methods for minimizing two proposed cost functions in the image domain and the PSF domain, shown in Fig. 4.1. The cost function of the restored true image $f(x)$ from Eq. (4.1) is deducted as:

$$P(f|g,h) \propto p(g|f,h)P(f) \tag{4.2}$$

the cost function of the PSF h(x) in Eq. (2) is deducted as:

$$P(h|g,f) \propto p(g|f,h)P(h) \tag{4.3}$$

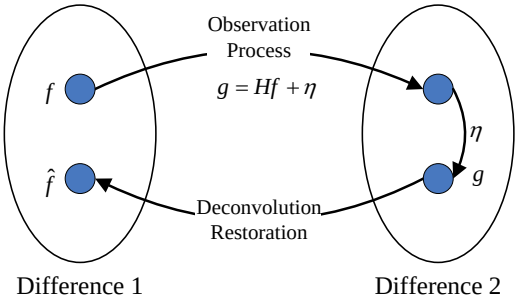


Figure 4.1: Formulation of blind image deconvolution problem into a double regularization approach. f is the unknown original image. H is the observed operator. g is the observed image. η is the noise. $\hat{f}$ is a restored image.

Some constraints are assumed for the application of these equations, e.g., the image pixel correlations are independent identically distributed (i.i.d.). However, manipulation of probability density functions in Bayesian estimation is difficult [63]. Several forms of the prior distribution, e.g., Gibbs distribution [303], smoothness prior or maximum entropy, pixel labeling prior [85], and some other pixel based prior (i, j are pixel neighbors) shown in Table. 4.1, have been suggested by researchers from different scientific disciplines. However, they are based on general knowledge about images. For most real blurred images, the prior knowledge is not descriptive models which can efficiently enhance the maximum a posterior (MAP) estimation. Therefore, we need to find representable information from potential functions which can characterize the observed blurred images. As we have discussed previously, in blurred images, power spectral densities vary considerably from low frequency domain in the uniform smoothing region to medium and high frequency domain in the discontinuity and texture regions. Moreover, most PSFs exist in the form of low-pass filters. The high frequency discontinuities are often diminished by vanishing blur multipliers [118]. Blur identification can be based on these characteristic properties. The proposed prior solution space supports parametric PSFs in a Bayesian MAP estimation, as PSFs of numerous real blurred images satisfy parametric blur kernels up to a certain degree.

4.2.1 Solution Space of Blur Kernel Priors

The proposed prior supports the parametric structure of PSFs in Bayesian estimation and reduces the effective search space. We define a set Θ as a solution space of Bayesian estimation which consists of primary parametric blur models as $\Theta = \{h_i(\theta), i = 1, 2, 3, ..., N\}$ presented in Fig. 4.2. $h_i(\theta)$ represents the ith parametric model of PSF with its defining parameters θ, and

Table 4.1: Prior distributions $p(\theta)$ in Bayesian Image Processing

Prior	Functionals	Explanation
Pixel priors:	$p(x) \propto \exp\{-\sum_{i\sim j} \phi(x_i - x_j)\}$, $x \in R^n$	pairwise differences
Gaussian prior:	$p(x) \propto \exp\{-\frac{1}{2k}\sum_{i\sim j}(x_i - x_j)^2\}$, $x \in R$	masked by Gaussian noise
Median prior:	$p(x) \propto \exp\{-\frac{1}{k}\sum_{i\sim j}\lvert x_i - x_j\rvert\}$, $x \in R$	attractive alternative to Gaussian
Labeling prior:	$p(x^L) \propto \exp\{-\beta \sum_{1\le k\le l\le c} n_{kl}\}$, $x^L \in C^n$	label at a coarse scale than pixels

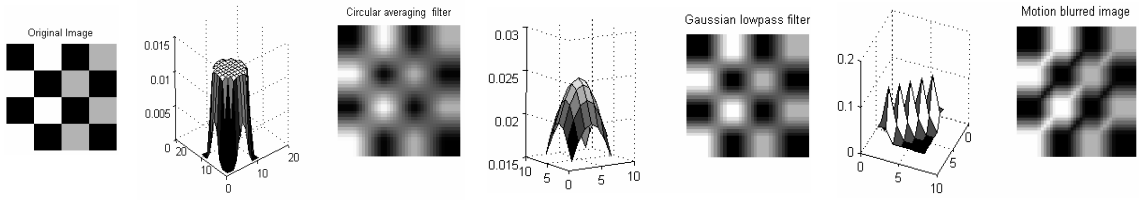


Figure 4.2: PSFs in the prior solution space. (a) Original synthetic image. (b) Pill-box PSF. (c) Gaussian PSF. (d) Linear motion PSF.

N is the number of blur kernel types.

$$h_i(\theta) = \begin{cases} h_1(\theta) \propto h(x,y;L_i,L_j) = 1/K, & \text{if } |i| \leq L_i \text{ and } |j| \leq L_j \\ h_2(\theta) \propto h(x,y) = K\exp(-\frac{x^2+y^2}{2\sigma^2}) & \\ h_3(\theta) \propto h\,(x,y,d,\phi) = 1/d, & \text{if } \sqrt{x^2+y^2} \leq D/2,\ \tan\phi = y/x \\ ..... & \end{cases} \tag{4.4}$$

$h_1(\theta)$ is a pill-box blur kernel with radius K. $h_2(\theta)$ is a Gaussian PSF characterized by its variance σ^2 and a normalization constant K. $h_3(\theta)$ is a simple linear motion blur PSF with a camera motion d and a motion angle ϕ. The other blur structures like out-of-focus and uniform 2D blur [19], [134] are also built in the solution space as *a priori* information. A set of parametric PSFs construct a predefined prior solution space for Bayesian MAP Estimation.

4.2.2 Weighted Space-Adaptive Regularization

To solve an adjustment optimization problem, the classical least squares based methods are mostly widely used, e.g., signal processing [112], image matching [207] and so on. However, the direct solution of the least squares problem is described in Eq. 4.5,

$$\sum\nolimits_{x\in\Omega} \left(h(x) * f(x) - g(x)\right)^2 = \min \tag{4.5}$$

Eq. (4.5) may leads to a vector $f(x)$ that is severely contaminated with noise. Tikhonov regularization [165], [241] can efficiently solve the ill-posed problem with additive noise as the following Eq. 4.6,

$$\frac{1}{2}\sum\nolimits_{x\in\Omega} \left(h(x) * f(x) - g(x)\right)^2 + \frac{1}{2}\lambda \sum\nolimits_{x\in\Omega} f(x)^2 = \min \tag{4.6}$$

The approach adds a penalty term multiplied by a regularization parameter λ for solving the linear least squares problem. However, for the image restoration, some ringing artifacts near sharp intensity transitions are still attributable to Tikhonov regularization. To reduce the ringing effects, Lagendijk et al. [135] made an extension of it by making use of the theory of the projections onto convex sets [126], [125] and the concepts of norms in a weighted Hilbert space. A weighted space-adaptive regularization equation seeks to minimize the following cost function as shown in Eq. 4.7,

$$\frac{1}{2}\sum\nolimits_{x\in\Omega} w_1\left(h(x) * f(x) - g(x)\right)^2 + \frac{1}{2}\lambda \sum\nolimits_{x\in\Omega} w_2\left(c(x) * f(x)\right)^2 = \min \tag{4.7}$$

where the cost function is minimized based on the degraded image data $g(x)$, original image $f(x)$, and PSF $h(x)$. $c(x)$ is called the regularization operator and traditionally is a second derivate Laplace filter. The issue of non-directional (Laplace) versus directional operator has been debated firstly by Marr and Hildreth. A directional operator can be shown to have better localization and preserving discontinuities than the isotropic Laplacian filter, and achieve better visual perception. The outputs of operators of different size is difficult to combine since the supports differ markedly. For a given operator width, both signal to noise ratio and localization improve as the length of the operator (parallel to the edge) increases, provided of course that the edge does not deviate from a straight line. λ is a regularization parameter that controls the trade-off between the fidelity to the observation and smoothness of the restored image. Normally, real images are piecewise smooth and additive noise is not spatially stationary. The trade-off should be spatially adaptive according to the local properties of image and noise. The adaptive space is adjusted by introducing two weights w_1 and w_2. Large w_1 emphasizes the fidelity of data where the noise is small or near sharp intensity transitions, otherwise it should be small. Large w_2 means smoothness near smooth areas or means large noise, otherwise it should be small.

4.2.3 Estimation in Image Domain

Based on the Bayesian form, our goal is to find the optimal $\hat{f}$ and $\hat{h}$ that maximizes the posterior $P(f, h|g)$ respectively.

$$\mathcal{J}(f|h, g) = -\log\{p(g|f, h)P(f)\}, \tag{4.8}$$
$$\mathcal{J}(h|f, g) = -\log\{p(g|f, h)P(h)\}, \tag{4.9}$$

express that the energy cost $\mathcal{J}$ is equivalent to the negative log-likelihood of the data. The priors $P(f)$ and $P(h)$ over the parameters are penalty terms added to the cost function to minimize the energy cost in a regularization framework for solving ill-posed problems [241], [86], [169], [28]. Another similar optimization framework called free energy is proposed by Jordan [121] in graphical models. Also, Opper et al. [183] proposed a variational free energy approach for approximate statistical inference. To avoid stochastic optimization (longer computing time)[85], [303], [209], we solve the optimization problem deterministically [86], [45], [289] in a convex manner with respect to the image and the PSF.

In the image domain, the cost function of image estimate can be minimized iteratively in the weighted space-adaptive regularized formulation. In this equation, $p(g|\hat{f}, h)$ follows a Gaussian distribution and $p(f)$ is prior knowledge with some constraint conditions.

$$\begin{aligned} \mathcal{J}(\hat{f}_{(g,h)}) &= \arg\max_{\hat{f}}[p(g|\hat{f}, h)P(\hat{f})] \\ &= \frac{1}{2}\sum\nolimits_{x\in\Omega} w_1(g(x) - h(x) * f(x))^2 + \frac{1}{2}\lambda\sum\nolimits_{x\in\Omega} w_2(c_1(x) * f(x))^2 \end{aligned} \tag{4.10}$$

where $p(g|\hat{f}, h) \propto \exp\left\{-\frac{1}{2}\sum_{x\in\Omega} w_1(g(x) - h(x) * f(x))^2\right\}$ and the prior of image is $p(\hat{f}) \propto \exp\left\{-\frac{1}{2}\lambda\sum_{x\in\Omega} w_2(c_1(x) * f(x))^2\right\}$. The first term is a fidelity term and the second is a smoothing term. Direct minimization of the cost function would lead to excessive noise magnification due to the ill conditioning of blur operator. A smoothness constraint $c_1(x)$ is an regularization operator and usually is a high-pass filter.

4.2.4 Estimation in PSF Domain

In PSF domain, PSF can be seen as maximizing the conditional probability. However, manipulation of probability density functions of PSFs in Bayesian estimation is difficult, and a decision must be made to attribute accurate initialization. The proposed prior solution space supports the parametrically structured PSFs in Bayesian estimation. A cost function for PSF from Eq. 4.3 is describing as the following:

$$\begin{aligned} \mathcal{J}(\hat{h}_{(g,f)}) &= \arg\max_{\hat{h}} \left\{ p\left(g \left| \hat{h}, f\right.\right) P_{\Theta}\left(\hat{h}\right) \right\} \\ &= \frac{1}{2}\sum\nolimits_{x\in\Omega} [g(x) - h(x) * f(x)]^2 + \frac{1}{2}\beta \sum\nolimits_{x\in\Omega} [c_2(x) * h(x)]^2 - \gamma \sum \log P_{\Theta}(\hat{h}) \end{aligned} \tag{4.11}$$

In Eq. 4.12, $p_{\Theta}(\hat{h})$ is the prior knowledge and needs to be computed. Θ is a set of primary parametric blur priors. Since both the original and observed images represent intensity distributions that cannot take negative values, the PSF coefficients are always nonnegative, $h(x) \geq 0$ Furthermore, since image formation systems normally do not absorb or generate energy, the PSF should satisfy $\sum_{x\in\Omega} \mathrm{h}\,(\mathrm{x}) = 1.0$, $x \in \Omega$, $\Omega \subset R^2$ is known. We need to compute:

$$\begin{aligned} p_{\Theta}\left(\hat{h}\right) &= \arg\max_{\theta} p\left(h_i(\theta) \left| \hat{h}\right.\right) \\ &= \arg\max_{\theta} \log \left\{ \frac{1}{(2\pi)^{\frac{LB}{2}} \left|\sum dd\right|^{\frac{1}{2}}} . \exp\left[-\frac{1}{2}\left(h_i\left(\theta\right) - \hat{h}\right)^T \sum\nolimits_{dd}^{-1} \left(h_i\left(\theta\right) - \hat{h}\right)\right] \right\} \\ &= \arg\min_{\theta} \left\{ \frac{1}{2} LB \log(2\pi) + \frac{1}{2}\log(\sigma_d^{2LB}) + \frac{1}{2\sigma_d^2}\left(h_i\left(\theta\right) - \hat{h}\right)^T \left(h_i\left(\theta\right) - \hat{h}\right) \right\} \end{aligned} \tag{4.12}$$

We define the likelihood of the neighbor $\hat{h}$ and in resembling the ith parametric model $h_i(\theta)$, $h_i(\theta) \in \Theta$. The first subscript i denotes the index of blur kernel. The modeling error $d = h_i(\theta) - \hat{h}$ is assumed to be a zero-mean homogeneous Gaussian distributed white noise process with covariance matrix $\sum_{dd} = \sigma_d^2 I$ independent of $f(x, y)$. LB is an assumed support size of blur. In reality, most of blurs satisfy up to a certain degree of parametric structure. A best fit model $h_i(\theta)$ for $\hat{h}$ is determined according to the density distribution.

The cost functions of image estimation and PSF estimation can be shown to be quadratic with positive semi-definite Hessian matrices. Therefore, the two cost functions are convex functions which ensure convergence in their respective domains. The resulting method attempts to minimize double cost functions subject to constraints such as non-negativity conditions of the image and energy preservation of PSFs. Our objective of the convergence is to minimize double cost functions by combing these two cost functions. We propose to solve the equation in the following:

$$\begin{aligned} \min_{\hat{h},\hat{f}} \mathcal{J}(\hat{f},\hat{h}) = &\underbrace{\frac{1}{2}\sum\nolimits_{x\in\Omega} w_1 (g(x) - h(x) * f(x))^2}_{\text{fidelity term}} + \underbrace{\frac{1}{2}\lambda \sum\nolimits_{x\in\Omega} w_2 (c_1(x) * f(x))^2}_{\text{penalty term of images}} \\ &+ \underbrace{\frac{1}{2}\beta \sum\nolimits_{x\in\Omega} w_3 (c_2(x) * h(x))^2}_{\text{penalty term of PSFs}} + \underbrace{\frac{1}{2}\gamma \sum\nolimits_{x\in\Omega} w_4 (h_i(\theta) - \hat{h})^2}_{\text{learning term}} \end{aligned} \tag{4.13}$$

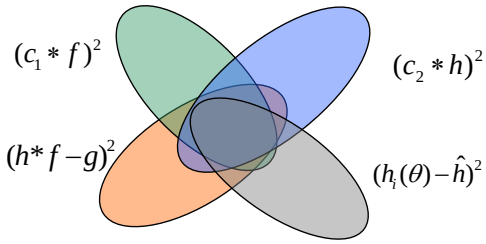


Figure 4.3: Representation of the intersection of the four convex sets. The proposed functional is presented in the set-theorey. Knowledge about the noise as well as other properties of the solution are directly incorporated into the restoration process, in terms of soft and hard constraints.

This double cost functional is quadratic in its variables so that it is a strictly convex functional. This functional can also be explained using the set theoretic approach followed by Katsaggelos [126]. The *a prior* knowledge constraints the solution to certain sets. Therefore, consistency with all *a prior* knowledge pertaining to the original image serves as an estimation criterion. Therefore, deterministic and/or statistical information about the undistorted image and statistical information about the noise are directly incorporated into the iterative procedure. The restored image is the center of an ellipsoid bounding the intersection of four convex sets, shown in Fig. 4.3.

However, the most tractable criterion-mean square error does not ensure human perceptual image restoration. Some soft constraints such as weight must be incorporated into the iterative minimization. If we ignore the effect of weights for the moment, the goal in the above minimization problem is to find an estimate image which makes the mean squared estimation error small and yet would not allow $\hat{f}$ and $\hat{h}$ to have much high frequency content in $\hat{f}$ and $\hat{h}$, respectively. The weight functions representing image local variances make their possible to allow high frequency content in $\hat{f}$ in the high activity (edge and texture) regions and to heavily penalize such content in low activity (smooth) regions. The weights are calculated according to [126], [280], [135]. Adaptive weights can be computed using fixed, variable and adaptive windows between zero and one.

1. $w_1 = 1$, if data at x is reliable, otherwise $w_1 = 0$;
2. The image weight $w_2 = 1/[1 + \alpha_2 \hat{\sigma}_f^2(x)]$, $\hat{\sigma}_f^2(x)$ is local variance of the observed image at x in a given window, and $\alpha_2 = 1000/\sigma_{max}^2$ is a tuning parameter designed so that $w_2 \to 1$ in the uniform regions and $w_2 \to 0$ near the edges.
3. Regarding to the weight of PSF, we take $w_3 = 1$, $w_4 = 1$. The reason is that most parametric blur kernels are homogeneous smoothness, the regularization operator c_2 can adjust the smoothness of PSF.

A similar idea was proposed by Polzehl et al. [196], they use an adaptive weights smoothing method for image restoration. In this method, nonparametric image estimation is based on locally constant smoothing with an adaptive choice of weights for every pair of data points.

The resulting method attempts to minimize double cost functions subject to constraints such as non-negativity conditions of the image and energy preservation of PSFs. During the implementation, λ, β, γ including diagonal matrices assign different emphases on the balance of the convergent PSF and image. The cost function of this equation is minimized in an alternating optimization approach via conjugate gradient descent.

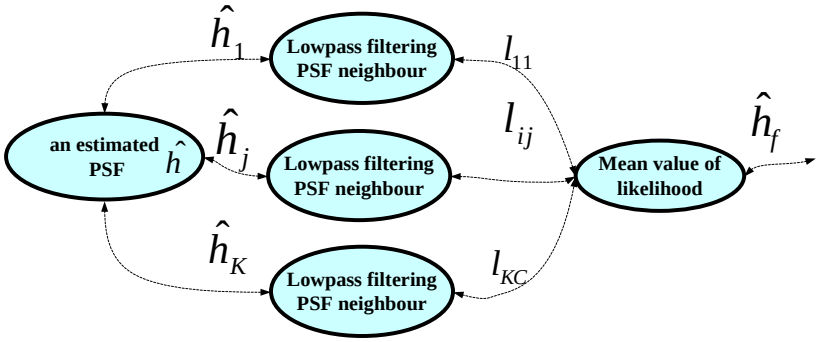


Figure 4.4: Diagram of K-nearest neighbors based nonparametric density estimation for PSF estimation.

4.2.5 Statistical Model Selection and Parametric Modeling

Blind image deconvolution methods can be classified into two basic categories: parametric and non-parametric methods. Parametric modeling methods assume that the PSF satisfies a known parametric structure. The overall flexibility of estimation is thus restricted and converted to the selection of a reasonable parametric model from the proposed solution space. In contrast, nonparametric modeling do not make any assumption of blur kernels, resulting in solving more natural blur identification, e.g., partially-blur, non-stationary or nonuniform blur. The underlying connections between these two methods is that nonparametric modeling can be represented by many parametric models. For example, Gaussian mixture model can be considered as a nonparametric modeling method. The dilemma can thus be resolved by some strategies, e.g., sampling methods, nonparametric model selection methods. Here, we propose a K-nearest neighbors (K-NN) method for the selection of a reasonable parametric model from the solution space. Then we use a weighted mean filtering method to estimate the PSF that can be locally optimized in further iterative regularization.

During the estimation process, each estimated PSF $\hat{h}$ for a sampling image region normally has some noise error $\hat{h}_j = \hat{h} + noise$. During the deconvolution, the coefficients of $\hat{h}$ are influenced by noises so that the high-frequency contents are unreliable. The reliable information of the PSF is its low-frequency contents. Since, we have a lot of sampling of PSF according to the principle in Fig. 4.5. To achieve a better estimated PSF with less noise, the estimated PSF is convolved with several low-pass filters to remove noise and get several PSF neighbors $\hat{h}_j, j \in (1, ..., K)$ in OTF format (Optical Transfer Function, OTF is PSF in frequency domain). The inverse discrete Fourier transform is thus performed to generate K number of PSFs.

In order to study the interaction between statistical blur kernel knowledge and blur degraded image information, we define the likelihood $P(\hat{h}_j)$ of the estimated PSF $\hat{h}$ of an observed image in resembling the ith parametric model $h_i(\theta)$ in a multivariate Gaussian distribution,

$$P(\hat{h}_j) \propto \arg\max_{\theta} \log \left\{ \frac{1}{(2\pi)^{\frac{LB}{2}} \left|\sum dd\right|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2} \left(h_i(\theta) - \hat{h}_j \right)^T \sum\nolimits_{dd}^{-1} \left(h_i(\theta) - \hat{h}_j \right) \right] \right\}$$

The first subscript i denotes the index of blur kernel. The modeling error $d = h_i(\theta) - \hat{h}$ is assumed to be a zero-mean homogeneous Gaussian distributed white noise process with covariance matrix $\sum_{dd} = \sigma_d^2 I$ independent of image. $L \times B$ is an assumed support size of blur. Then the Gaussian

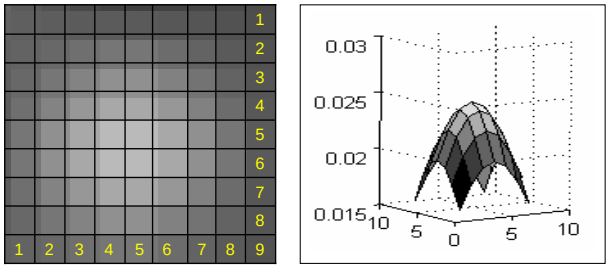


Figure 4.5: An example of a blur kernel with 9×9 pixel support size

probability corresponds to a PSF learning likelihood:

$$l_{ij}(\hat{h}_j) = \frac{1}{2} \exp\left\{(h_i(\theta) - \hat{h}_j)^T \sum\nolimits_{dd}^{-1} (h_i(\theta) - \hat{h}_j)\right\} \tag{4.14}$$

In reality, most of blurs satisfy up to a certain degree of parametric structures. A best fit model $h_i(\theta)$ for $\hat{h}$ is selected according to the Gaussian distribution and a weighted mean filter. The mean value of PSF learning likelihood $l_i(\hat{h})$ is that $l_{ij}(\hat{h}_j)$ is weight divided by $d(\hat{h}, \hat{h}_j)$. $d(\hat{h}, \hat{h}_j)$ is the Euclidean distance between $\hat{h}$ and its neighbor $\hat{h}_j$,

$$l_i(\hat{h}) = \sum\nolimits_{j=1}^{K} [l_{ij}(\hat{h}_j) d^2(\hat{h}, \hat{h}_j)] / [d^2(\hat{h}, \hat{h}_j)] \tag{4.15}$$

The weighted mean likelihood $l_i(\hat{h})$ depends on two conditions using a weighted mean filter. The first condition is the likelihood value of the blur manifold $l_{ij}(\hat{h}_j)$, and the second is the distance between $\hat{h}$ and its neighbor $\hat{h}_j$. The estimated output blur model $\hat{h}_f$ is obtained from the parametric blur models using

$$\hat{h}_f = [l_0(\hat{h})\hat{h} + \sum\nolimits_{i=1}^{C} l_i(\hat{h}) P(\hat{h}_j)] / [\sum\nolimits_{i=1}^{C} l_i(\hat{h})] \tag{4.16}$$

where $l_0(\hat{h}) = 1 - max(l_i(\hat{h}))$, $i = 1, ..., C$. The main objective is to assess the relevance of current estimated blur $\hat{h}$ with respect to parametric PSF models, and to integrate such knowledge progressively into the computation scheme. If the current blur $\hat{h}$ is close to the estimated PSF model $\hat{h}_f$, that means $\hat{h}$ belongs to a predefined parametric blur structure. Otherwise, if $\hat{h}$ differs from $\hat{h}_f$ significantly, this means that current blur $\hat{h}$ may not belong to the predefined PSF priors. The solution space of PSF kernels supports stronger parametric prior for the next iterative regularization. This method allows the construction of a representative solution for any special data acquisition environments.

4.3 Alternating Minimization

You and Kaveh [280] introduced a joint L^2 norm regularization method of the image and blur kernel for nonparametric blur identification. Later, Chan and Wong [44] demonstrated this method in TV (L^1 norm) based joint regularization for image and blur kernel. To achieve the joint results, a scale problem arises between the minimization of the PSF and the image via

steepest descent. The reason is that the $\partial\mathcal{J}/\partial\hat{h}$ is $\sum_{x\in\Omega}\hat{f}(x)$ times larger than $\partial\mathcal{J}/\partial\hat{f}$. Also, the dynamic range of the image $[0, 255]$ is larger than the dynamic range of the PSF $[0, 1]$. The scale factor changes dynamically with space coordinates (x, y).

To avoid the scale problem, an alternate minimization method following the idea of coordinate descent [280], [151], [152] is applied. The alternating minimization decreases complexity. The formulation is derived from the double cost functional in the following:

$$p(x) = \frac{\partial\mathcal{J}(\hat{f},\hat{h})}{\partial\hat{f}(x)} \text{ and } q(x) = \frac{\partial\mathcal{J}(\hat{f},\hat{h})}{\partial\hat{h}(x)}$$

1. Initialization:

$$\hat{f}^0(x) = g(x),\ \hat{h}^0(x) \text{ get from Eq. (14)} \tag{4.17}$$

2. nth iteration: restoration step under a fixed $h(x)$

$$\hat{f}_n(x) = \arg\min_{\hat{f}} \mathcal{J}_f(\hat{f}|\hat{h}_{n-1}, g) \tag{4.18}$$

3. (n+1)th iteration: identification under a fixed $f(x)$

$$\hat{h}_{n+1} = \arg\min_{\hat{f}} \mathcal{J}_h(\hat{h}|f_n, g),\ h(x) \geq 0 \tag{4.19}$$

4. If convergence is reached, then stop iterating.

The global convergence of the algorithm to the local minima of cost functions can be established by noting the two steps in Eq. 4.18 and Eq. 4.19. Since the convergence with respect to the PSF and the image are separate and optimized alternatively, the flexibility of this proposed algorithm allows us to use conjugate gradient algorithm for computing the convergence. Conjugate gradient method utilizes the conjugate direction instead of local gradient to search for the minima. Therefore, it is faster and also requires less memory storage when compared with quasi-Newton method. To get a convergent value of PSF, let $v(x)$ be the element at x of the conjugate vector. The conjugate gradient descent is given by the following:

- Initialize the conjugate vector from $q(x)$:

$$v_0(x) = -q_0(x) \tag{4.20}$$

- Step size for updating the PSF in iteration k:

$$\alpha_k = \frac{\sum_{x\in\Omega}[q_k(x)]^2}{\sum(v_k * \hat{f}_k)^2 + \beta\sum(c_2 * v_k)^2 + \gamma\sum(c_2 * v_k)^2}$$

- Update the PSF:

$$\hat{h}_{k+1}(x) = \hat{h}_k(x) + \alpha_k v_k(x) \tag{4.21}$$

- Step size for updating the conjugate vector:

$$\beta_k = \frac{\sum_{x\in\Omega}[q_{k+1}(x)]^2}{\sum_{x\in\Omega}[q_k(x)]^2} \tag{4.22}$$

- Update the conjugate vector:

$$v_{k+1}(x) = -q_{k+1}(x) + \beta_k v_k(x) \tag{4.23}$$

The above steps should be stopped after n steps. To compute convergent image, the conjugate gradient descent algorithm is described as:

- Initialize the conjugate vector:

$$u_0(x) = -p_0(x) \tag{4.24}$$

- Step size for updating the image in k iteration:

$$\alpha_k = \frac{\sum_{x\in\Omega}[p_k(x)]^2}{\sum(\hat{h}_k * u_k)^2 + \lambda\sum(c_1 * u_k)^2}$$

- Update the estimated image

$$\hat{f}_{k+1}(x) = \hat{f}_k(x) + \alpha_k u_k(x) \tag{4.25}$$

- Step size for updating the conjugate vector:

$$\beta_k = \frac{\sum_{x\in\Omega}[p_{k+1}(x)]^2}{\sum_{x\in\Omega}[p_k(x)]^2} \tag{4.26}$$

- Update the conjugate vector:

$$u_{k+1}(x) = -p_{k+1}(x) + \beta_k u_k(x) \tag{4.27}$$

If an image has $M \times N$ pixels, the above conjugate method will converge to the minimum of $L_f\left(\hat{f}\,|g,h\right)$ after $m \ll MN$ steps based on partial conjugate gradient method. The update of weights w_1, w_2, w_3 and w_4 is done after the conjugate gradient descent algorithm in order not to influence the conjugacy of the descent vectors. Real images or video have only a few very large frequency components and the others are very close to zero. Thus the Hessian matrices become sparse, and only a small number of $n = (5-15)$ iterations can get the convergence.

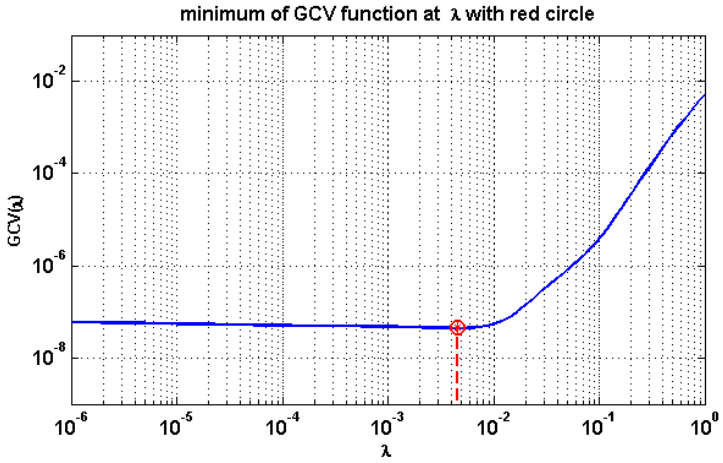


Figure 4.6: Generalized cross validation for the estimation of regularization parameter λ. The corner of the GCV curve is the best estimated regularization parameter λ.

4.4 Parameters Selection of Iterative Regularization

The parameter selection of regularization is discussed intensively in the literature due to its importance, especially in deblurring. The reason is that the deblurring is related to the estimation of image f and the blur identification h. There is a "balance" between the estimation of PSFs and images. We discuss three types of solution for regularization [105]. The first one is the original Tikhonov regularization which has a penalty term to the problem to filter out unwanted components. The second method is the truncated singular value decomposition regularization (TSVD) [104]. This method is projected into a specific subspace without the unwanted components. The third one is a "hybrid" regularization approach which combines Tikhonov and TSVD into one approach. The subspace methods have one regularization parameter, namely the subspace dimension k corresponding to the iteration count for iterative methods. On the other hand, the penalty methods can have multiple regularization parameters - one for each penalty term.

In our approach, a single Tikhonov regularization is extended to double regularization with respect to the image and the PSF. Therefore, the regularization parameters with respect to the image and the PSF need to be estimated. We have studied three types of parameter-selection methods which are described in the following part.

4.4.1 Generalized Cross-Validation

Generalized cross-validation (GCV) is a method [94] that does not depend on a priori knowledge about the noise variance. It is also a rotation-invariant form of ordinary cross-validation [54] and has some robust behavior for certain situations. The regularization parameter can be chosen to minimize the GCV function in the case of single regularization.

$$C(\lambda) = \frac{\|h\hat{f} - g\|_2^2}{[\text{trace}(I - hh^I)]^2} \tag{4.28}$$

where h^I is the matrix which produce the regularized solution, i.e., $\hat{=}h^I g$. For a single Tikhonov regularization, $h^I = (h^T h + \lambda^2 I)^{-1} h^T$. The cross-validation can be used to estimate the regular-

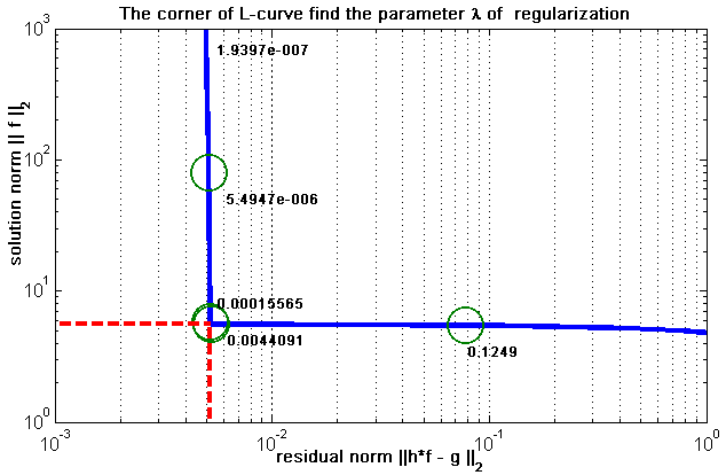


Figure 4.7: L-curve method for the estimation of regularization parameter λ. The corner of L-curve is the best estimated regularization parameter λ.

ization parameters [202]. For example, in the space-invariant blur case, each set h_i is considered as a single pixel. Thus, a "restored" image is determined using the values from the observed image g at a fixed regularization parameter, and the "restored" image is re-blurred in order to predict the blurred and noisy observation that was left out of the restoration. A different "restored" image is formed for each observation. The regularization parameter that minimizes the mean square prediction error over all the observation is chosen as the estimated optimal parameter. However, the performance suffers from the assumption of a stationary image, and the algorithmic complexity and computation load are still high.

One noted difficulty with GCV is that G can have a very flat minimum in the GCV curve, making it difficult to determine the optimal λ numerically. The solution estimates fails to converge to the true solution as $n \to \infty$ or as the error norm goes to zero, shown in Fig. 4.6.

4.4.2 L-Curve Method

The L-curve method is based on heuristic observations that are directly used as a parameter-selection methods and named by Hansen [106]. The L-curve attempts to balance the penalty term and the fidelity term for the regularized solution with regularization parameters with respect to the image and the PSF. The L-curve comes from the characteristic shape of the curve $(\log \|h\hat{f} - g\|^2, \log \|\hat{f}\|^2)$. Functions are presented in logarithmic scaling. A small parameter λ yields a large penalty term and a small model fit norm. Similarly, a large λ gives a small penalty term and a poor fit. The idea is to find a balance value between two terms at the corner of the L-curve, shown in Fig. 4.7.

However, in our case, we use double regularization with respect to the blur kernel and the image which are interleaved constraints and prior knowledge. In this case, we can directly use other methods to estimate the regularization parameters, e.g., the L-curve method [106]. The L-curve criteria are likely to terminate prematurely and return a regularization parameter λ which is significantly too large. GCV may require more Lanczos iterations but it provides reliable answers in return. The plot of the norm of the regularization parameter values, was introduced by Lawson and popularized by Hansen [106]. Intuitively, the best regularization parameter should lie on the corner of the L-curve, since the residual increases without reducing the norm of the solution

much. The norm of the solution increase rapidly without much decrease in residual. In practice, only a few points on the L-curve are computed and the corner is located by estimating the point of maximum curvature [106].

To find a useful algorithm we need a precise definition of the corner of the L-curve. Hansen et al. suggested using the point of maximum curvature. This approach is invariant to scaling of the equations, but the computation needs derivatives of the penalty and residual fit functions.

4.4.3 Morozov's Discrepancy Principle

The Morozov's discrepancy principle [171], [172] selects the regularization parameter so that the model fit $\|hf - g\|_2$ is equal to an upper bound on the error δ in the following,

$$\|h\hat{f} - g\|_2 = \delta_e \tag{4.29}$$

where $\|e\|_2 \leq \delta^2$. If we know the norm of the noise $\|e\|_2 = \delta$ (the noise level), it does not make sense to ask for a solution $\hat{f}$ where $\|h\hat{f} - g\| < \delta$. The iterative methods GMRES and LSQR (available in matlab tool box) all have monotonically decreasing residuals and iterations can be stopped when the residual norm passes the error δ_e. This mechanism makes the discrepancy principle a perfect choice for these methods in case we know the noise-level.

Tikhonov regularization is a half quadratic functional and it is strictly convex in the case of blur identification. The residual norm of the regularization solutions are also monotonically converging with respect to the regularization parameter. Therefore, the Morozov discrepancy principle is well defined for searching parameters in Tikhonov regularization.

However, the noise is not always available and an estimation may be unreliable for an observed blurred noisy image. The GCV and L-curve method do not require the noise-level. Because these two methods do not consider the noise level to influence the regularization parameters. An experimental comparison in [105] is done to show how good the the optimal regularization parameters are. In practical environment, we have tried these different methods for finding a best regularization parameter for blur identification and image restoration.

4.4.4 Self-Adjusting PSF Support

The support size of blur kernel is more important than the coefficients of blur kernel. For real blurred image or video data, one of the main differences between stationary blur and non-stationary blur is the size of blur kernel changing continuously and randomly. Therefore, the restoration of non-stationary blur needs reasonable sampling methods and estimation methods to follow the changing of blur kernels continuously. In this section, we introduce an accurate estimation method for estimating the support size of stationary blur.

From frequency point of view, spectrum or cepstrum of blurred images are used to perform blur identification [135], [136] However, these methods are sensitive to additive noise and changes of image structure and texture due to the restriction of the Fourier transforms of PSFs. Autoregressive (AR) and moving average (MA) processes are used to model the true image and blur kernel respectively. Under this autoregressive moving average method (ARMA) framework, statistical methods are employed to estimate the blur parameters for the objective of blind image restoration such as maximum-likelihood (ML) estimation [135], generalized cross-validation

(GCV) [202]. ML method is used to maximize the log-likelihood function for getting the parameter set in ARMA solution space. GCV determines parameters by minimizing a weighted sum of predictive errors. Chen et al. [46] proposed a method to identify the support size using maximum average square difference and maximum average absolute difference based on ARMA.

In our method, in order to ensure the actual PSF support size, the boundary of the assumed PSF support is decreased at each iteration. After several iterations, the approximate PSF support size can be reached till the convergence of support size is stable. Because of the nonnegative constraints of the PSF, the boundary size is adjusted by giving a positive size threshold. Although different parametric PSFs have different kernel, the self-adjusted PSF support is always rectangular or circular.

4.5 Experimental Results

Experiments on synthetic and real data are carried out to demonstrate the effectiveness of our algorithm.

4.5.1 Adaptively Weighted Image Smoothing Parameters

The choice of regularization parameters is crucial. You and Kaveh propose a way of setting these parameters. However, as they have pointed out in their paper, these values are meant to be used as guidelines only and not as exact values. To find a good way of determining the values of regularization parameters, we use the L-curve method [106] due to its robustness for correlated noise. It is a graphical tool for analysis of discrete ill-posed problems in a log-log plot for all valid parameters using the compromise between minimization of these quantities. The novelty is that no prior knowledge about the properties of the noise and the image (other than its "smoothness") is required, and required parameters are computed through our proposed double cost functions. There is a relative scale relation between λ and β. It is formulated as

$$\beta/\lambda = \sum\nolimits_{x\in\Omega} \hat{f}\,(\mathrm{x}) \max_{x\in\Omega} \hat{f}\,(\mathrm{x}) \tag{4.30}$$

The order-of-magnitude of two parameters are given using the normalized local variance of image and PSF, $\lambda_i = 0.5/(1+10^3\mathrm{var}(f(i))$, $\beta_i = 10^6/(1+10^3\mathrm{var}(h(i))$ and $\gamma_i = 10^6/(1+10^3\mathrm{var}(d(i))$, where $d = h_i\,(\theta) - \hat{h}$. We have also tried different weights to determine the suitable weighting scheme. Piecewise smooth and ringing reduction can slightly compensate the error of estimation so that $w(x,y) \to 1$ in the uniform regions and $w(x,y) \to 0$ near the edges. A meaningful measure called normalized mean square-error (NMSE) is used to evaluate the performance of the identified blur,

$$NMSE = \frac{\left(\sum_x \sum_y (h(x,y) - \hat{h}(x,y))^2\right)^{\frac{1}{2}}}{\sum_x \sum_y h(x,y)} \tag{4.31}$$

where $h(x,y)$ and $\hat{h}(x,y)$ are the true and estimated blur. We test the approach in different blur status. Accurate parameters of PSF can be adjusted based on the minimized cost function and NMSE measure. The closed PSFs of NMSE normally has a range [0, 0.1] depends on the different PSFs.

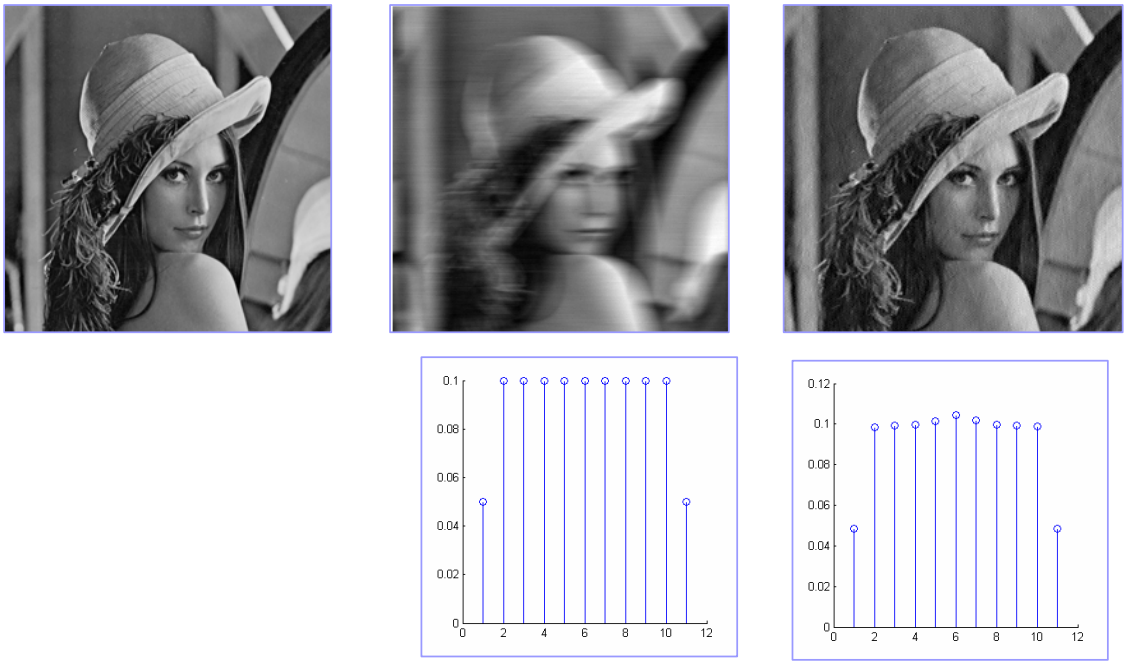


Figure 4.8: $a|b|c$ Recovered PSF and restored image. The first row (left to right): (a) Original image. (b) Synthetic motion blurred image without any additive noise. (c) Restored image using toeplitz-circular block matrix approximation weak noise. The second row (left to right): original PSF, identified PSF. From this experiment, without noise, SNR=$+\infty$, the restoration has very weak ringing effects for the motion blur. We may find most ringing effects and influences coming from noises and blur. Gaussian blur and out-focus blur has more stronger ringing effects than motion blur.

4.5.2 Blind Deconvolution of Degraded Image

To evaluate this algorithm, the performance of the approach is investigated by using simulated blurred images and real video data at different signal-to-noise ratios. The performance of image restoration is measured by SNR improvement and formulated as the following,

$$ISNR = 10\log_{10}(||f-g||^2/||f-\hat{f}||^2)(dB) \tag{4.32}$$

Simulated experiments are performed in standard images. The identified PSFs and restored images are illustrated in Fig. 4.8. A MRI image has been degraded by three different blur kernels with quantization noise SNR 20dB. The proposed algorithm was applied to the degraded image. The final restored image and the identified blur are given in Fig. 4.9, respectively. It can be observed that the overall textured and edge region of the image has been recovered. This second

Table 4.2: ISNR results on test data

SNR (dB)	SNR IMPROVEMENT (dB)					
	Motion blur		Gaussian blur		Uniform	
	5x5	7x7	5x5	7x7	5x5	7x7
30	5.32	4.98	5.32	4.63	5.76	5.72
noiseless	5.88	5.12	5.56	4.86	5.87	5.97

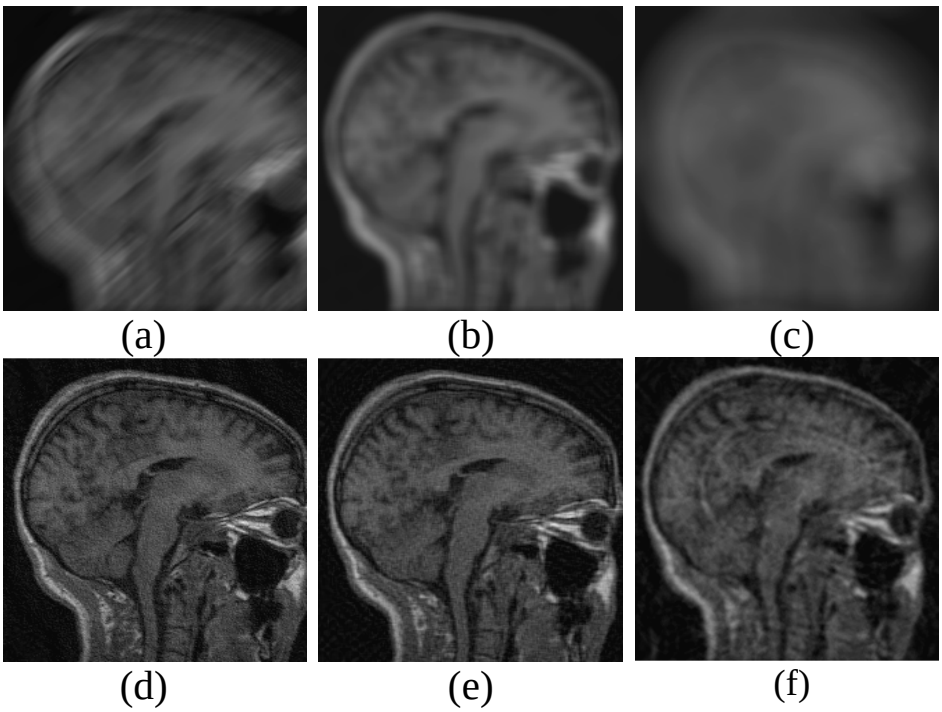


Figure 4.9: (a)(d) Blurred image and result of blind deconvolution, ISNR = 5.29dB. (b)(e) Blurred image and result of blind deconvolution, ISNR = 5.27dB. (c)(f) Blurred image and result of blind deconvolution, ISNR = 4.79dB.

experiment presents blind deconvolution of a degraded image to demonstrate the flexibility of the proposed algorithm. The original "Lena" image has a dimension of [256, 256] with 256 gray levels. It was degraded by 20 pixel linear motion kernel and additive SNR 30dB noise in Fig. 4.8 and Fig. 4.10. Comparison between Fig. 4.10(b) and (c) reveals the good performance of our algorithm. The ringing reduction is efficient while preserving the fine details of eyes and feather. Fig. 4.10 shows the efficiency and accuracy of our proposed algorithm comparing with Lucy-Richardson algorithm.

The third experiment tests the robustness of the proposed method in different blurs. The "Lena" is simulated in different degraded images. Table 4.2 summarize the results and demonstrates that the method is effective in restoring images under different sizes and types of blur with different noise levels.

4.5.3 Blind Deconvolution of Degraded Objects in Video Data

In this experiment, we illustrate the capability of the proposed algorithm to handle real-life video data degraded by non-standard blur in Fig. 4.12. The video frames are captured from films or video test data. The degraded video objects are separated into RGB colour channels and each channel is performed respectively. Based on the estimated PSFs and parameters, piecewise smooth and accurate PSF model helps to suppress the ringing effects.

Figure 4.10: (a) Blurred noisy image. (b) Restored image based on Lucy-Richardson algorithm 100 iterations with known PSF, ISNR=5.35 dB (c) Blind image deconvolution using our algorithm, ISNR = 6.16 dB.

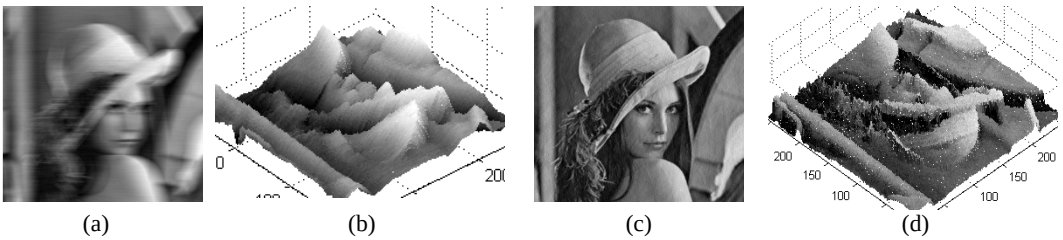


Figure 4.11: Example of blind image restoration and surface, 512×512. (a) Blurred noisy image. (b) Corresponding surface. (c) Restored image. (d) Corresponding surface.

4.5.4 Effects of Boundary Conditions

The proposed method is space adaptive weighted double regularization which has advantages of piecewise smoothness and stronger suppression of ringing effects. However, in the experiments, we still observe there are some ringing effects in Fig. 4.12c and Fig. 4.13c. The reason is that we use the periodic boundary conditions during the blur identification and image deconvolution. The periodic boundary conditions introduce discontinuities which entail ringing artifacts or some false discontinuities and edges in the boundaries of the restored image frequently. We propose three approaches to solve this problem. One way is to mitigate these artifacts as well as the undesired wrap-around of image information in the deblurring with periodic boundary conditions, the image can be extended continuously to a larger image with equal gray-values at opposing boundaries. Periodic boundary conditions will not introduce the false discontinuities or edges any more. The wrap-around influences the amended parts of the image. Periodic extension of this larger image is equivalent to reflecting extension of the original image. Fortunately, the periodic boundary conditions are compatible with any shift-invariant blur, without imposing symmetry constraints on the blur kernel. The second way is to use Neumann boundary condition during the image deconvolution. For example, Ng et al. [175] proposed to establish similar results in the two-dimensional case for deblurring, where the blurring matrices will be block Toeplitz-plus-Hankel matrices with Toeplitz-plus-Hankel blocks (BTHTHB). Finally, we can observe the restored image is of a relatively large size so that the problem has transformed to one of the deconvolution of large-scale images. Golub et al. introduced a method based on Morozov's discrepancy principle which largely solve the large-scale regularization problem.

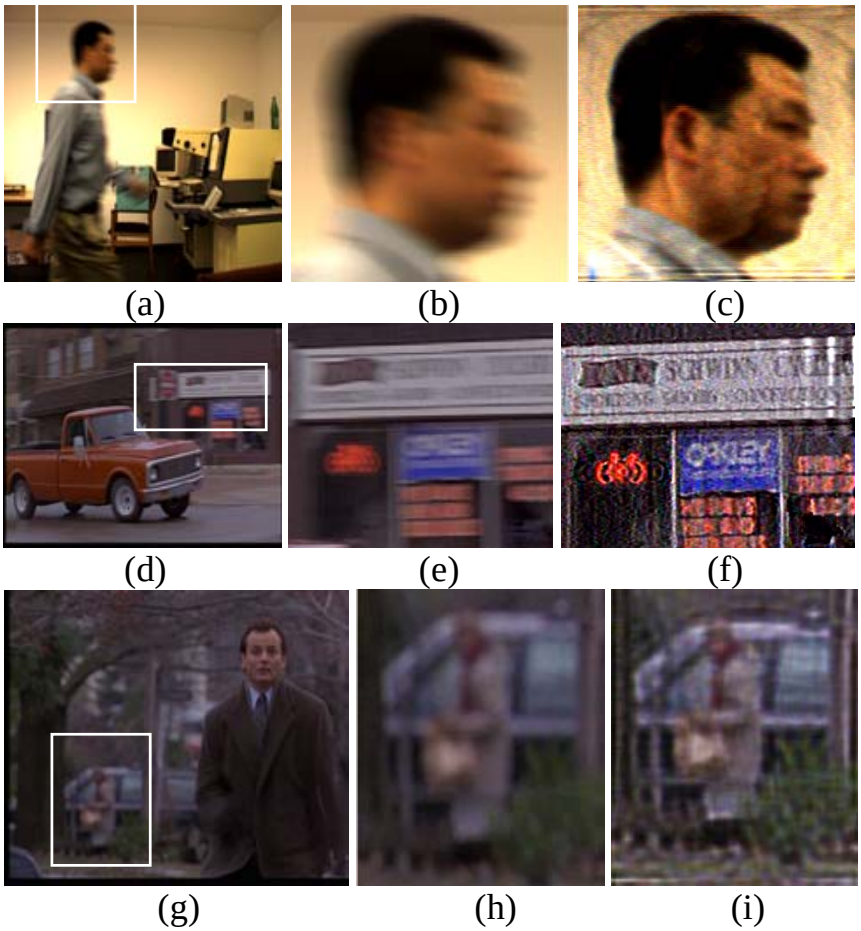


Figure 4.12: (a)(d)(g) Real video frames. (d)(e)(h) Blurred parts in video. (c)(f)(i) Results of blind deconvolution

4.5.5 Effects of Non-stationary Blur

When the blur kernel is changed continuously, blur identification becomes more difficult. From experimental results in Fig. 4.12f and Fig. 4.13d, we can observe that there are some gaps in restored images due to different sampling areas for blur identification. The reason of this effects might comes from three points. The first point is the influence of non-stationary blur kernel, i.e., the blur kernel of this region is a little different from the sampling region. The second reason is that the boundary condition problem, i.e., this image is restored using periodic boundary condition in BCCB matrix FFT discretization. The parameters of the blur kernel in Fig. 4.13d is not accuracy for the region in the green color framework.

4.5.6 Effects of Noises

The functional is constructed based on the assumption of additive Gaussian noise. The information about the noise is incorporated into the algorithm with the use of regularization parameter, which controls the tradeoff between noise amplification and deconvolution.

The proposed algorithm has been analyzed and experimental results have been shown. Based on these results, we concluded that the performance of the algorithm is satisfactory for synthetic

Figure 4.13: $\frac{a|b}{c|d}$ (a) Original video. (b) blurred background. (c) Restored image (d) Restored image based on different sampling area.

and real images with a wide range of noise, SNR's $(15 \rightarrow +\infty)$ dB. The image restored by the adaptive weighted double regularization have better visual quality than the image restored by the traditional Wiener filters and by the non-adaptive algorithms.

Finally, a direction of the next step work is the development of iterative algorithms in removing nonlinear distortion in the presence of noises, based on recent results on regularization theory. This work will be presented in the next chapter.

4.6 Discussion

In classical image restoration methods, A Wiener filter replaces the inverse filter with some penalties at denominator. However, the Wiener filter lacks rigorous justification with respect to the iterative convergence. Most of traditional methods like GCV, ARMA are space-invariant methods, the proposed method is space-adaptive with piecewise smoothness of images. The piecewise smoothness of both the image and the PSF are incorporated into the unsupervised image restoration process via space-adaptive regularization with the constraints of nonnegativity, interleaved prior knowledge and the good initial value. A self-pruning algorithm can automatically estimate the PSF support.

4.6.1 From Global Nonparametric Estimation to Local Parametric Optimization

The traditional nonparametric methods have the error of statistical estimation and the error of approximating the underlying function by the given functional family. Although a number of asymptotic minimax approaches [231] try to "balance" these error, some limitations of the optimality in the asymptotic minimax sense do not ensure good sample properties from the finite

sampling feature space. The sampling feature space is "never fullfilled" (in some sense) in that the asymptotic results are usually assuming an infinite sampling number $n \to \infty$. More over, the quality of sampling features also need to represent the characteristic properties of targets. For example, inhomogeneous regions including discontinuities are more useful than homogeneous sampling regions for blur identification. Although there are many different measures of descriptive information such as entropy, Gibbs distribution etc., the measure of blurred images is still difficult.

Practically, to achieve high accuracy blur identification in the case of non-stationarity for a sampling area, the first step is to classify and predicate the blur kernels in a family of right parametric PSF models. The second step is to optimize the parameters of the predicated parametric blur kernel PSF. Global nonparametric estimation can estimate the right distribution density in a sampling feature space. However, the choosing of right parameters is impossible in nonparametric regression methods or kernel density estimation methods. On the other hand, in the nonparametric theory there exist a number of modern welled developed methods such as automatic parameter selection like Cross-Validation [94] and Generalized Cross-Validation [94], Acaike [3], or Schwarz methods (Bayesian Information Criterion, BIC) [223]. These methods are not based on asymptotic minimax considerations but they are still hard for adjusting parameters of blur kernels without the constraints of image restoration.

The iterative structure of the suggested algorithm offers a number of advantages over non-iterative and recursive techniques, including the possibility of directly incorporating deterministic knowledge and soft statistical learning models into the restoration process, with the use of hard constraints. These constraints actually represented by projection onto convex sets. The soft or statistical constraints is in turn a function of regularization parameter. The weight of local variances representing human visual system were incorporated into the algorithm according to the observed noises and image discontinuities.

Furthermore, the convenience of the suggested algorithm is its interleaved prior property and double constraints. The estimation with respect to the PSF and the image is locally parametric optimized in the alternating minimization. The image estimation step estimates the true image assuming that the current estimates of PSF is correct prior knowledge, and vice versa. The predicated parametric model as an accurate initial value in regularization is obtained using the nonparametric Bayesian model selection technique. Thus, the approach is actually based on a local parametric optimization in nonparametric estimation strategy which was described by V. Spokoiny [231]. The nonparametric estimation is adaptation of the parametric methods to the situation when the parametric structural assumption is not fulfilled.

4.6.2 Discussion of Related Optimization Approaches

Constrained Optimization

The important property of the Bayesian approach is that the Bayesian method which minimizes the deviation depends on an *a priori* distribution. It is both the main advantage and the main disadvantage of the Bayesian approach. The advantage is that we can develop methods in accordance with average properties of the function to be minimized. The disadvantage is the arbitrariness and uncertainty of how to fix the *a priori* distribution. To solve this problem, we use nonparametric estimation techniques to find a prior distribution for Bayesian estimation.

Also we put the Bayesian estimation in a convex energy optimization functional which can ensure the global convergence.

D. Geman and G. Reynolds developed a constrained restoration approach for the recovery of discontinuities [86]. The idea was developed with a somewhat different coupled objective function. The model is also "half quadratic " and the auxiliary variables are also noninteracting but there is one crucial difference: the quadratic form is not block circulant FFT transform in the frequency domain. As a result, optimization must rely on updating pixels one by one in the spatial domain in the usual fashion. In contrast, the optimization method in our algorithm is to update pixel values in the FFT frequency domain based on a global optimization strategy.

Furthermore, D. German and G. Reynolds combine the first and second order terms that give consistently good results in their experiments than using the first and second order alone. With only first order terms, the objective function would favor regions of constant grey level. This suggests that purely first-order models would introduce an artificial patchiness or mottling, which is exactly what has been covered in a variety of studies. To the extent that grey-level images of real scene have homogeneous regions, these regions are better defined by constant gradient, or even constant curvature, then by constant grey level (This would become a TV method). These analysis give us some hints to develop a visual perception based data-driven image restoration approach.

On-line learning

Normal optimization techniques such as gradient ascent are undesirable because of their slow convergence. Alternatively, conjugate gradient or various preconditioned forms of gradient ascent techniques can be used due to their rapid convergence for quadratic optimization problems. The cost function being minimized is strictly convex and the cost function converges to the global minimum. Therefore, the exact restoration will be identical to a reconstruction compute using the modified EM algorithm.

On the other hand, much interest was devoted to the problem of on-line learning in pattern recognition. When data are presented sequentially to the estimator, on-line algorithms change their hypothesis and use the most recent data only. Hence the storage of the entire set of data is avoided. As discussed by Opper previously [180], [**?**] when one applies a smooth realizable stochastic rule to a random data sets. The on-line algorithm can achieve similar asymptotic generalization rates as the more complicated optimal batch algorithms.

Furthermore, Amari et al. [10] proposed a different on-line learning algorithm which minimizes a statistical dependency among outputs. The dependency is measured by the averaging mutual information (MI) of the outputs. A natural Riemannian gradient in structured parameter spaces is developed to minimize the MI based on information geometry theory [8], [9]. The on-line learning method based on the natural gradient is asymptotically as efficient as the optimal batch algorithm. This algorithm is transformation invariant and can be directly applied to Independent Component Analysis (ICA) problems and can be further extended in solving computer vision and pattern recognition problems.

4.7 Conclusions

This chapter presents a weighted space-adaptive regularized Bayesian approach for blind blur identification and image restoration. The proposed algorithms are very flexible, since there are a number of parameters which control the final solution. First, the approach improves the accuracy of PSF estimation. Bayesian MAP estimation can then speeds up the minimization of related cost functions progressively based on the initialization of accurate prior models. The double cost functions are then projected and converged to image and blur domains precisely. During the alternating minimization procedure, piecewise smooth reconstruction of both image and PSF is adopted to improve the quality of restoration. It is clear that the proposed method is instrumental in blind image deconvolution and can be can be extended for signals of any dimensionality, as well as space-invariant and space-varying (nonuniform) distortions in practical environments.

5 Data-Driven Regularization for Variational Image Restoration in the BV Space

The problem of learning is arguably at the very core of the problem of intelligence, both biological and artificial - T. Poggio and C.R. Shelton

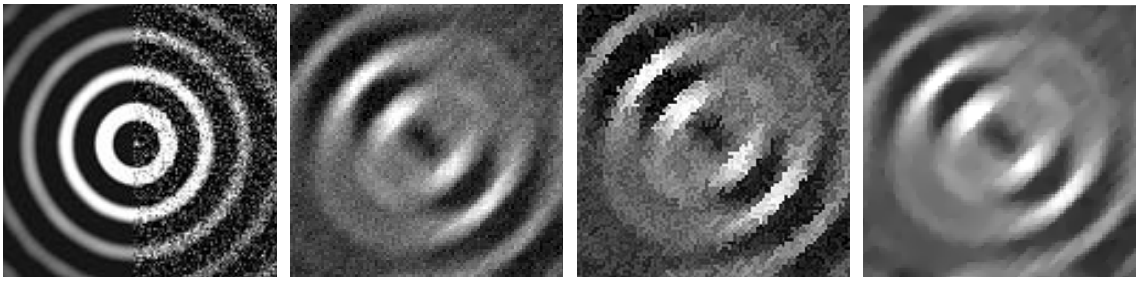
How to represent an image in a reasonable mathematical model in the spatial domain? This question is still a hot topic in mathematics, computer vision and other related research communities. This chapter presents a novel mathematical model represent an image in the space of functions of Bounded Variation (BV). It can be directly used for adaptive data-driven variational image restoration. As we know, the discontinuities are important features in image processing. The BV space is well adapted for the measure of gradient and discontinuities. More over, the degradation of images includes not only random noises but also multiplicative, spatial degradations, i.e., blur. To achieve high-quality image deblurring and denoising, a variant exponent linear growth functional in the BV space is extended in Bayesian estimation with respect to deblurring and denoising. The selection of regularization parameters is self-adjustable based on spatially local variances. Simultaneously, the linear and non-linear smoothing operators are continuously changed following the strength of discontinuities. The time of stopping the process is optimally determined by measuring the signal-to-noise ratio. The algorithm is robust in that it can handle images that are formed with different types of noises and blur kernels. Numerical experiments show that the algorithm achieves more encouraging perceptual image restoration results.

5.1 Introduction

5.1.1 Problem Formation and Proposed Approach

In classical Sobolev spaces, we can not make detailed analysis and measure for discontinuities. A simple image including a white disk on a black background is not in any Sobolev space, but belongs to the BV space. The BV space is the space of functions for which the sum of the perimeters of the level sets is finite. Therefore, the BV space is well adapted for the measure of discontinuities across edges. Compared to wavelet based methods in the frequency domain [95], the assumption of BV space is still too restrictive to represent the tiny detailed textures and infinite discontinuities [5]. However, currently, the BV space is still a much larger space than the Sobolev space for modeling images in the spatial domain.

Since the seminal work from Rudin, Osher and Fatemi (ROF) [213], the BV space based functionals have been widely applied to image restoration, super-resolution approaches, segmentation and related early vision tasks, e.g., modeling of oscillatory components from Meyer [164], modeling of inpainting and super-resolution approaches from Chan and Shen [42]. Other more work such as Mumford and Shah model [173] and its PDE version [40], Weickert [259], [260],



(a) (b) (c) (d)

Figure 5.1: (a) Ground truth. (b) Spatially degraded blurring image, Gaussian noise 15dB. (c) Restored image b using shock filter. (d) Restored image b using the normal TV.

and Schnörr [267], [220] have been proven to be very effective for image restoration and image enhancement.

Recently, [249], [16] propose a convex linear growth functional in the BV space for deblurring and denoising using Γ-convergence approximation. [48], [49] suggest a more general variable exponent, linear growth functional in the BV space for image denoising. Due to the complexity of image degradation, e.g., shown in Fig. 5.1, including spatial geometry distortion, additive or multiplicative noises and multiplicative spatial degradations (blur), we need to design more flexible and robust algorithms for denoising, deblurring and towards perceptual image restoration. Generally, a scheme of image restoration is to get a trade-off between the noise suppression and discontinuity (or edge) preservation, since noise reduction is achieved by constraining the image to be smooth. The spatial-invariant (nonstationary) image constraints are used to emphasize noise reduction in the "flat" region and preserve the discontinuities, edges and structures in "non-flat" regions in images. The regularization parameters can be chosen based on the reduction of noise. Furthermore, through literature study, we find that only little work is done on how to determine regularization parameters, and optimal diffusion operators for achieving optimal image restoration results.

In this chapter, we extend the variable exponent, linear growth functional from Chen, Levine and Rao [48], [49] to double regularized Bayesian estimation for simultaneously deblurring and denoising. The Bayesian framework provides a structured way to include prior knowledge concerning the quantities to be estimated [80]. Different from traditional "passive" edge-preserving methods [45], [35], [86], [205], [281], our method is an "active" data-driven approach which integrates self-adjusting regularization parameters and dynamic computed gradient prior to self-adjusting the fidelity term and multiple image diffusion operators. A new scheme is designed to select the regularization parameters adaptively on different levels based on the measurements of local variances. The chosen diffusion operators are automatically adjusted following the strengths of edge gradient. It has several important effects: firstly, it shows a theoretically and experimentally sound way of how local diffusion operators are changed automatically in the BV space. Secondly, the self-adjusting regularization parameters also control the diffusion operators simultaneously for image restoration. Finally, this process is relatively simple and can be easily extended for other regularization or energy optimization approaches. The experimental results show that the method yields encouraging results under different kinds and amounts of noise and degradation.

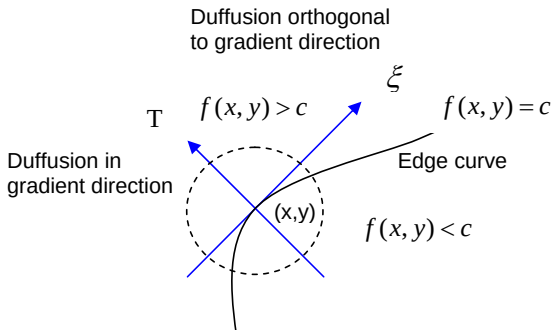


Figure 5.2: Orthogonal decomposition for image geometric analysis and an edge curve C separating homogeneous regions.

5.1.2 Total Variational Regularization for Inverse Problems

According to the image degradation model that has been defined previously, we have the following form,

$$g = hf + \eta \tag{5.1}$$

where an observed image in the image plane g is formed by two unknown conditions h and η.

Let us reconsider the energy function in Eq.(5.2)

$$\mathcal{J}(f) = \frac{1}{2}\int_{\Omega}(g - hf)^2 dxdy + \lambda \int_{\Omega} |\nabla f|^p dxdy \tag{5.2}$$

with $p = 1$. We would like to find the (unique) minimizer of f. Let $\Omega \subset \mathbb{R}^2$ denote the open image domain. The total variation (TV) prior model is defined in the distributional sense $TV(f) = \int_{\Omega} |Df| dxdy$ in the BV space [213].

$$BV(\Omega) = f \text{ such that } TV(f) < +\infty \tag{5.3}$$

The $TV(f)$ is often denoted by $\int_{\Omega} |Df| dxdy$, with the symbol D referring to the conventional differentiation ∇. The absence of the Lebesgue measure element dx (1D) indicates that $|Df|$ is a general Radon measure. A Radon measure is a Borel measure that is finite on compact sets. If $|Df|$ is the Borel sigma-algebra on some topological space, then a measure $m : |Df| \rightarrow \mathbb{R}$ is said to be a Borel measure (or Borel probability measure). For a Borel measure, all continuous functions are measurable. However, due to the complexity of the functions of BV space, one uses $f \in L^1(\Omega)$ to simplify the numerical computation (see [92], for instance),

$$\int_{\Omega} |Df| dxdy = \int_{\Omega} |\nabla f| dxdy \tag{5.4}$$

where ∇f belongs to L^1 which is simply the ordinary L^1 integral in the sense of Sobolev norm. Adopting the TV measure for image regularization, the posterior energy for Tikhonov deblurring takes the form which also appears in the TV functional [213],

$$\mathcal{J}(f) = \frac{1}{2}\int_{\Omega}(g - hf)^2 dxdy + \lambda \int_{\Omega} |\nabla f| dxdy \tag{5.5}$$

Table 5.1: Convex and nonconvex functions (edge-preserving)

Functions	$\phi(t)$	$\phi'(t)/(2t)$	convexity
Geman and Reynolds [86]:	$\frac{t^2}{1+t^2}$	$\frac{1}{(1+t^2)^2}$	no
Hebert and Leahy [108]:	$\log(1+t^2)$	$\frac{1}{1+t^2}$	no
Tikhonov [241]:	t^2	1	yes
Total Variation [213]:	$\lvert t\rvert$	$\frac{1}{a\lvert t\rvert}$(if $t \neq 0$)	yes
Green [97]:	$\log(\cosh(t))$	$\frac{\tanh(t)}{2t}$ (if $t \neq 0$)	yes
Hybersurface [45]:	$2\sqrt{1+t^2}-2$	$\frac{1}{2\sqrt{1+t^2}}$	yes

where g is the noisy image, f is an ideal image and $\lambda > 0$ is a scaling regularization parameter. A general bounded total variational function can be written in the following,

$$\mathcal{J}(f_{(g,h)}) = \frac{1}{2}\int_{\Omega}(g-hf)^2 dxdy + \lambda \int_{\Omega}\phi(|\nabla f(x,y)|)dxdy \tag{5.6}$$

In such type of energy functional, the choice of the function $\phi(\cdot)$ is crucial. It determines the smoothness of the resulting image function f in the space $V = \{f \in L^2(\Omega); \nabla f \in L^1(\Omega)\}$ which is not reflexive. For example, In Eq. 5.6, the first term on the right side is quadratic (convex). The second term including $\phi(\cdot)$ function (could be convex or non-convex) has been intensively investigated by researchers. According to the related work of Weickert and Schnörr [267], Aubert and Vese [17], Rudin, Osher and Fatemi [213], Chambolle and Lions [39], the ϕ-functions are usually classified in two categories, the nonconvex ones and the convex ones. The theoretical study shows that the convex term $\phi(\cdot)$ can lead the total energy function to an existing global convergence, while the nonconvex has non-uniqueness of the minimum, if it exists. Nevertheless, the non-convex functions are often used because they usually provide better results, e.g., shown in Table. 5.1 [86], [108] while they use special methods to solve these non-convex functions. Geman and Reyold [86] proposed to update the pixels one by one in the spatial domain in the usual fashion, [108] developed a Bayesian reconstruction based upon locally correlated Markov random field priors in the form of Gibbs function and upon the Poisson data model in discrete spatial domain.

In order to study more precisely the influence of the term $\phi(\cdot)$ in the regularization, we need to make an insight observation of geometric diffusion behavior which can help us to understand the convexity criteria in variational regularization.

Supposing that the integral in $\mathcal{J}(f_{(g,h)}$ in Eq. 5.6 have the form of $\phi(|\nabla f(x,y)|)$, the minima of $\mathcal{J}(f_{(g,h)}$ must formally verifies the Euler equation $\mathcal{J}'(f_{(g,h)} = 0$ or,

$$-\frac{\lambda}{2}\mathrm{div}\left(\frac{\phi'(|\nabla f|)}{|\nabla f|}\nabla f\right) + h^*hf = h^*g \tag{5.7}$$

where h^* denotes the adjoint operator of h. Since h^*h is not always invertible and the problem is often unstable (could have many wrong solutions of PSF h), λ is then chosen to regularize the problem. It is also necessary to remove the noise. To do this, for each pixel point(x,y) where $\nabla f(x,y) \neq 0$, the vector $T(x,y) = (\nabla f)/|\nabla f|$ in the gradient direction, and $\xi(x,y)$ in the orthogonal to $T(x,y)$, as shown in Fig. 5.2. With the usual notation f_x, f_y, f_{xx}, f_{yy} for the first

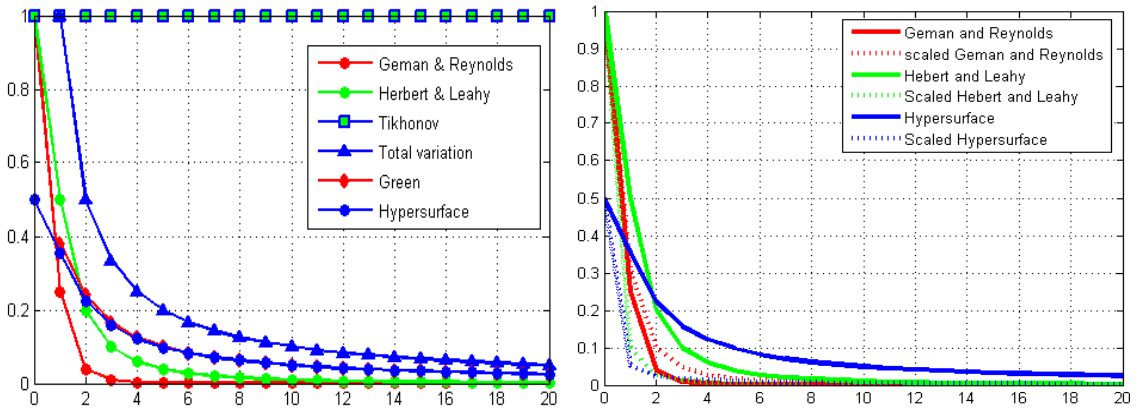


Figure 5.3: *a|b.* Convex and decreasing curves. (a) Functions $\phi'(t)/(2t)$ with different choice of ϕ. (b) Scaled Functions $\phi'(t)/(2st)$ with different choice of ϕ and scale s.

and second partial derivatives of f, and by formally developing the divergence operator, Eq. 5.7 can be formed in the following,

$$-\frac{\lambda}{2}\mathrm{div}\underbrace{\left(\frac{\phi'(|\nabla f|)}{|\nabla f|}\right)}_{\text{coefficient 1}} f_{\xi\xi} - \frac{\lambda}{2}\underbrace{\phi''(|\nabla f|)}_{\text{coefficient 2}} f_{TT} + h^* h f = h^* g \tag{5.8}$$

where $f_{\xi\xi}$ and f_{TT} ($f_{\xi\xi} + f_{TT} = \Delta f$) denote the second derivatives of f in the direction of $\xi(x, y)$ and $T(x, y)$, respectively. Through geometric analysis, we can get criteria of $\phi(\cdot)$ in most exiting variational methods [17] which achieve edge-preserving in convex optimization. It is also useful for determining how the function $\phi(\cdot)$ be chosen.

1. The local edge curve separates the region part into two homogeneous regions of the image. In the interior of the homogeneous regions $\{(x, y)|f(x, y) > 0\} \cup \{(x, y)|f(x, y) < 0\}$, where the variations of f are weak, smoothing is encouraged, $\phi'(0) = 0$ and $\phi''(0) > 0$ is supposed. The function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ is of class C^2 due to nonnegative constraints of images.

2. Normally, the variations of the intensity are weak at homogeneous regions in an image. Assuming that the function $\phi(\cdot)$ is regular, the *isotropic* smoothing condition can be achieved by imposing,

$$\phi'(0) = 0, \lim_{t\to 0^+} \frac{\phi'(t)}{t} = \lim_{t\to 0^+} \phi''(t) = \phi''(0) > 0 \tag{5.9}$$

Therefore, in this homogeneous regions, ∇f is small, Eq. 5.8 becomes

$$-\lambda\phi''(0)(f_{\xi\xi} + f_{TT}) + h^* h f = h^* g \tag{5.10}$$

since $f_{\xi\xi} + f_{TT} = \Delta f$, f locally satisfies the equation $-\lambda\phi''(0)\Delta f + h^* h f = h^* g$ in this region. It is a uniformly elliptic equation having strong regularizing properties in all directions.

3. In a neighborhood of an edge curve C, the image presents stronger gradients. For preserving the edge curves, it is preferable to diffuse along the the direction of ξ of the curve and not across it. To do this, it is sufficient to annihilate the coefficient of f_{TT} in Eq. 5.8, $\lim_{t\to+\infty} \phi''(t) = 0$ and keep the coefficient of $f_{\xi\xi}$ does not vanish: $\lim_{t\to+\infty} \frac{\phi'(t)}{t} > 0$. However, both conditions are incompatible, e.g., different weak intensities (different low gradients in one region), one must make a compromise between these two diffusions. The strategy is to make both coefficients converge to zero as $t \to +\infty$, but at different rate. The function is used $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ which has the properties

$$\lim_{t\to+\infty} \phi''(t) = \lim_{t\to+\infty} \frac{\phi'(t)}{t} = 0 \text{ and } \lim_{t\to+\infty} \frac{\phi''(t)}{\frac{\phi'(t)}{t}} = 0 \tag{5.11}$$

   Notice that many functions Φ in Table. 5.1 satisfying these conditions. These qualitative conditions have been imposed in order to describe the regularization conditions.

4. Furthermore, these conditions are not sufficient to ensure that the model is well posed. Other hypothesis such as convexity, and linear growth conditions are necessary to obtain the solution of well posed properties in calculus of variations.

We make a short summarization for Table. 5.1. This table presents different $\phi(t)$-functions that are commonly used. It is interesting to observe that some ϕ functions are non-convex. But their $\phi'(t)/2(t)$ exists in always convex and decreasing manner, shown in Fig. 5.3. This means that non-convex function can be converted to convex during the computation. For example, non-quadratic and non-convex functions to be minimized can be split into a sequence of half-quadratic problems that is convex and easier to solve numerically. Thus, we will be able to give a convergence result only for convex functions.

For example, in Fig. 5.3, we present six $\phi'(t)/(2t)$ curvatures of their ϕ functions including two non-convex functions Herbert and Leahy (H-L)[108] and Geman and Reynolds (G-R) [86]. Scaled G-R function takes $1/\left[(1+(3t/2)^2)^2\right]$, scaled H-L function takes $1/\left[1+(3t)^2\right]$, scaled hypersurface takes $1/\sqrt{1+(10t)^2}$. Scaled hypersurface minimal function is close to 0.1 for $t = 1$. It shows a better comparison from a numerical point of view. Different from Tikhonov function ϕ, the other ϕ are all edge-preserving functions. These ϕ satisfy the edge-preserving hypotheses $\lim_{t\to+\infty} \phi'(t) = 0$ and $\lim_{t\to+0^+} \phi'(t) = 0$. However, we call these methods are "passive" edge-preserving methods which is totally different from our proposed "active" data-driven methods in the BV space.

5.2 Description of Models in the BV Spaces

In this section, according to [58], [310], firstly, we discuss the basic properties of the Lebesgue measure and integration. The main reason for providing the review of Lebesgue measure and integration is to compare its difference with that of Riemann integration, and Hausdorff measure. The Hausdorff measure (in Appendix D) is not as well known as Lebesgue measure but yet is extremely important in geometric image analysis in the bounded variation spaces.

Secondly, we introduce the Sobolev space (also called Sobolev functions). This includes the discussion of continuity properties of functions with first derivatives in L^p in terms of Lebesgue

measure, as well as the higher order Sobolev functions by means of L^p-derivatives. While the Lebesgue theory for Sobolev functions is relatively straightforward, the corresponding development for BV functions is much more demanding. The intrinsic nature of BV functions requires a more involved exposition than does Sobolev functions.

Furthermore, we focus on the introduction of the functions of bounded variation and their characteristic properties in related functionals. A function of bounded variation of one variable can be characterized as an integrable function whose derivative in the sense of distributions is a signed measure with finite total variation. The multivariate analog of these functions is the class of L^1 functions whose partial derivatives are measures in the sense of distributions. Just as absolutely continuous functions form a subclass of BV functions, so it is that Sobolev functions are contained within the class of BV function of several variables. While the BV functions of one variable have a relatively simple structure that is easy to expose, the multivariate theory produces a rich and beautiful structure based on geometric measure theory. An interesting and important aspect of the geometric measure theory is the analysis of sets whose characteristic function are BV (called sets of finite perimeter). These sets have applications in a variety of settings because of their generality and utility.

Lastly, since the total variational functional has been introduced into image processing, the functions of bounded variation become more important. we discuss several recently developed linear-growth functionals in the $BV(\Omega)$ space and our proposed Bayesian estimation based double variational regularization functional.

5.2.1 Spaces of Functions and Lebesgue Integration

Before embarking the functions of bounded variation, we first introduce two important concepts of spaces (sets) of functions: (a) these are the spaces of *continuous functions* $C^m(\Omega)$ etc. in the space of $C(\Omega)$ etc. (b) the *Lebesgue spaces* L^p etc. whose pth powers are integrable in an open set of $\mathbb{R}$ or $\mathbb{R}^n$, shown in Fig. 5.4.

Continuous Functions

Roughly speaking, a continuous function of a single variable may be characterized as one whose graph is an uninterrupted curve. On the other hand, a function is discontinuous in that its graph has a break. Another type of discontinuous function is one that is unbounded at some point. Continuous functions are defined on a subset Ω of $\mathbb{R}^n$ and can be categorized:

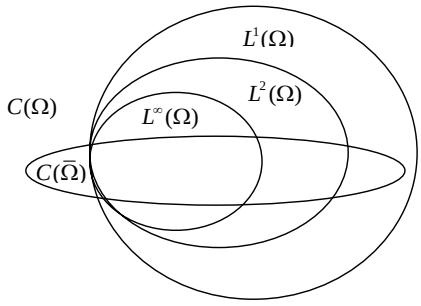


Figure 5.4: The relationship between the L^p spaces and spaces of continuous functions.

1. **The space** $C(\Omega)$. For any domain Ω in $\mathbb{R}^n$, the collection of all continuous functions defined on Ω forms a set, or space, which is denoted by $C(\Omega)$, shown in Fig. 5.4. The space of functions that are continuous on the closest set $\bar{\Omega} = \Omega \cup \Gamma$ (Γ and its boundary Γ) is denoted by $C(\bar{\Gamma})$ and by $C[a, b]$ for functions on the closed interval $[a, b]$.

2. **The spaces** $C^m(\Omega)$ **and** $C^\infty(\Omega)$. Among all the continuous functions defined on a subset Ω of $\mathbb{R}^n$, some have the properties that their first derivatives and possibly some derivatives of higher order are also continuous. It is very important to identify such functions with their derivatives (of order m) are continuous on Ω. That is, $C^m(\Omega) = \{u : u, \partial u/\partial x, \partial u/\partial y, ..., \partial^m u/\partial x^k \partial y^{m-k} (k = 0, ..., m)$ are all continuous functions$\}$ for $\Omega \subset \mathbb{R}^2$ and so on. Clearly, the inclusions $C^\infty(\Omega) \subset ... \subset C^m(\Omega) \subset C^{m-1}(\Omega) \subset ... \subset C^0(\Omega) = C(\Omega)$ hold, so that C^m constitutes a gradation which permits continuous functions to be classified according to their degree of smoothness: for any function in $C^m(\Omega)$, the higher the value of m, the smoother the function.

3. **Continuous functions on compact sets**. It turns out that continuous functions defined on compact sets (closed and bounded sets in $\mathbb{R}^n$) may be characterized necessarily on such bounded sets. A function f defined on a set Ω in $\mathbb{R}^n$ is said to be bounded if it is possible to find a number $M > 0$ such that $f(x) \leq M$ for all $x \in \Omega$. In other words, the function does not "blow up" anywhere. Continuous functions behave in a special way on compact sets, shown in the definition.

   Definition 5.2.1.1 *Let Ω be a bounded domain (that is, a bounded open, connected set) in $\mathbb{R}^n$, and f a continuous function defined on the compact set $\bar{\Omega}$. Then,* (**a**) *f is bounded on $\bar{\Omega}$ and, furthermore, f achieves its supremum and infimum on $\bar{\Omega}$.* (**b**) *f is uniformly continuous on $\bar{\Omega}$.*

   The definition shows the function has a maximum for a given point $z = \sup f(\Omega) = \max f(\Omega)$ for all points $x \in \Omega$. A similar interpretation applies with respect to the infimum.

4. **Lipschitz continuous functions**. A function f defined on a set Ω in $\mathbb{R}^n$ is said to be Lipschitz continuous or Lipschitz, if there exists a constant $L > 0$ such that $f(x) - f(y) \leq L|x - y|$ for all $x, y \in \Omega$. The definition of Lipschitz continuity does not require that the derivative exists at every point. It is straightforward to show that every Lipschitz function is uniformly continuous, although the converse is not true. If Ω is a compact set, then every continuously differentiable function on Ω is Lipschitz.

Measures of Sets in $\mathbb{R}^n$

However, many functions in practical applications are not continuous, and cannot therefore be accommodated in one of the spaces $C^m(\Omega)$, such as discontinuities, unconnected edges in images. A simple example is to use Heaviside step function,

$$H(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Although these functions are not continuous, they do possess the important property that they are integrable. Our aim is to set up a space of functions that may be classified according to their

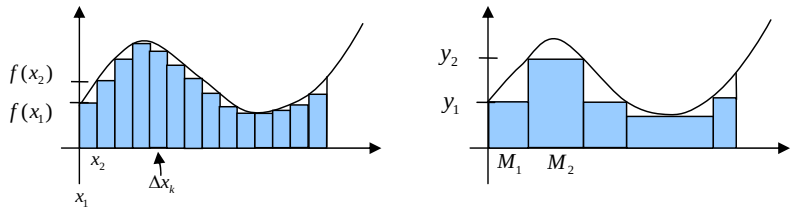


Figure 5.5: *a|b.* The basic idea behind (a) Riemann integration and (b) Lebesgue integration.

integrable powers, e.g., $\int_a^b |f(x)|^p dx$, $p \geq 1$. This permits the introduction of the space $L^p(a,b)$ or, generally, $L^p(\Omega)$. For such demands, the case of the spaces C^m is possible to obtain a precise idea of the degree of smoothness of a function by determining the largest value of m for which it belongs to C^m. The smoothness of two functions may then be compared by determining the largest numbers m of the spaces C^m of which they are members. In the same way, we will see that the L^p spaces are also *nested* in the sense of $L^p \subset L^q$ for the case in which $p > q$. Therefore, we note that these spaces also provide a means of comparing functions during the period of the integrability.

In order to give such spaces a proper treatment, it is necessary to introduce the notion of *Lebesgue measure*. This in turn allows us to introduce the notion of *Lebesgue integration*, which is a generalization of the "standard" Riemann integration. Then we define the spaces $L^p(\Omega)$. Lebesgue measure is an important measure method in the well-established measure theory in mathematics. In order to introduce the Lebesgue integral, we return first to the definition of the *Riemann integral*. The basic idea of the Riemann integral is to divide $[a,b]$ into a finite number N of subintervals, the kth subintervals having length Δx_k, and the approximation area under the graph f is the sum of the forms: $f(x_1)\Delta x_1 + f(x_2)\Delta x_2 + ... + f(x_N)\Delta x_N$, shown in Fig. 5.5(a). Thus, the Riemann integral is denoted by $\int_a^b f(x)dx$ which is used widely and adequate for most purposes.

However, the Riemann integral suffers from certain deficiencies, e.g., it is unable to deal with the function $f(x) = 1, x$ is rational (while $f(x) = 0, x$ is irrational) on the interval $[0,1]$. Secondly, in contrast to the Riemann integral, for a more general *Lebesgue integral*, the approximation to the integral of f can be progressively improved, not by further subdivisions of the domain, but by refining the approximation to f. The approximating functions that serve this purpose are indeed known as simple functions, and are defined to be functions that take on a finite number of values. Provided that with the subsets M_k on their *constant values*, the integral of f can be approximated by a sum of the form, $y_1\mu(M_1) + y_2\mu(M_2) + ... + y_N\mu(M_N)$, shown in Fig. 5.5(b), where $\mu(M_k)$ is a measure of M_k. The limit number N is a nice improvement for the approximation of f. Therefore, in this measurable space Ω, the Lebesgue measure is defined to satisfy those four criteria for measurable sets: (1) Ω itself; (2) $\Omega - M$, for $M \in \mathcal{M}$; (3) all open sets in Ω; and (4) $M_1 \cup M_2 ...$, for any countable family $\{M_1, M_2, ..., \}$ of disjoint sets in $\mathcal{M}$. Also, functions that are Riemann-integral are also Lebesgue-integral, and the two integrals coincide.

Lebesgue Integration and the Space $L^p(\Omega)$

We say that a function defined on a measurable set Ω in $\mathbb{R}$ is measurable if the inverse image $f^{-1}(M)$ of any measurable set M in $\mathbb{R}$ is itself measurable. Therefore, we can verify any

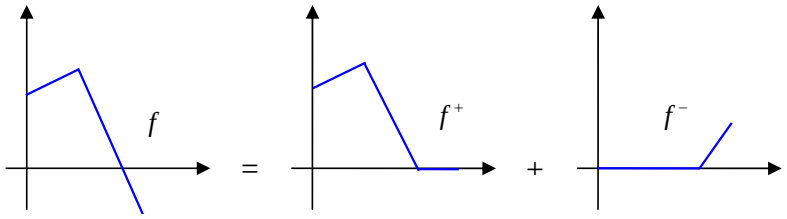


Figure 5.6: *a|b|c.* The positive and negative parts of a function in the Lebesgue integral.

continuous function is measurable. Heaviside function is a measurable function. Since sums of measurable functions are measurable, we can conclude that every step function is measurable. Based on the intuitively obvious character, the Lebesgue integral of a simple function s on Ω is defined by

$$\int_{\Omega} s dx = a_1\mu(M_1 \cap \Omega) + a_2\mu(M_2 \cap \Omega) + ... + a_N\mu(M_N \cap \Omega) \tag{5.12}$$

where M_k are measurable and pairwise disjoint. To obtain the Lebesgue integral of a measurable function f we first set up a sequence of nondecreasing simple functions that approximate f. Next, we evaluate the integrals of these simple functions and take the limit to obtain the integral of f. Of course, f is a nonnegative measurable function on $\mathbb{R}^n$ with a nondecreasing sequence S of simple functions on $\mathbb{R}^n$ such that $\lim_{n\to\infty} s_n(x) = f(x)$ at all points x in $\mathbb{R}^n$. Therefore, when f is a measurable function defined on a measurable set Ω and f is nonnegative on Ω, then the Lebesgue integral of f over Ω is defined by

$$\int_{\Omega} f dx = \lim_{k\to\infty} \int_{\Omega} s_k dx \tag{5.13}$$

where s_k are nondecreasing simple functions that approximate f. Indeed, for well-behaved functions, e.g., piecewise continuous functions: it is clear that the Lebesgue integral like the Riemann integral, amounts to the area under the graph of the function. However, there are Lebesgue-integrable functions which are not Riemann-integrable. To complete the theory of the Lebesgue integral, we extend the treatment to include functions that are not necessary nonnegative. Suppose that f is any measurable function. Then f may be decomposed into positive part $f^+(x) = f(x)$, if $f(x) \geq 0$ (otherwise $f(x) = 0$) and negative part $f^-(x) = 0$, if $f(x) \geq 0$ ($f^-(x) = -f(x)$, otherwise) shown in Fig. 5.6. More concisely, we can write $f^+ = \frac{1}{2}(f + |f|)$ and $f^- = \frac{1}{2}(|f| - f)$ so that $f = f^+ - f^-$. It is possible to show that f^+ and f^- are both measurable. The summable function $\int_\Omega f$ (Lebesgue integral exists) can be decomposed as the sum of two nonnegative functions,

$$\int_{\Omega} f dx = \int_{\Omega} f^+ dx - \int_{\Omega} f^- dx$$

Now we note that it is possible to have $\int_\Omega f dx = +\infty$ for a nonnegative function. This lemma is very useful in the function of bounded variation for image deblurring and denoising.

The space of Lebesgue integration $L^p(\Omega)$ is defined in an open set Ω. Let p be a real number with $p > 1$. A function $f(x)$ defined on a subset Ω of $\mathbb{R}^n$ is said to belong to $L^p(\Omega)$, if f is measurable

and if the (Lebesgue) integral $\int_\Omega |f(x)|^p dx$ exits, i.e., is finite. The case $p = 2$ is special in many ways, and is referred to as square-integrable. Therefore, every bounded continuous function defined on a bounded set Ω belongs to L^p. If we let $p \to \infty$, then we may define the space $L^\infty(\Omega)$ to be the space of all measurable functions on Ω that are bounded almost everywhere on Ω.

We note that although $L^\infty(\Omega) \subset ... \subset L^p(\Omega) \subset ... \subset L^1(\Omega)$, the space $C(\Omega)$ of continuous functions is not a subset of any of the L^p spaces, shown in Fig. 5.4. For example, the function $f(x) = x^{-1}$ belongs to $C(0,1)$ but not to $L^\infty(0,1)$ since it is not bounded. But the space of bounded continuous functions, equivalently, the space $C(\bar{\Omega})$ of continuous functions defined on a compact set $(\bar{\Omega})$ is a subset of $L^\infty(\Omega)$. Fig. 5.4 also shows schematically how the spaces $C^m(\Omega)$ and $L^p(\Omega)$ are related.

Distributions and Sobolev Spaces

For m an integer, $1 \leq p \leq \infty$ and $\Omega \subset \mathbb{R}$, we define the *Sobolev space*,

$$W^{m,p}(\Omega) \stackrel{def}{=} \{f \in L^p(\Omega);\, D^\alpha f \in L^p(\Omega),\, 0 \leq |\alpha| \leq m\} \tag{5.14}$$

where for $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_n\} \in \mathbb{N}^n$, we put the partial derivative

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_n^{\alpha_n}} \text{ and } |\alpha| = \alpha_1 + \alpha_2 + ... + \alpha_n \tag{5.15}$$

Thus if $|\alpha| = m$, then $D^\alpha f$ denotes one of the mth derivatives of f. The space is a normed space when endowed with the Sobolev norm $\|\cdot\|_{m,p}$, The *Banach space* for the norm becomes

$$\|f\|_{m,p} = \left(\sum_{0 \leq |\alpha| \leq m} |D^\alpha f|^p_{L^p(\Omega)} \right)^{1/p}, \; 1 \leq p < \infty \tag{5.16}$$

and in the case $p = \infty$

$$\|f\|_{m,\infty} = \max_{0 \leq |\alpha| \leq m} |D^\alpha f|_{L^\infty(\Omega)} \tag{5.17}$$

In the particular case $p = 2$, we have

$$W^{m,2}(\Omega) = H^m(\Omega) \tag{5.18}$$

where the Sobolev space $H^m(\Omega)$ has been defined by taking as a point of departure in the Hilbert space $L^2(\Omega)$. The results concerning the spaces $W^{m,p}(\Omega)$ are analogous to that obtained for the space $H^m(\Omega)$. The definition of the spaces $W^{s,p}(\Omega)$ for non-integral values of s, can be given by interpolation between $L^p(\Omega)$ and $W^{m,p}(\Omega)$. Fig. 5.7 shows the relationship between Hilbert spaces and Banach spaces, and the others. The theory of Hilbert spaces forces the use of the Lebesgue integral. As we have discussed previously, the widely used Riemann integral is only valid under very *restrictive* assumptions, in contrast to the Lebesgue integral. The Riemann integral leads only to pre-Hilbert spaces for which the fundamental Cauchy criterion is not valid.

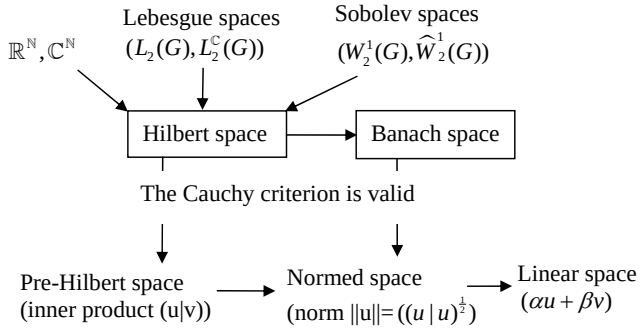


Figure 5.7: The relationship between Hilbert spaces and Banach spaces, and others. Each Hilbert space is a Banach space; the most important Hilbert spaces are from the Lebesgue spaces $L_2(G)$, $L_2^C(G)$ and the related Sobolev spaces $W_2^1(G)$ and $\widehat{W}_2^1(G)$. Roughly speaking, the real Lebesgue space $L_2(G)$ (resp. the complex Lebesgue space $L_2^C(G)$) consists of all functions. The theory of Hilbert spaces forces the use of the Lebesgue integral.

1. Real Lebesgue space $L_2(G)$ is applied in Fourier series, integral equations, and partial differential equations.
2. Sobolev spaces $W_2^1(G)$ and $\widehat{W}_2^1(G)$ are applied mainly in Dirichlet principle and the calculus of variations.
3. Complex Lebesgue space $L_2^c(\mathbb{R}^N)$ is applied mainly in quantum mechanics and Fourier transformation.

First of all, there is a category of Sobolev spaces that are Hilbert spaces. In *Hilbert spaces*, an inner product $(u|v)$ is defined, allowing us to introduce the fundamental notion of orthogonality. The *Sobolev spaces* provide a very natural setting for boundary value problems. This *Banach space* is a complete *normed space*, while Hilbert space is complete inner product space. Since every inner product defines a norm, every Hilbert space is a Banach space. Second, it is possible to obtain quite general results regarding existence and uniqueness of solutions in a variational setting, using these spaces. A third advantage is that, like the space $C^m(\Omega)$, Sobolev spaces provide a means of characterizing the degree of smoothness of functions. Finally, perhaps most important, is the fact that approximate solution methods such as the Galerkin and finite element methods. These methods are most conveniently and correctly formulated in finite-dimensional subspaces of Sobolev spaces.

5.2.2 The Space of Functions of Bounded Variation

The idea of function of bounded variation (BV) developed along different streams, both in an analytical and in a geometrical vein. From the classical analysis, BV functions were singled out as a possible control on the oscillations and suitable insurance of the convergence of the Fourier series. The functions of BV have been firstly introduced by C. Jordan in 1881 in connection with Dirchlet's test for the convergence of Fourier series [11]. The geometric counterpart is that rectifiable curves, i.e., images of continuous parametrization with finite length, can be precisely parameterized by BV functions.

However, it is not so easy to describe the situation of functions of several variables. Many attempts have been needed to clarify the links between the possible extensions of the concepts

of variation of a function and the finiteness of the area of its graph [11]. Fortunately, another point of view came to the fore in connection with Schwartz'distribution theory, leading to a definition of BV functions in terms of distributional derivatives. Here, the understanding of the definition BV functions can be based on the definition of total finite variation and its distributional gradients. The more properties in the BV spaces are discussed in the following subsection of the convex linear-growth functional.

Normed Linear Vector Space and TV

The vector spaces of particular interest in both abstract analysis and applications have a good deal more structure than that implied solely by the seven principle axioms such as commutative law, associative law, distributive law and so on. The vector space axioms only describe algebraic properties of the elements of the space: addition, scalar multiplication, and combinations of these. What are missing are the topological concepts such as openness, closure, convergence, and completeness. These concepts can be provided by the introduction of a measure of distance in a normed linear vector space[151].

Definition 5.2.2.1 *A normed linear vector space is a vector space X on which is defined a real-valued function which maps each element x in X into a real number $\|x\|$ called the norm of x. The norm satisfies the following axioms:*

1. $\|x\| \geq 0$ *for all* $x \in X$, $\|x\| = 0$ *if and only if* $x = 0$.
2. $\|x + y\| \leq \|x\| + \|y\|$ *for each* $x, y \in X$, *it is also triangle inequality.*
3. $\|\alpha x\| = |\alpha| \cdot \|x\|$ *for all scalars* α *and each* $x \in X$.

The norm is clearly an abstraction of the usual concept of length. Based on the normed linear space, the function of bounded variation (BV) is one of useful consequences of the triangle inequality. The space $BV[a, b]$ consists of functions of bounded variation on the interval $[a, b]$. By a partition of the interval $[a, b]$, we mean a finite set of points $t_i \in [a, b]$, $i = 0, 1, 2, ..., n$, such that $a = t_0 < t_1 < t_2 < ... < t_n = b$. A function x defined on $[a, b]$ is said to be of bounded variation if there is a constant K so that for any partition of $[a, b]$

$$\sum_{i=1}^{n} |x(t_i) - x(t_{i-1})| \leq K. \tag{5.19}$$

The total variation of x is defined as

$$TV(x) = \sup \sum_{i=1}^{n} |x(t_i) - x(t_{i-1})| \tag{5.20}$$

where the supremum is taken with respect to all partitions of $[a, b]$. A convenient and suggestive notation for the total variation is

$$TV(x) = \int_a^b |Dx(t)| \tag{5.21}$$

The total variation of a constant function is zero and the total variation of a monotonic function is the absolute value of the difference between the function values at the end point a and b. The $BV[a,b]$ space is defined as the space of all functions of bounded variation on $[a,b]$ together with the norm defined in the following,

$$\|x\| = |x(a)| + TV(x) \tag{5.22}$$

In general, let Ω be a bounded open subset of $\mathbb{R}^N$, $N = 1, 2, 3, ...$, whose boundary $\partial\Omega$ is Lipschitz continuous. The Euclidean norm on $\mathbb{R}^N$ is $|x| = \sqrt{\sum_{i=1}^{N} x_i^2}$. The norm on the Banach spaces $L^p(\Omega)$ is denoted by $\|\cdot\|_{L^p(\Omega)}$, $1 \leq p \leq \infty$. Let Ω denote the Lebesgue measure of Ω. Let an image f be a function in $L^1(\Omega)$, we set

$$\begin{aligned} TV(f) &= \int_\Omega |Df| dx \\ &= \sup\left\{ \int_\Omega f \cdot \mathrm{div}\varphi dx \ ; \ \varphi \in C_0^1(\Omega), \Omega \in \mathcal{R}^n \text{ and } |\varphi(x)|_{L^\infty(\Omega)} \leq 1 \right\} \end{aligned} \tag{5.23}$$

where $d\varphi = \sum_{i=1}^{N} \frac{\partial \varphi_i}{\partial x_i}(x)$, dx is the Lebesgue measure, and $C_0^1(\Omega)$ is the space of continuously differentiable function with compact support in Ω. The inequality $|\varphi(x)|_{L^\infty(\Omega)} \leq 1$ means that all the components of the vector-values function φ have a $L^\infty(\Omega)$-norm less than one. If $f \in C^1(\Omega)$, then $\int_\Omega f \cdot \mathrm{div}\varphi dx = -\int_\Omega \nabla f \cdot \varphi dx$ and $\int_\Omega |Df| dx = \int_\Omega |\nabla f(x) dx$. By a standard denseness argument, this also applies for f in the Sobolev space $W^{1,1}(\Omega)$. The equation is similar to Eq. 5.4 which is a special case in the $BV(\Omega)$ space.

Functions of Bounded Variation

Unlike Sobolev spaces, one of the main advantages of the BV space is that includes characteristic functions of *sufficiently regular sets* and *piecewise smooth functions* (more generally) [11]. The space of functions of bounded variation on Ω is defined by

$$BV(\Omega) = \{f \in L^1(\Omega); TV(f) < \infty\} \tag{5.24}$$

The BV norm is given by

$$\|f\|_{BV} = \|f\|_{L^1(\Omega)} + TV(f) \tag{5.25}$$

BV is complete with respect to this norm, and hence a Banach space. The Sobolev space $W^{1,1}(\Omega)$ is a proper subset of $BV(\Omega)$. Note that for Ω bounded, $L^p(\Omega) \subset L^1(\Omega)$ for $p > 1$. For the definition, $BV(\Omega) \subset L^1(\Omega)$. It is shown below that $BV(\Omega) \subset L^p(\Omega)$ for $1 \leq p \leq N/(N-1)$.

We define $BV(\Omega)$ in the image domain, the space of functions of bounded variation,

$$BV(\Omega) = \left\{ f \in L^1(\Omega); \ \int_\Omega |Df| < \infty \right\}. \tag{5.26}$$

If $f \in BV(\Omega)$, then Df (the distributional gradient of f; notice that in this case Df is a function and we can also denote it ∇f) can be identified to be a Radon vector-valued measure.

In the next section, we are going to show that Df can be decomposed as the sum of a regular measure and a singular measure with the Hausdorff measure in the BV space.

5.2.3 Convex Linear-Growth Functional

Following Rudin, Osher and Fatemi [213], Chambolle and Lions [39], Weickert and Schnörr [267], Chan et al. [41], Aubert and Vese [17], [249] etc., we study the total variation functional in the Bounded Variations (BV) space. The TV functional is strictly convex and is lower semicontinuous with respect to the weak-star topology of BV. Therefore, the minimum exists and is unique. The decomposition of the TV model heavily depends on the specific norm which is chosen on BV. However, the TV functional is a special example of a more general smoothing algorithm [164].

We relax the TV functional to a more general convex functional in the space $BV(\Omega)$ where$|Df| \rightarrow \phi(|Df|)$. Let Ω be an open, bounded, and connected subset of $\mathbb{R}^n$ and the Lipschitz boundary Γ. We use standard notations for the Sobolev $W^{1,p}(\Omega)$ and Lebesgue spaces $L^p(\Omega)$. A variational function can be written in the form,

$$\mathcal{J}(f_{(g,h)}) = \frac{1}{2}\int_{\Omega}(g-hf)^2 dA + \lambda \int_{\Omega}\phi(Df(x,y))dxdy \tag{5.27}$$

where the function $\int_{\Omega}\phi(Df)dxdy$ is finite on the space $W^{1,1}$ which is a nonreflexive Banach space. Nonreflexive property obviously does not satisfy the boundary conditions. As a result, the minimization of $\int_{\Omega}\phi(Df)$ may not have a solution. On the other side, we can observe the importance of reflexivity. For these reasons, functions of bounded variation, the notions of convex functions of measures and relaxed functionals on measures are used to obtain the existence of a minimum. Furthermore, the space of BV-functions is the proper class for many basic image processing tasks, because it allows discontinuities along or across the curves or edges.

On the $BV(\Omega)$ space, we recall the notation of lower semicontinuity of functionals defined on this space. We denote by $\mathcal{L}_N$ the Lebesgue N-dimensional measure $\mathbb{R}^N$ and by $\mathcal{H}^{\alpha}$ the α-dimensional Hausdorff measure. We say that $f \in L^1(\Omega)$ is a function of bounded variation ($f \in BV(\Omega)$) if its distributed derivative $Df = (D_1 f, ..., D_n f)$ belongs to a weak topology on $\mathcal{M}(\Omega)$. $\mathcal{M}(\Omega)$ is the set of all signed measures on Ω with bounded total variation. Furthermore, the space $BV(\Omega)$ endowed with the norm,

$$\|f\|_{BV(\Omega)} = \|f\|_{L^1(\Omega)} + |Df|(\Omega)$$

is a Banach space. The product topology of the strong topology of $L^1(\Omega)$ for f and of the $weak*$topology of measures for Df is called the *weak*topology* of BV [66],

For any function $f \in L^1(\Omega)$, we denote by S_f the complement of the Lebesgue set of f. Normally, the set S_f is of zero Lebesgue measure and is also called the *jump set of* f. If $f \in BV(\Omega)$, then f is differentiable almost everywhere on $\Omega \setminus S_f$. Moreover, the Hausdorff dimension of S_f is at most $(N-1)$ and for $\mathcal{H}^{N-1}$, $x \in S_f$ it is possible to find unique $f^+(x), f^-(x) \in \mathbb{R}$, with $f^+(x) > f^-(x)$ and $\nu \in S^{n-1}$ of unit sphere in $\mathbb{R}^n$, such that

$$\lim_{r\rightarrow 0^+} r^{-N}\int_{B_r^{\nu}(x)}|f(y)-f^+(y)|dy = \lim_{r\rightarrow 0^+} r^{-N}\int_{B_r^{\nu}(x)}|f(y)-f^-(y)|dy = 0 \tag{5.28}$$

where $B_r^{\nu}(x) = \{y \in B_r(x)\}|(y-x)\cdot\nu > 0$ and $B_r^{-\nu}(x) = \{y \in B_r(x)\}|(y-x)\cdot\nu < 0$. The normal ν means that they points toward the larger value in the image f. We denote by $B_r(x)$

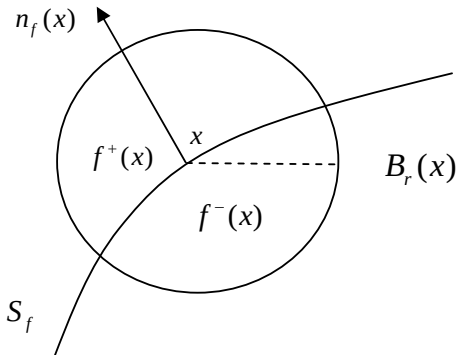


Figure 5.8: Definition of f^+, f^-, and the jump set S_f in the BV space. $B_r(x)$ be the ball of center x and radius r. f^+ and f^- is the positive part $f \vee 0$ and negative part $-(f \wedge 0)$ of f.

the ball centered in x of radius r, shown in Fig. 5.8. The detailed definition of Hausdorff measure has been described in Appendix D.

We have the Lebesgue decomposition,

$$Df = \nabla f \cdot \mathcal{L}_N + D^s f \tag{5.29}$$

where $\nabla f \in (L^1(\Omega))^N$ is the Radon-Nikodym derivative of Df with respect to $\mathcal{L}_N$. Generally, by the Radon-Nikodym theorem we set $Df = D^a f + D^s f$ where $D^a f \ll \mathcal{L}_N$ is the absolutely continuous part of Df with respect to the Lebesgue measure, and $D^s f$ is singular part of Df with respect to $\mathcal{L}_N$. In other words, ∇f is the density of the absolutely continuous part of Df with respect to the Lebesgue measure. We also have the decomposition for $D^s f$,

$$D^s f = C_f + J_f, \tag{5.30}$$

where

$$J_f = (f^+ - f^-) n_f \cdot \mathcal{H}_S^{n-1}$$

is *Hausdorff part* or *jump part* and C_f is the *Cantor part* of Df. The measure C_f is singular with respect to $\mathcal{L}_N$ and it is diffuse, that is, $C_f(S) = 0$ for every set S of Hausdorff dimension $N-1$. $\mathcal{H}^{N-1}_{|S_f}$ is called the perimeter of related edges in Ω. Finally, we can write Df and its total variation on Ω, $|Df|(\Omega)$, in the following,

$$Df = \nabla f \cdot \mathcal{L}_N + C_f + (f^+ - f^-)\nu \cdot \mathcal{H}^{N-1}_{|S_f} \tag{5.31}$$

$$|Df|(\Omega) = \underbrace{\int_\Omega |\nabla f| dx}_{\text{Lebesgue measure}} + \underbrace{\int_{\Omega \setminus S_f} |C_f|}_{\text{Cantor measure}} + \underbrace{\int_{S_f} (f^+ - f^-) d\mathcal{H}^{N-1}_{|S_f}}_{\text{Hausdorff part}} \tag{5.32}$$

It is then possible to define the convex function of measure $\phi(|\cdot|)$ on $\mathcal{M}(\Omega)$, which is for Df,

$$\phi(|Df|) = \phi(|\nabla f|) \cdot \mathcal{L}_N + \phi^\infty(1)|D^s f|, \tag{5.33}$$

and the functional following [93],

$$\int_{\Omega}\phi(|Df|)=\int_{\Omega}\phi(|\nabla f|)dx+\int_{\Omega}\phi^{\infty}(1)|D^{s}f|, \tag{5.34}$$

where the functional $\phi(|\cdot|)(\Omega)$ is proved in weakly*lower semi-continuous on $\mathcal{M}(\Omega)$. That is to say that $\int_{\Omega}\phi(|Df|)$ is convex on $BV(\Omega)$, ϕ is convex and increasing on $\mathbb{R}^{+}$.

By the decomposition of $D^{s}f$, the properties of C_f, J_f, and the definition of the constant c, the functional $\int_{\Omega}\phi(|Df|)$ can be written as,

$$\int_{\Omega}\phi(|Df|)=\int_{\Omega}\phi(|\nabla f|)dx+c\int_{\Omega\backslash S_f}|C_f|+c\int_{S_f}(f^{+}-f^{-})d\mathcal{H}^{N-1}_{|S_f}, \tag{5.35}$$

Based on this equation, Vese [249] proposed an energy functional for image deblurring and denoising in the BV space,

$$\inf_{f\in BV(\Omega)}\mathcal{J}(f_{(g,h)})=\frac{1}{2}\int_{\Omega}(g-hf)^{2}dxdy+\lambda\int_{\Omega}\phi(|Df(x,y)|)dxdy \tag{5.36}$$

where

$$\int_{\Omega}\phi(|Df(x,y)|)dxdy=\int_{\Omega}|\nabla f|dxdy+c\int_{\Omega\backslash S_f}|C_f|+c\int_{S_f}(f^{+}-f^{-})d\mathcal{H}^{N-1}_{|S_f}$$

Although some characterization of the solution is possible in the distributional sense, it remains difficult to handle numerically. To circumvent the problem, Vese [249] approximate the BV solution by Sobelev functions, using the notion of Γ-convergence which is also an approximation for the well-known Mumford-Shah functional [173]. The Mumford-Shah functional [173] and its extended Mumford-Shah functional [291] have similar underlying mathematic concepts with the variational energy modeling in the BV spaces.

The target of studying these functionals in the BV space is to understand and deduce a more general variation functional. In the following section, we study a more general variable exponent L^{p} linear growth functional $\phi(|Df(x,y)|)\rightarrow\phi(x,Df(x,y))$, which is a deductive functional in the BV space.

5.2.4 Convex Linear-Growth Variable Exponent Functional

The fundamental goal is to find $\hat{f}$ given an observed image g and an estimated PSF $\hat{h}$ by minimizing the image cost function. In the image domain, the cost function can be minimized according to the following formulation,

$$\mathcal{J}(f_{(g,h)})=\frac{1}{2}\int_{\Omega}(g-h*f)^{2}dxdy+\lambda\min_{f\in BV_{B}\cap L^{2}(\Omega)}\int_{\Omega}\phi(x,Df)dxdy \tag{5.37}$$

where $p(g|\hat{f},\hat{h})\propto\exp\left\{-\frac{\alpha_1}{2}\int_{\Omega}(g-\hat{h}*\hat{f})^{2}dxdy\right\}$ and $p(\hat{f})$ is extended to a nonlinear diffusion functional with variable exponent [48]. Edge-driven piecewise smoothing can be considered

as *a priori* knowledge for the estimation of image, then $p(\hat{f}) \propto \exp\Big\{-\int_\Omega \phi(x, D\hat{f})dxdy\Big\}$. $BV_B(\Omega) := \{f \in BV(\Omega) | f = B \text{ on } \partial\Omega\}$, and its associated flow,

$$-\frac{\lambda}{2}\text{div}(\phi_r(x, Df)) + h^*hf = h^*g \text{ , in } \partial\Omega \tag{5.38}$$

where $f(x,t) = B(x)$, on $\partial\Omega^\top$, and $f(0) = g$, in Ω. $\Omega^\top := \Omega \times [0,T]$ and $\partial\Omega^\top := \partial\Omega \times [0,T]$. This very general case where the functional has a variable exponent and $\phi = \phi(x, Df)$ are proved according to Chen, Levine and Rao [48], [49], Chan et al. [41], [107]. We integrate this more general variable exponent L^p, linear growth convex $\phi = \phi(x, Df)$ function in the bounded variations (BV) space to our double variational regularization. More related work on linear growth functionals an their flows in [21] and alternate variational approach [39] for reducing stair-casing by minimizing second order functionals.

For the definition of a convex function of measures, we refer to the works of Goffman-Serrin [93] Demengel-Temam [59], and Aubert [15]. Therefore, based on the previous analysis for $f \in BV(\Omega)$ space, we have,

$$\int_\Omega \phi(x, Df)dxdy = \int_\Omega \phi(x, \nabla f)dxdy + \int_\Omega |D^s f|dxdy \tag{5.39}$$

where

$$\phi(x, \nabla\hat{f})dxdy = \begin{cases} \frac{1}{q(x)}|\nabla\hat{f}|^{q(x)}, & |\nabla\hat{f}| < \beta \\ |\nabla\hat{f}| - \frac{\beta q(x) - \beta^{q(x)}}{q(x)}, & |\nabla\hat{f}| \geq \beta \end{cases} \tag{5.40}$$

where $\beta > 0$ is fixed, and $1 \leq q(x) \leq 2$. The term $q(x)$ is chosen as $q(x) = 1 + \frac{1}{1+k|\nabla G_\sigma * I(x)|^2}$ based on the edge gradients shown in Fig. 5.9, $I(x)$ is the observed image $g(x)$, $G_\sigma(x) = \frac{1}{\sigma}\exp[-|x|^2/(2\sigma^2)]$ is a Gaussian filter. $k > 0$, $\sigma > 0$ are fixed parameters. The detailed proof in the functions of BV space is available in appendix B.

The main benefit of this equation is that the local image information are computed as prior information for guiding image diffusion. As we have presented previously, TV-based diffusion is a stronger noise smoothing method but it can not performs "smoothly" for the homogeneous regions and the regions that have weak discontinuities. This filter integrates TV-based filter (L^p, $p = 1$), Gaussian filter, Laplace filter (L^p, $p = 2$) and continuous (L^p, $0 \leq p < 1 \cup 1 < p < 2$) filters in the BV space.

We extend this functional into Bayesian estimation based double variational regularization for simultaneous image deblurring and denoising. There are several main differences between this approach (TV based approach) and the suggested approach in the last chapter(Tikhonov based approach). Firstly, different from the Tikhonov based double regularization approach in the last chapter, the proposed approach in this chapter is extended from the total variational functional in the BV space. It is an "active" data-driven variational image diffusion and variational image restoration approach. Second, this variational functional has similar convexity properties with the total variational functional. Third, this approach has more advantages for image denoising than the Tikhonov based regularization approaches. Furthermore, in this approach, we focus on perceptual and high-fidelity image restoration. Finally, although our approach also uses the alternating minimization method for the solution. It mainly focus on PDE based numerical approximation in the spatial domain for image processing.

5.3 Bayesian Data-Driven Variational Image Deblurring and Denoising

From Bayesian point of view, we get the joint regularization for the estimation of image and PSF. The resulting method attempts to minimize double cost functions subject to constraints such as non-negativity conditions of the image and energy preservation of PSFs. The objective of the convergence is to minimize double cost functions by combing the energy function of the estimation of PSFs and images. Following a Bayesian paradigm, the ideal image f, the PSF h and an observed image g fulfill

$$P(f,h|g) = \frac{p(g|f,h)P(f,h)}{p(g)} \propto p(g|f,h)P(f,h) \tag{5.41}$$

Based on this form, our goal is to find the optimal $\hat{f}$ and $\hat{h}$ that maximize the posterior $p(f,h|g)$. $\mathcal{J}(f|h,g) = -\log\{p(g|f,h)P(f)\}$ and $\mathcal{J}(h|f,g) = -\log\{p(g|f,\ h)P(h)\}$ express that the energy cost $\mathcal{J}$ is equivalent to the negative log-likelihood of the data.

The proposed double variational regularization functional in a Bayesian framework in the BV spaces is formulated according to

$$\mathcal{J}(\hat{f},\hat{h}) = \underbrace{\int_\Omega (g-\hat{h}*\hat{f})^2 dxdy}_{\text{fidelity Term}} + \lambda \underbrace{\int_\Omega \phi(x,D\hat{f})dxdy}_{\text{image Smoothing}} + \beta \underbrace{\int_\Omega |\nabla\hat{h}|dxdy}_{\text{psf Smoothing}} \tag{5.42}$$

where $dxdy = dxdy$. The estimates of the ideal image f and the PSF h are denoted by $\hat{f}$ and $\hat{h}$, respectively, which can be iteratively alternating minimized (AM) [289]. The image smoothing term is a variable exponent, nonlinear diffusion term [48]. The PSF smoothing term represents the regularization of blur kernels.

5.3.1 Alternating Minimization of PSF and Image Energy

During the numerical computation, we compute ∇f instead of Df. Furthermore, we remove the singularity when $|\nabla f| = 0$, by approximating $\mathcal{J}(f)$ by $\mathcal{J}_\varepsilon(f)$ with $\varepsilon > 0$ a small parameter. Although the most common algorithm has been based on the *lagged-diffusivity* technique [39], [43], [250] using an iterative procedure, we can also use it for solving the denoising and deblurring respectively. Therefore, the data-driven variant exponent, linear growth div operator becomes,

$$\operatorname{div}\left(\phi\left(x,\sqrt{\varepsilon^2+|\nabla\hat{f}|^2}\right)\right) = \operatorname{div}(\phi(x,\nabla\hat{f})) \tag{5.43}$$

As we have discussed in the previous chapter, the scale problem between the minimization of the PSF and the image is avoided using the alternate minimization approach. We propose to solve the joint regularization equations in an alternate minimization approach with decreased complexity. The formulation is derived from Eq. (5.42) in the following:

$$\inf_{f\in BV(\Omega)} \mathcal{J}_\varepsilon(\hat{f},\hat{h}) = \tfrac{1}{2}\int_\Omega (g-\hat{h}*\hat{f})^2 dxdy + \lambda\int_\Omega \phi_\varepsilon(x,\nabla\hat{f})dxdy + \beta\int_\Omega (\nabla\hat{h})dxdy \tag{5.44}$$

This equation is in strictly convexity. We can also solve this equation in the alternating minimization (AM) algorithm for the *augmented* energy using two partial differential equations with respect to the image f and PSF. Furthermore, the lagged-diffusivity algorithm corresponds to exactly the AM algorithm for the augmented energy with $f^{(n)} \rightarrow H^{(n)} \rightarrow f^{(n+1)}$. This algorithm is used the continuous section for achieving blur identification, and data-driven image restoration. The two equations derived from Eq. (5.44) are using finite differences which approximate the flow of the Euler-Lagrange equation associated with it,

$$\partial \mathcal{J}_\varepsilon / \partial \hat{f} = \alpha_1 \hat{h}(-x,-y) * (\hat{h} * \hat{f} - g) - \lambda \mathrm{div}\left(\phi(x, \nabla \hat{f})\right) \tag{5.45}$$

$$\partial \mathcal{J}_\varepsilon / \partial \hat{h} = \alpha_2 \hat{f}(-x,-y) * (\hat{f} * \hat{h} - g) - \beta \nabla \hat{h} \cdot \mathrm{div}\left(\frac{\nabla \hat{h}}{|\nabla \hat{h}|}\right) \tag{5.46}$$

In the alternate minimization, blur identification including deconvolution, and image restoration including denoising are processed alternatingly for the estimation of the image and the PSF. The partially recovered PSF is the prior for the next iterative image restoration and vice versa. The algorithm is described in the following:

`Initialization:` $g(x) = g(x), h_0(x)$ `is random numbers`
`while` $(nmse > threshold$
(1). nth `it.` $\hat{f}_n(x) = \arg\min(\hat{f}_n | \hat{h}_{n-1}, g)$, `fix` $\hat{h}_{n-1}(x)$
(2). $(n+1)$th `it.` $\hat{h}_{n+1} = \arg\min(\hat{h}_{n+1} | \hat{f}_n, g)$, `fix` $\hat{f}_n(x)$, $h(x) > 0$
`end`

The data-driven diffusion term in Eq.5.46 is numerically approximated in the following,

$$\begin{aligned} &\mathrm{div}(\phi(x, \nabla \hat{f})) = \\ &\underbrace{|\nabla \hat{f}|^{p(x)-2}}_{\text{Coefficient}} [\underbrace{(p(x)-1)\Delta \hat{f}}_{\text{IsotropicTerm}} + \underbrace{(2-p(x))|\nabla \hat{f}| div(\frac{\nabla \hat{f}}{|\nabla \hat{f}|})}_{\text{CurvatureTerm}} + \underbrace{\nabla p \cdot \nabla \hat{f} \log |\nabla \hat{f}|}_{\text{HyperbolicTerm}}] \\ &\text{with} \\ &p(x) = \begin{cases} q(x) \equiv 1 + \frac{1}{1+k|\nabla G_\sigma * I(x)|^2}, & |\nabla \hat{f}| < \beta \\ 1, & |\nabla \hat{f}| \geq \beta \end{cases} \end{aligned} \tag{5.47}$$

We indicate with div the divergence operator, and with ∇ and Δ respectively the gradient and Laplacian operators, with respect to the space variables. The Neumann boundary condition [2] $\frac{\partial \hat{f}}{\partial N}(x,t) = 0$ on $\partial\Omega \times [0,T]$ and the initial condition$\hat{f}(x,0) = f_0(x) = g$ in Ω are used, where n is the direction perpendicular to the boundary, g is the observed image. The numerical implementation of the nonlinear diffusion operator is based on *central differences* for coefficient and the isotropic term, *minmod scheme* for the curvature term, and *upwind finite difference scheme* in the seminal work of Osher and Sethian for curve evolution [213] of the hyperbolic term based on the hyperbolic conservation laws. We use here the minmod function, in order to reduce the oscillations and to get the correct values of derivatives in the case of local maxima and minima.

The image is restored by denoising in the process of edge-driven image diffusion as well as deblurring in the process of image deconvolution. Firstly, the chosen variable exponent of $p(x)$

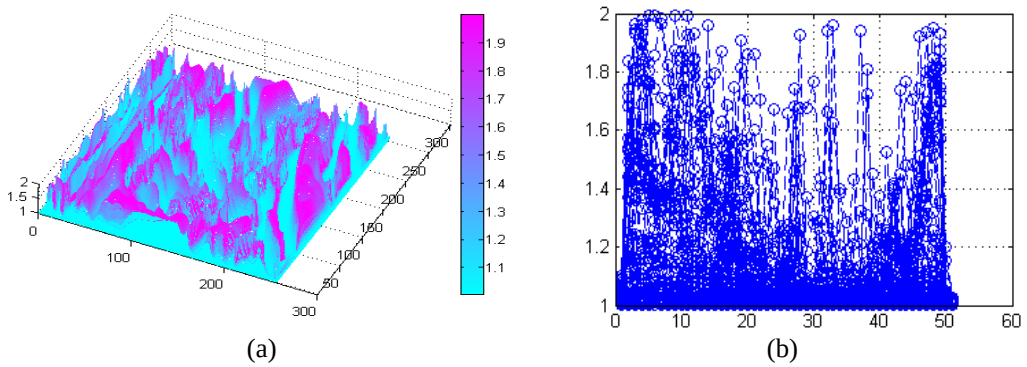


Figure 5.9: Strength of $p(x)$ in the Lena image. (a) Strength of $p(x)$ between [1,2] in the Lena image. (b) Strength of $p(x)$ is shown in a cropped image with size $[50, 50]$.

is based on the computation of gradient edges in the image. In homogeneous flat regions, the differences of intensity between neighboring pixels are small; then the gradient ∇G_σ become smaller ($p(x) \rightarrow 2$). The isotropic diffusion operator (Laplace) is used in such regions. In non-homogeneous regions (near edge or discontinuity), the anisotropic diffusion filter is chosen continuously based on the gradient values ($1 < p(x) < 2$) of edges. The reason is that the discrete chosen anisotropic operators will hamper the recovery of edges [177]. Secondly, the nonlinear diffusion operator for piecewise image smoothing is processed during image deconvolution based on a previously estimated PSF. Finally, coupling estimation of PSF (deconvolution) and estimation of image (edge-driven piecewise smoothing) are alternately optimized applying a stopping criteria. Hence, over-regularization or under-regularization is avoided by pixels at the boundary of the restored image.

5.3.2 Self-Adjusting Regularization Parameter

We have classified the regularization parameters λ in three different levels. Here, we present the method for the selection of window-based regularization parameters λ_w (window w based λ_w, 1st level). When the window size is amplified to the size of the input image, λ becomes a scale regularization parameter for the whole image (2nd level). If we fix λ for the whole process, then the selection of regularization parameter is conducted on the level of one fixed λ for the whole process (3nd level). We assume that the noise is approximated by an additive white Gaussian noise with standard deviation σ to construct a window-based local variance estimation. Then we focus on the adjustment of parameter λ and the operators in the smoothing term ϕ. These two computed components can be prior knowledge for preserving discontinuities and detailed textures during the image restoration. The Eq. 5.44 can be formulated in the following,

$$\arg\min \int_\Omega \phi(x, Df(x,y))dxdy \text{ subject to } \int_\Omega (g - hf)^2 dxdy$$

where the noise is Gaussian distributed with variance σ^2. λ can be a Lagrange multiplier in the following form,

$$\lambda = \frac{1}{\sigma^2 |\Omega|} \int_\Omega \text{div}\left[\phi(x, Df(x,y))\right](g - hf)dxdy \tag{5.48}$$

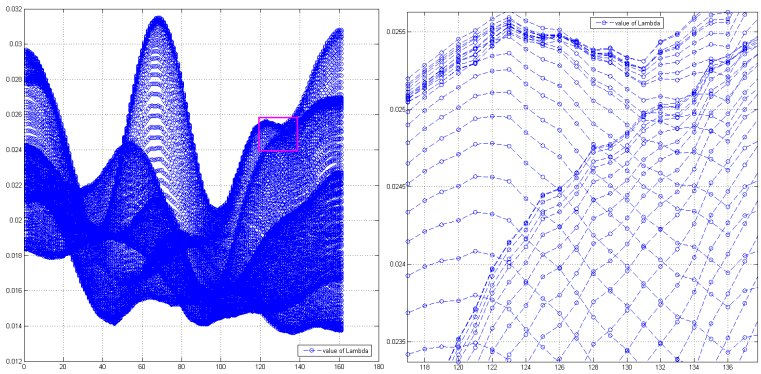


Figure 5.10: *a|b.* (a) Computed $\lambda \in [0.012, 0.028]$ values in sampling windows for the image with size [160, 160]. (b) Zoom in (a) for showing the distribution of the regularization parameters λ_w.

λ is a regularization parameter controlling the "balance" between the fidelity term and the penalty term. The underlying assumption of this functional satisfies $\|f\|_{BV(\Omega)} = \|f\|_{L^1(\Omega)} + TV(f)$ in the BV space. The distributed derivative $|Df|$ generates an approximation of input "cartoon model"and oscillation model [164]. Therefore, this process preserves discontinuities during the elimination of oscillatory noise. We note that the term $\int_\Omega (g - hf)^2 dxdy$ is the power of the residue. Therefore, there exists a relationship among the non-oscillatory sketch "cartoon model" [173], [33], oscillation model [164] and the reduced power of the original image with some proportional measure.

We formulate the local variance $L_w(i, j)$ in a given window w based on an input image.

$$L_w(i,j) = \frac{1}{|\Omega|} \int_\Omega [f_w(i,j) - E(f_w)]^2 w(i,j) didj \tag{5.49}$$

where $w(i, j)$ is a normalized and symmetric small window, $E(f_w)$ is the expected value with respect to the window $w(i, j)$ on the size of the estimated image f in each iteration. The local variance in a small window satisfies $var(f_w) = L_w(i, j)$. Thereby, we can write λ for a small window w according to Euler-Lagrange equation for the variation with respect to f Therefore, the regularization equation with respect to the windows becomes

$$J\varepsilon(f) = \sum \lambda_w L_w(i,j) + S_p(f) \tag{5.50}$$

where $\lambda_w$is a λ in a small window w. C is a constant is a small window. g_w and f_w is the observed image and the estimated image in a small window w. Thus, we can easily get many λ_w for moving windows which can be adjusted by local variances, shown in Fig. 5.10. These λ_w are directly used as regularization parameters for adjusting the balance during the energy optimization. They also adjust the strength of diffusion operators for keeping more fidelity during the diffusion process. The related regularization parameters β and γ incorporate λ, while the parameter λ of the fidelity term needs to be defined.

During image restoration, the parameter λ can be switched among three different levels. The window-based parameter λ_w and the scale-based (entire image) parameter can be adjusted to find the optimal results. Simultaneously, λ thus controls the image fidelity and diffusion strength of each selected operator in an optimal manner.

5.4 Numerical Approximation

We studied the problem of image reconstruction when the PSF operator $h = I_n$ (corresponding to a denoising problem, I_n is the identity matrix $I_n = diag(1, 1, ..., 1)$. If $h \neq I_n$ (generally a convolution operator), the existence and uniqueness results remain true, if h satisfies the following hypotheses: (a) h is a continuous and linear operator on $L^2(\Omega)$. (b) h does not annihilate constant functions. (c) h is injective.

5.4.1 Numerical Approximation of Image Denoising

Let for the moment $h = I_n$. The proposed double variational regularization functional Eq. 5.42 becomes only denoising. For numerical reasons, we need to compute $\mathcal{J}_\varepsilon$, the continuous approximation of the BV solution f, with $\varepsilon > 0$ small enough. The Eq. 5.44 becomes only for image denoising without including deconvolution process

$$g = f - \lambda \underbrace{|\nabla \hat{f}|^{p(x)-2}}_{\text{Coefficient}} [\underbrace{(p(x)-1)\Delta \hat{f}}_{\text{IsotropicTerm}} + \underbrace{(2-p(x))|\nabla \hat{f}| div(\frac{\nabla \hat{f}}{|\nabla \hat{f}|})}_{\text{CurvatureTerm}} + \underbrace{\nabla p \cdot \nabla \hat{f} \log |\nabla \hat{f}|}_{\text{HyperbolicTerm}}]$$

where $\frac{\partial f}{\partial N} = 0$, along the boundary $\partial\Omega$.

$$p(x) = \begin{cases} q(x) \equiv 1 + \frac{1}{1+k|\nabla G_\sigma * I(x)|^2}, & |\nabla \hat{f}| < \beta \\ 1, & |\nabla \hat{f}| \geq \beta \end{cases} \tag{5.51}$$

We describe the extension to the two-dimensional problem with $\nabla f = (f_x, f_y)$, $\Omega \subset \mathbb{R}^2$ and with $\partial f/\partial n = 0$ on $\Gamma = \partial\Omega$. Assume spatial step $\tau > 0$, and let $x_i = i\tau,\ y_j = j\tau,\ \tau = 1/M$, for $0 \leq i,\ j \leq M$, be the discrete points. We recall the following usual notations:

1. $f_\tau(x_i, y_j) = f_{ij} \approx f(x_i, y_j),\ f_{0,\tau}(x_i, y_j) = f_{0,ij} \approx g(x_i, y_j)$.

2. $m(a, b) = \min \mathrm{mod}(a, b) = ((\mathrm{sign}a + \mathrm{sign}b)/2) \min(|a|, |b|)$.

3. $\nabla^x_\mp f_{ij} = \mp(f_{i\mp1,j} - f_{ij})$ and $\nabla^y_\mp f_{ij} = \mp(f_{i,j\mp1} - f_{ij})$.

with the boundary conditions

$$f^{n+1}_{0,j} = f^{n+1}_{1,j},\ f^{n+1}_{Mj} = f^{n+1}_{M-1,j},\ f^{n+1}_{i0} = f^{n+1}_{i1},\ f^{n+1}_{i,M} = f^{n+1}_{i,M-1}.$$

We use here the minmod function, in order to reduce the oscillations and to get the correct values of derivatives in the case of local maxima and minima. The detailed evolution can be classified into three types such as explicit scheme $f^{k+1} - f^k/\tau = A(f^k)f^k$, implicit scheme $f^{k+1} - f^k/\tau = A(f^{k+1})f^{k+1}$, recently developed semi-implicit scheme $f^{k+1} - f^k/\tau = A(f^k)f^{k+1}$ by Weickert et al. [265]. The semi-implicit scheme are stable for all time steps in arbitrary dimensions using a discrete nonlinear diffusion scale-space framework based on *Thomas algorithm* [240].

5.4.2 Numerical Approximation of Image Denoising and Deblurring

Now we consider the case $h \neq I_n$. In many cases the degradation operator h, the blur, is a convolution type integral operator.

In the numerical approximations, $(h_m n)_{m,n=0,d}$ is a symmetric matrix with

$$\sum_{m,n=1}^{d} h_{mn} = 1$$

and an approximation of h_u can be

$$hf_{ij} \sum_{m,n=1}^{d} h_{mn} f_{i+d/2-m,j+d/2-n}$$

Since h is symmetric, then $h^* = h \; and \; h^* hf = hhf$ is approximated by

$$hhf_{ij} = \sum_{m,n=1}^{d} \sum_{r,t=1}^{d} h_{mn} h_{rt} f_{i+d-r-m,j+d-t-n.}$$

Then we use the same approximation of the divergence term and the same iterative algorithm, with a slight modification.

5.5 Experiments and Results

5.5.1 Denoising and Image Restoration for Noisy Images

Firstly, we have studied the importance of diffusion in the regularization based image deconvolution, shown in Fig. 5.11. The second experiments demonstrate the efficiency of the suggested edge-driven diffusion method. From visual perception and denoising viewpoint, our unsupervised edge-driven method favorably compares to some state-of-the-art methods: the TV [213], a statistic-wavelet method (GSM) [198] and a Markov random field based filter learning method (FoE) [209] using a PIII 1.8GHz PC. In Fig. 5.12 and Fig. 5.14, the structure of the restored fingerprint is largely enhanced than the original image in our method and more recognizable than the restored image using the GSM method [198]. Fig. 5.13 shows the advantage of our method, while the TV method [213] has some piecewise constant effects during the denoising. Table 5.3 shows the different properties of different methods and also shows our method outperforms most of these methods. To achieve similar results, FoE [209] needs more time. Our method (100 iter.) is faster than the TV method (30 iter.) in that our method that does not over-smooth and generate redundant image discontinuities. The GSM [198] method is relatively faster due to the computation in the Fourier domain. However, the GSM is only designed for denoising. The dual-purpose edge-driven method is not only for denoising but also for compensating the "ringing" and "staircase" effects and for protecting the image structure and textures during the image deconvolution.

Figure 5.11: The role of smoothing operators in regularization based image deblurring. Even with known PSF, the staircasing effects are generated during the deconvolution process.

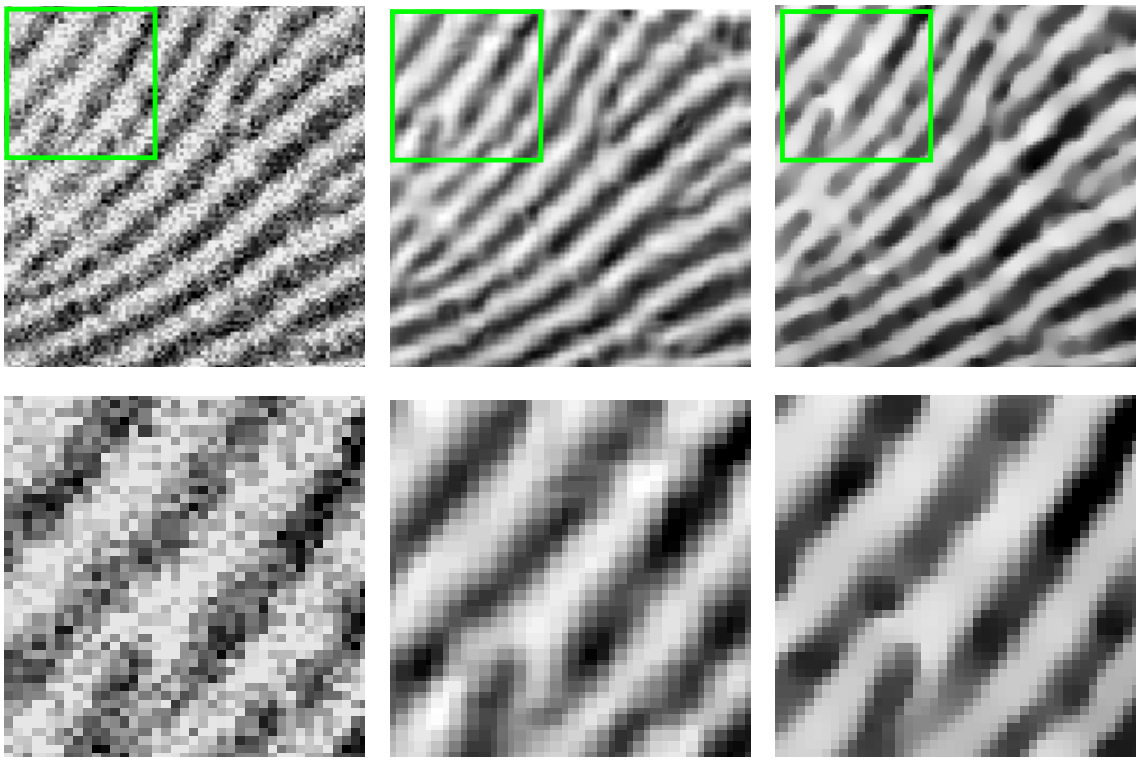


Figure 5.12: $\frac{a|b|c}{d|e|f}$. Compare two methods in fingerprint denoising. (a)(d) Cropped noisy image, $SNR = 8$ dB. (b)(e) GSM method[198] PSNR=27.8. dB. (c)(f) The suggested method PSNR= 28.6 dB

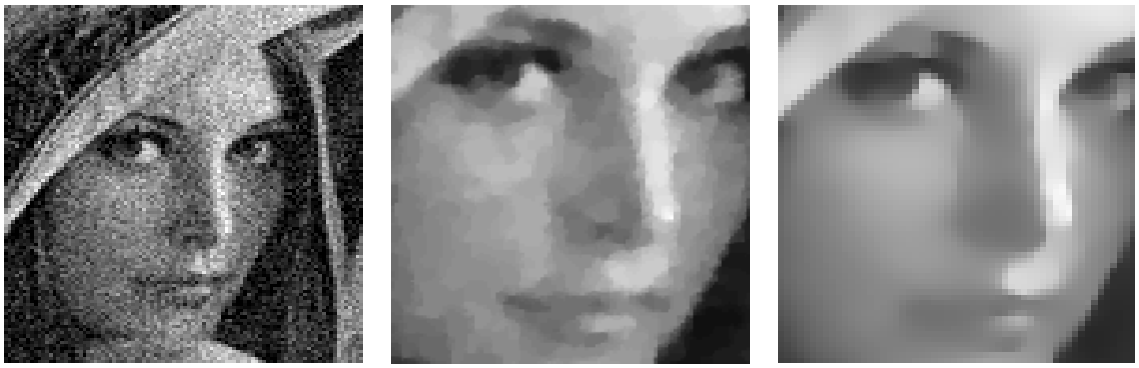


Figure 5.13: Denoising. *a:* Unblurred noisy image, SNR=8dB, size: $[256, 256]$. *b:* Normal TV method, $PSNR = 27.1$ dB. *c:* data-driven diffusion, $PSNR = 30.2$ dB.

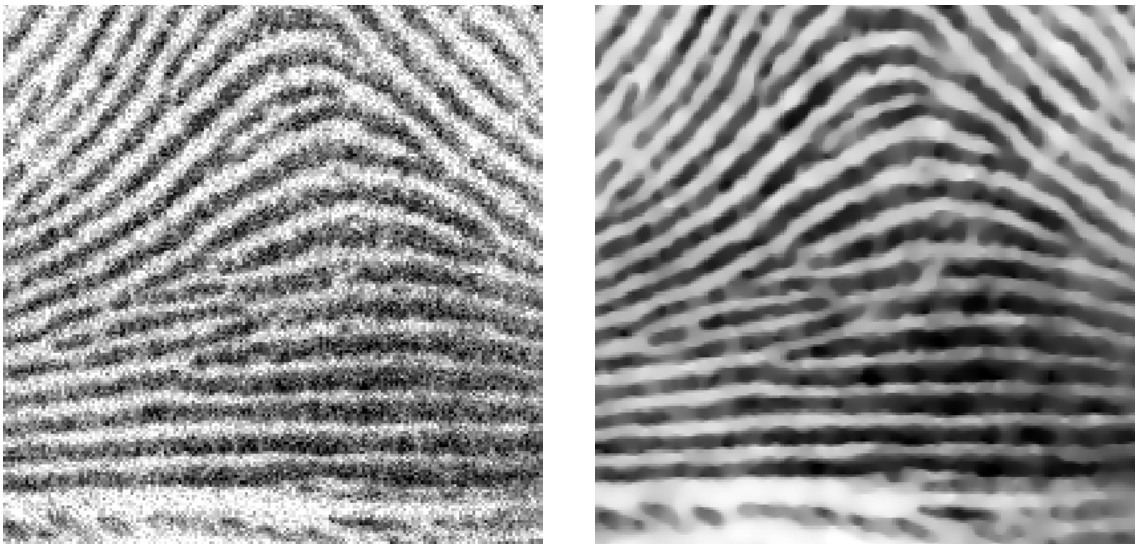


Figure 5.14: *a|b.* Data-driven image denoising using the suggested method. (a). Additive Gaussian noise. $SNR = 8$ dB. (b). Restored using the suggested method PSNR= 28.6 dB

Table 5.3 shows the different properties of different methods and also shows that our method outperforms the total variation methods in signal-to-noise-ratio improvement (SNRI)(dB). The advantage of our method is its high-fidelity and smoothness of visual perception so that the SNRI is higher than that of the related TV methods. From these experiments, we conclude that the regularization functional in the BV space has some advantages for image denoising.

Table 5.2: Denoising performance of different methods on PSNR (dB)

PSNR	$\sigma = 17.5$, $SNR \approx 8.7$ dB, size [512,512]						Iter(n)	Time(s)
(dB)	Lena	Barbara	Boats	House	Pepper	fingerprint	Number	Second
Our Met.	32.26	31.25	31.01	31.85	30.61	28.81	100	$600 \sim 650$
TV.[213]	31.28	26.33	29.42	31.33	24.57	27.29	30	$800 \sim 820$
FoE[209]	32.11	27.65	30.26	32.51	30.42	26.41	$1 \sim 3 \times 10^3$	$3 \sim 9 \times 10^3$
GSM[198]	32.72	30.12	30.58	32.69	30.78	28.59	100	$140 \sim 180$

Table 5.3: ISNR (dB) Results on Test Data

SNR	TV-fixed λ	TV- adaptive λ	Our met.
13.8	15.39	17.85	19.16
12.5	14.42	17.12	18.14
8.7	11.58	15.03	16.26
8.6	11.34	15.02	16.09

5.5.2 Denoising and Unsupervised Deblurring for Blurred Noisy Images

For blind deconvolution, synthetic and real natural images are processed for the testing. First, a synthetic blurred image (noise = 30 dB) is considered for blur identification and restoration, shown in Fig.5.15. The results show the efficiency of the proposed approach.

Second, we compare the classical Lucy-Richardson (L-R) deconvolution method with known PSF to the suggested method with unknown PSF. A MRI image is heavily blurred with two levels of noise 20 dB and 12 dB, shown in the first column of Fig. 5.16. The noise is amplified during the L-R deconvolution with known PSF, shown in the middle column. In the suggested method, the self-initialized PSF is iteratively parametric optimized in the AM algorithm. Diffusion operators vary with the coefficient $p(x)$ in the interval $[1, 2]$ continuously. The estimated PSF supports the image smoothing coefficients progressively till the best recovered image is reached, shown in the right column. From the restored images, we can observe that the low frequency regions are more smooth while the fine details of discontinuities (high frequency regions) are preserved during the image deconvolution. The experiment demonstrates the flexibility of Bayesian based double regularization method which can accurately identify the blur and restore images using edge-driven nonlinear image operators. The results also show that the denoising and debluring can be achieved simultaneously even under the presence of stronger noise and blur.

Our experiments show that blur identification is sensitive to the noise level in the observed image. However, the blur kernel (PSF) in a given blurred noise image cannot be identified and recovered using deconvolution methods. The reason is that the blur kernels are totally modified after the denoising procedure including linear and nonlinear diffusions. The best strategy is to achieve blur identification and image restoration in an interleaved manner. Thus, the alternating minimization with respect to the estimation of PSFs and images can avoid such difficulties.

5.5.3 Effects of Different Types and Strengths of Noise and Blur

We have also tested this approach in different types of noises, speckle, impulsive noise, Poisson, Gaussian noise in different level of strength. Fig. 5.17 and Fig. 5.18 show that the image denoising can be successfully achieved even on the very strong noise level $SNR = 1.5dB$. The intermediate restoration results with detailed diffusion effects can be observed in Fig. 5.17. Fig. 5.20 shows that the suggested approach in the BV space is robust for different types of noise. The impulsive noise (salt-and-pepper) with different strength can also be successfully eliminated, while structure and main textures are still preserved. We have also tested this approach in different types of noise, speckle, impulsive noise, Poisson, Gaussian noise in different strength levels, shown in Fig. 5.19, Fig. 5.20, and Fig. 5.21. Some more results are shown in Fig. 5.22, Fig. 5.23, and Fig. 5.24 to demonstrate that the suggested method keeps high-fidelity

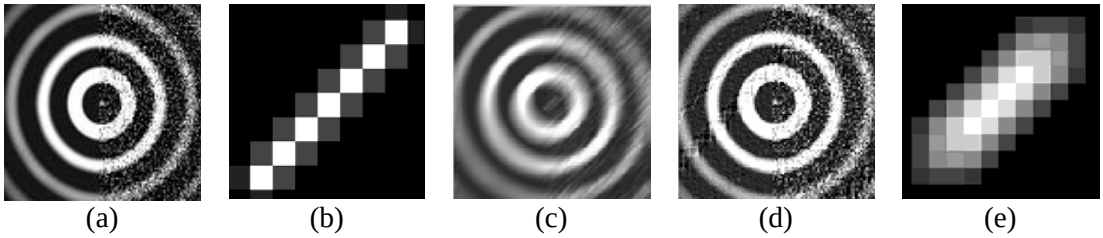


Figure 5.15: (a) Ground truth image. (b) Ground truth PSF. (c) Blurred image with white Gaussian noise 30dB. (d) Blind deconvolution for image c. (e) Estimated PSF

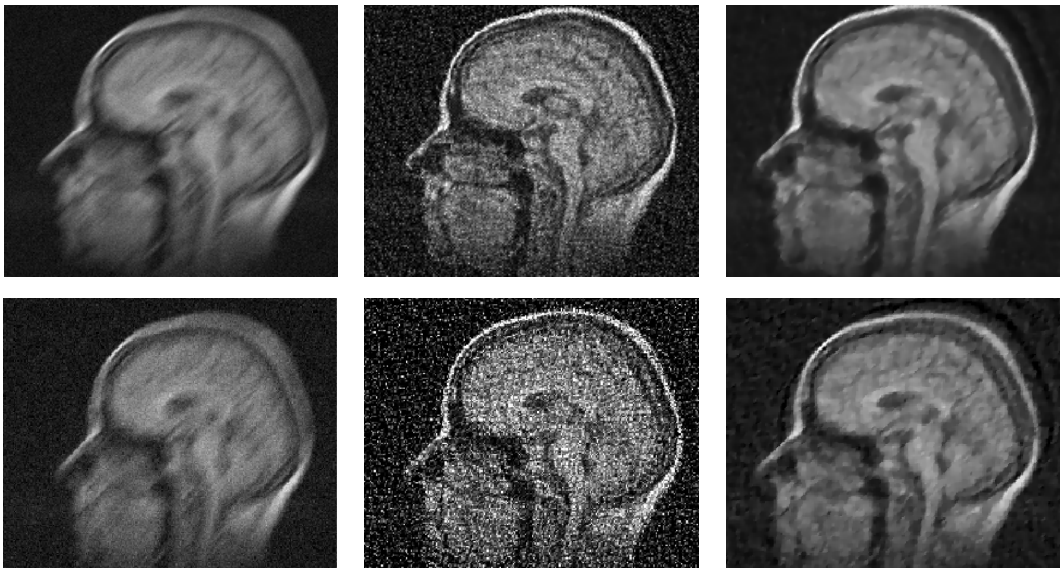


Figure 5.16: $a|b|c$ Deconvolution and denoising. (a) From top to bottom: $SNR = 20$dB and 12dB, size: $[256, 256]$. (b) L-R method with known PSF. (c) The suggested method with unknown PSF.

during image restoration.

Improvement of signal-to-noise ratio (SNR) sometimes might not match human visual perception. For example, for the salt-pepper noise, while the SNR value becomes larger (Normally, the restoration result goes well), the visual perception can not be improved continuously but becomes worse, e.g., shown in Fig. 5.20 and Fig. 5.21.

5.6 Discussion

This chapter introduces a novel Bayesian based variational image restoration model incorporating the dynamic computed prior knowledge. This model can achieve adaptive data-driven image restoration in an integrated mathematical functional in the BV space. This functional is derived using Lebesgue integral based on the total variation functional in the BV space. It is a more accurate approximation of images in the spatial domain. Moreover, although this functional can be considered as a sibling of the well-known Mumford-Shah functional [220], it has different computation mechanism. In Mumford-Shah functional, we need to simultaneously compute the length of curves and piecewise regions by using Γ-convergence. It increases the difficulties in discrete computation. The proposed functional and method can be directly

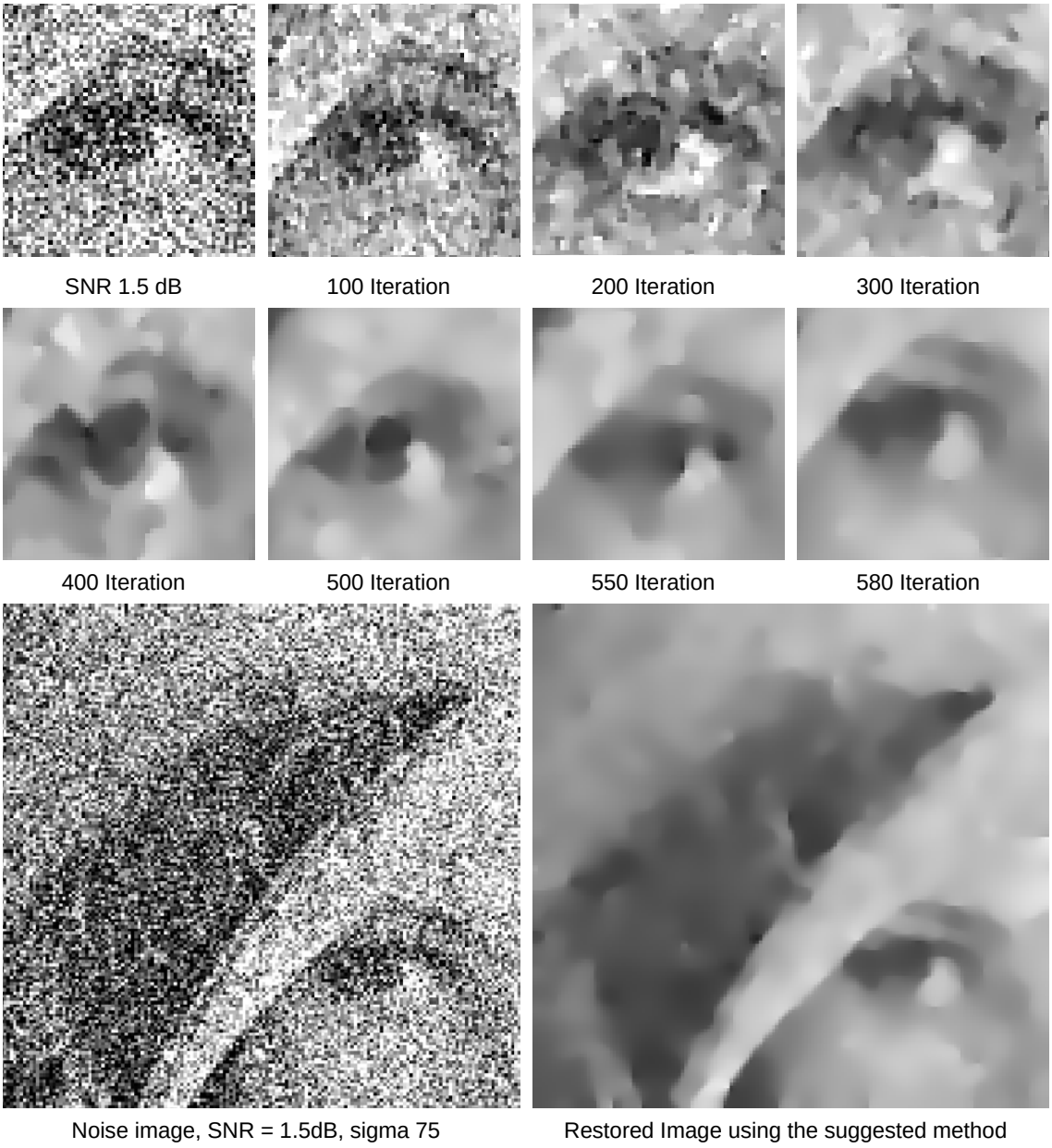


Figure 5.17: Restored image using the suggested method. The noise image has stronger distributed noise level, $SNR = 1.5dB$. In this figure, we can observe that the number of iteration is dependent on the noise strength. If the noise is stronger, the number of iteration is bigger.

computed using different discrete image diffusion methods following the gradient of edges and discontinuities. Therefore, the computation of regions and discontinuities are "separated" and more well-posed to achieve high fidelity image restoration. As these experiments show, the regularization functional in the BV space has some advantages on image denoising, deblurring and image restoration. The suggested method can also be easily extended to other regularization functionals for solving image restoration and other related early vision problems.

From another point of view, inverse scale space interpreted regularization methods introduced by Scherzer, Grotesch [217] and Weickert [218] based on a different paradigm are gradually applied to image restoration. Simultaneously, recent methods like combination of nonlinear diffusion

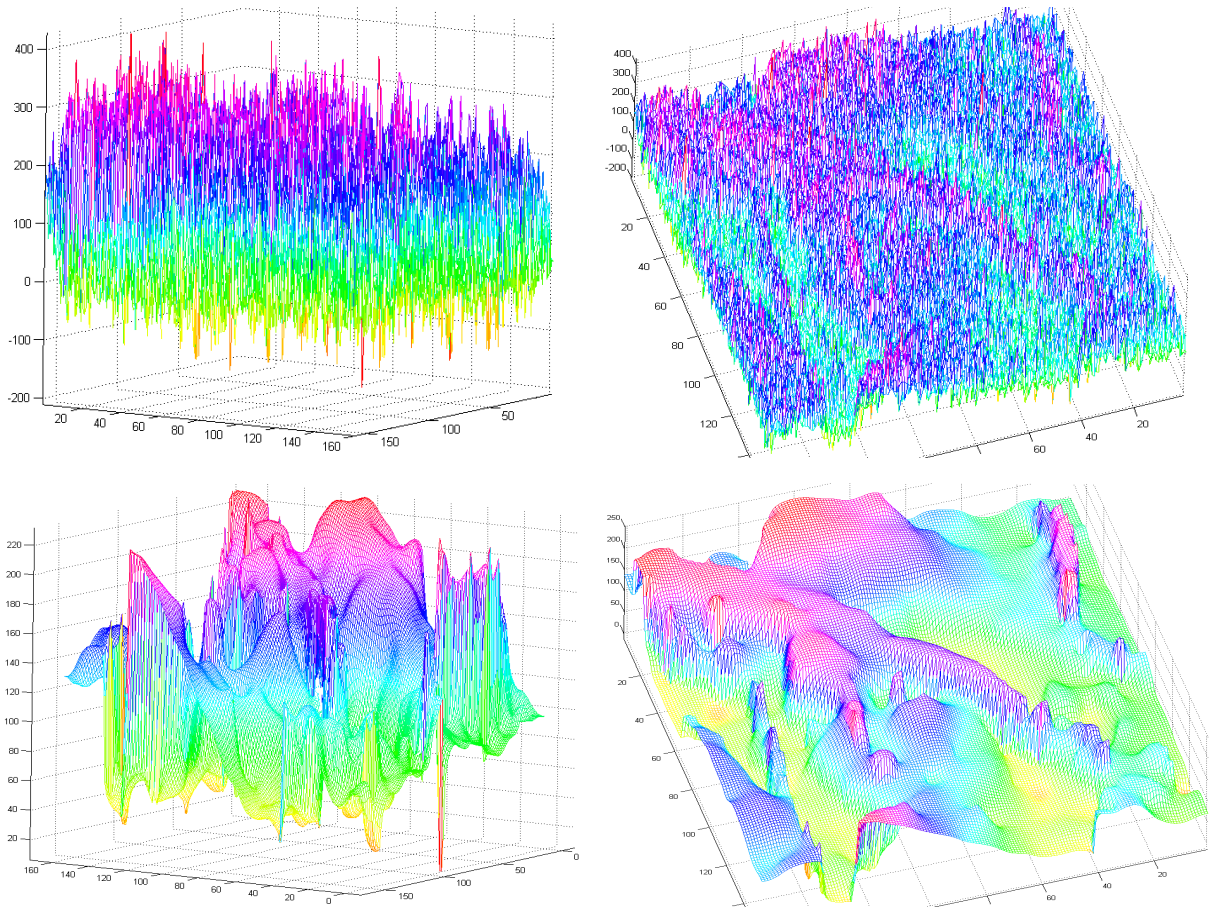


Figure 5.18: $\frac{a|b}{c|d}$. The surface of restored images using the suggested method. The noise image has stronger distributed noise level, $SNR = 1.5dB$. (a)(b) Noisy surfaces. (c)(d) Surfaces of the restored image.

and wavelet shrinkage from signal scale to multiscale [61], [277], [276] make some progress on image denoising. Different from these methods, our proposed method is based on a more general function which is a deduction in the BV space. Our approach is also one kind of "active" image restoration method based on the Bayesian framework.

Further improvements in performance of state-of-the-art algorithms might be possible through a further reduction of unknowns in a Bayesian estimation based optimization framework. More over, the initialization of the location of the centers of the basis functions is crucial for regularization based optimization. On the other hand, Stochastic optimization approaches from Winkler [273], Hellwich [111], [113] uses *a priori* information concerning line continuity expressed as neighborhood relations between pixels. This method can be extended to supply more descriptive and generative prior knowledge to the suggested approach for solving such ill-posed inverse problems.

5.7 Conclusions

The main structure and skeleton of images are well approximated in the BV space. In order to preserve textures and detailed structures, more constraints or generative prior information

Figure 5.19: $\frac{a|b|c|d}{e|f|g|h}$. Image denoising using the suggested method. (a)(b)Speckle noise image and denoising. (c)(d) Zoom in from (a)(b) respectively, 100 iterations. (e)(f) Poisson noise image and denoising. (g)(h) Zoom in from (e)(f) respectively, 100 iterations.

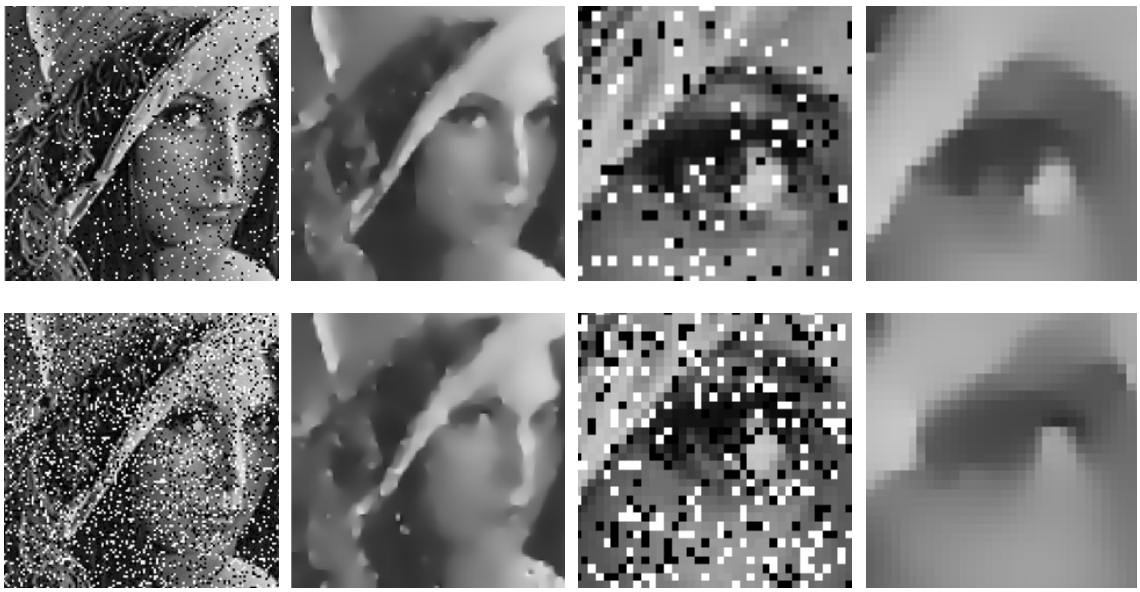


Figure 5.20: $\frac{a|b|c|d}{e|f|g|h}$. Restoration of impulsive noise images. (a) 10% salt-pepper noise image. (b) Restored image, 200 iterations. (c)(d) Zoom in from (a)(b) respectively. (e)25% salt-pepper noise image. (f) Restored image, 900 iterations. From visual perception viewpoint, 700 iteration is better than 900 iteration. However, the SNRI value is less than that of 900 iterations. (g)(h)Zoom in from (e)(f) respectively.

are called for. We develop a self-adjusting scheme that controls the level of denoising by local variances based on the edge-driven convex semi-continuous functionals. The performance of image denoising is not only based on the computed gradient but also based on the computed local variances of the residues. Therefore, linear and nonlinear smoothing operators in the smoothing term are continuously self-adjusting to the gradient power. Also, the fidelity term

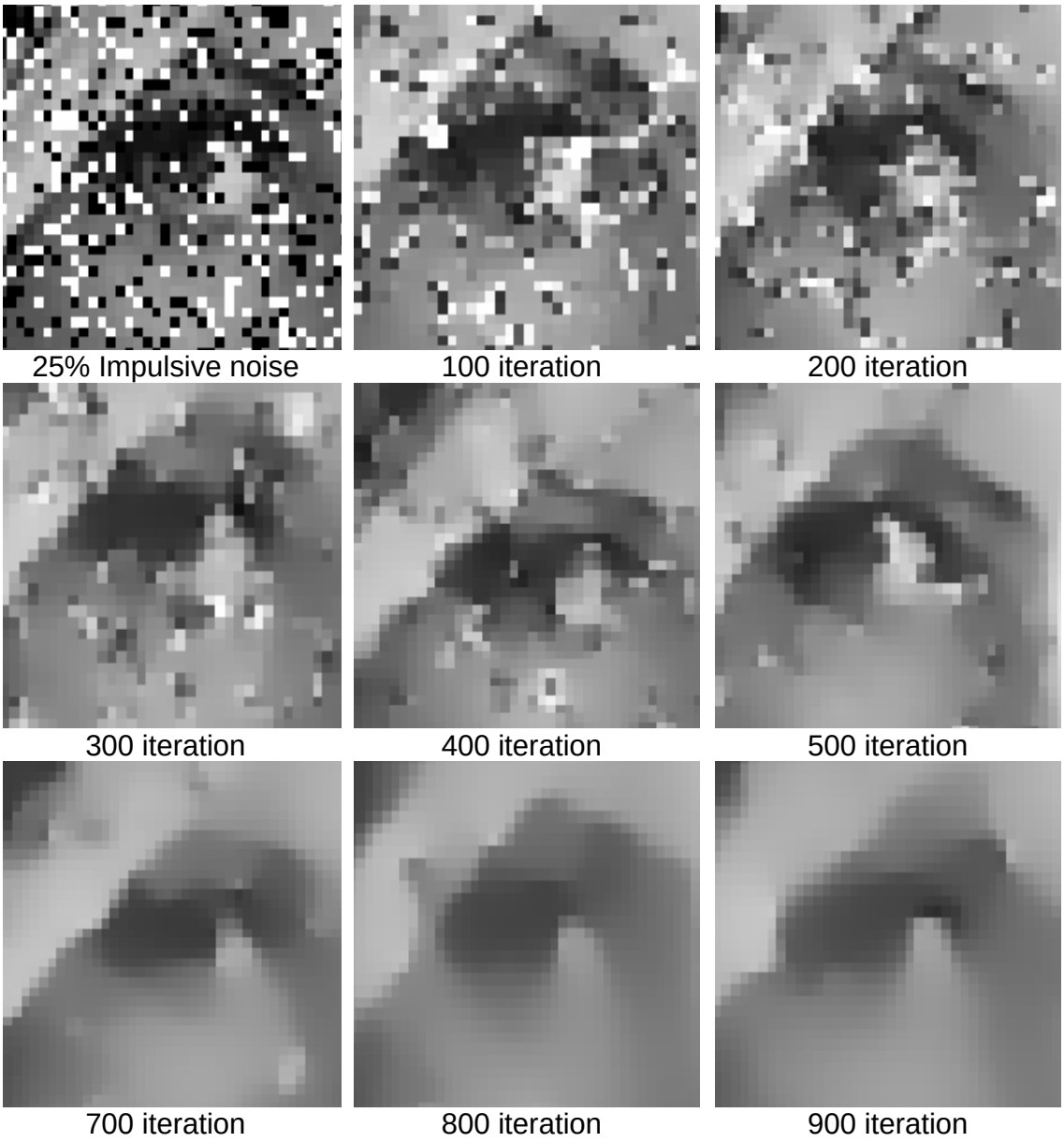


Figure 5.21: $\frac{a|b|c|d}{e|f|g|h}$. Restoration of impulsive noise images. (a) 10% salt-pepper noise image. (b) Restored image, 200 iteration. (c)(d) Zoom in from (a)(b) respectively. (e)25% salt-pepper noise image. (f) Restored image, 900 iteration. (g)(h)Zoom in from (e)(f) respectively.

in the functional is self-adapting the fidelity value of the input image. The consistency of self-adjusting local variances and the global convergence can be achieved in the iterative convex optimization approach. We have shown that this algorithm has relatively robust performance for different types of noise and different noise levels. The restoration keeps high fidelity to the original image.

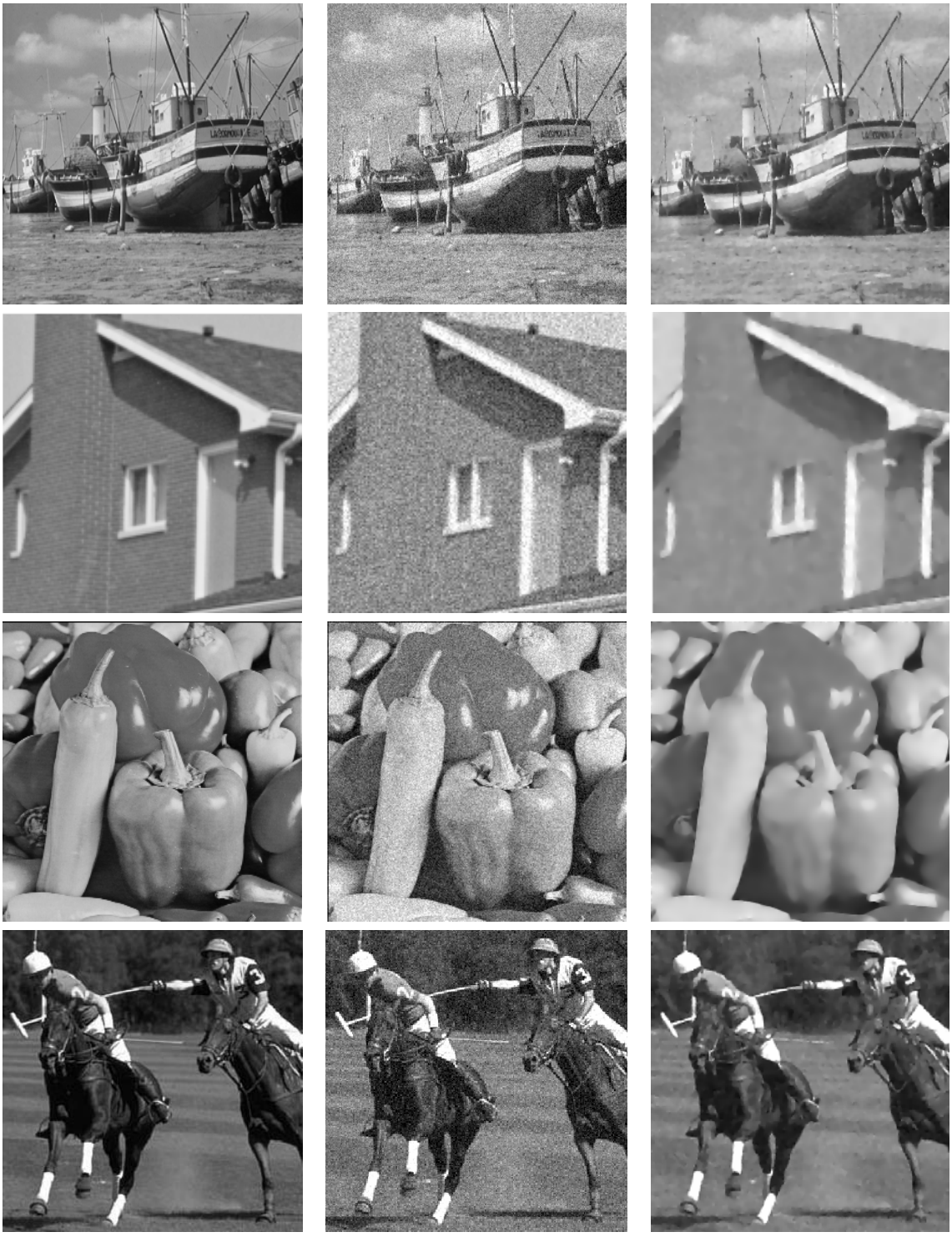


Figure 5.22: Image denoising using the suggested method. Image denoising using the suggested method. (a) column: Original images. (b) column: Noisy images with SNR = 10 dB . (c) column: Restored images (100 iterations) using the suggested method.

Figure 5.23: Edge-driven denoising. (a) Noisy image with PSNR =25.38dB, sigma = 25. (b) Restored image after 120 iterations. (c) Restored image after 150 iterations

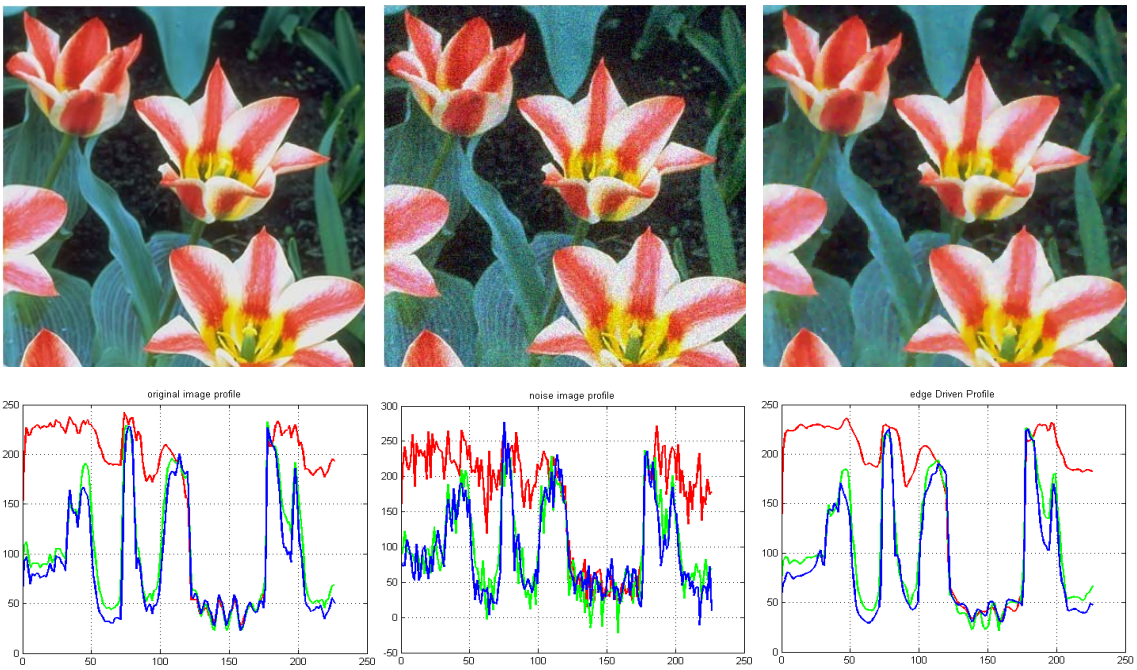


Figure 5.24: $\frac{a|b|c}{d|e|f}$. (a)(d) The original color image and its R,G,B color profile. (b)(e) The noisy image and its R,G,B color profile, $SNR = 8.6dB$. (c)(f) The restored image and its R,G,B color profile. SNR Improvement =16.1 dB.

6 Nonuniform Blurred Image Identification, Segmentation and Restoration

We must envision the present state of the universe as the effect of its anterior state and as the cause of the following state. - Laplace(1795)

Since the restoration of a nonuniform-blurred (e.g., partially-blurred) image is to restore blurred regions or objects without influencing unblurred regions or objects, we can not directly apply traditional methods for this task. We derive a regularized spectral graph clustering approach on discrete graph spaces for partially-blurred image restoration, and show that it is possible to achieve high-quality blurred regions segmentation, and perceptual blind image restoration. Based on the assumption of image foreground and background, natural image learning help us to find differences between foreground and background regions in given images. These differences are labeled as prior information for smoothing spectral clustering energy in an iterative regularization framework. Nonlinear diffusion methods and Hausdorff distance are used to enhance and maintain the learning and labeling for optimal image segmentation. We then show how the global optimization can be efficiently found by combining bottom-up and top-down principle via learning and sparse labeling in an iterative regularization approach.

Furthermore, these identified and segmented blurred regions or objects are mostly non-stationary blurred. It means that we cannot directly to represent these real blur kernels using some simple parametric blur kernels. Therefore, based on previous work, we extend our previous double regularized Bayesian estimation to a more tractable variational Bayesian learning approach. This approach allows the true posterior to be approximated by a simpler approximate distribution for which the required inference are tractable. Moreover, reasonable and effect prior probability is important in Bayesian learning. Natural image learning can help us find translation and scale-invariant spatial prior distribution. In particular, the approach makes effective use of the natural image statistics through the whole variational learning scheme. Our experiments show that the results derived from the algorithm are superior to this type of blurred images. The scheme can be further extended to other types blurred image restoration in real environments.

6.1 Introduction

6.1.1 Problem Formation

The regularization theory [241] has been recognized as a unified framework for studying several problems in computer vision and image processing [195]. It also presents numerous challenges as well as opportunities for further statistical and mathematical modeling, e.g., Markov random fields based regularization [85], neural networks based regularization [91], kernel based regularization [222], variational regularization [173], [184], [267], and discrete analogue of Tikhonov regularization [24], [300]. Although some of these regularization approaches can be used for

Figure 6.1: $a|b|c$ columns. (a) Entirely, uniform and relatively stationary blurred image . (b) uniform and nonstationary blurred image. (c) nonuniform, partially blurred image.

entirely linear-invariant (stationary) blurred image restoration, we can not directly apply these approaches to partially-blurred restoration.

Due to the complexity of blurring, we classify blurred images into three main groups so that we can design related methods for reconstructing these images, shown in Fig. 6.1. The first group in Fig. 6.1(a) is uniform and stationary blurring. The blur kernel for the entire image can be approximated by only one parametric blur kernel like Gaussian blur kernel, motion blur kernel and so on. The second group in Fig. 6.1(b) is uniform but nonstationary blurring. Such blurred images are entirely blurred and the blur kernel can not be represented by a single parametric model. The blur kernel of such images can be considered as a generalization from parametric to nonparametric approximation. The third group in Fig. 6.1(c) is partially-blurring. Such blurred images are nonuniform partially-blurred and the restoration should not influences unblurred regions. In this paper, we focus on entirely unform, partially nonuniform and nonstationary blurred image restoration in real environments.

Partially-blurred image restoration is to restore blurred regions without influencing unblurred regions for achieving better visual perception based on the Gestalt theory [270]. It generates an interesting question. From the mathematical viewpoint the question is, how to get a global convergence of multi-levels of local distributions. These multi-levels of local distributions include local pixel gray level distributions, randomly distributed local blurry regions and unblurred regions or objects. Therefore, it becomes a challenging partial convergence problem [248]. A novel mathematical model needs to be constructed for the solution.

6.1.2 Prior Work

Image segmentation is an important but large topic. Here, we limit our discussion on closely related work in discrete spaces and continuous partial differential equation spaces, respectively. Then we present our proposed algorithm based on an integration of regularization and spectral clustering methods.

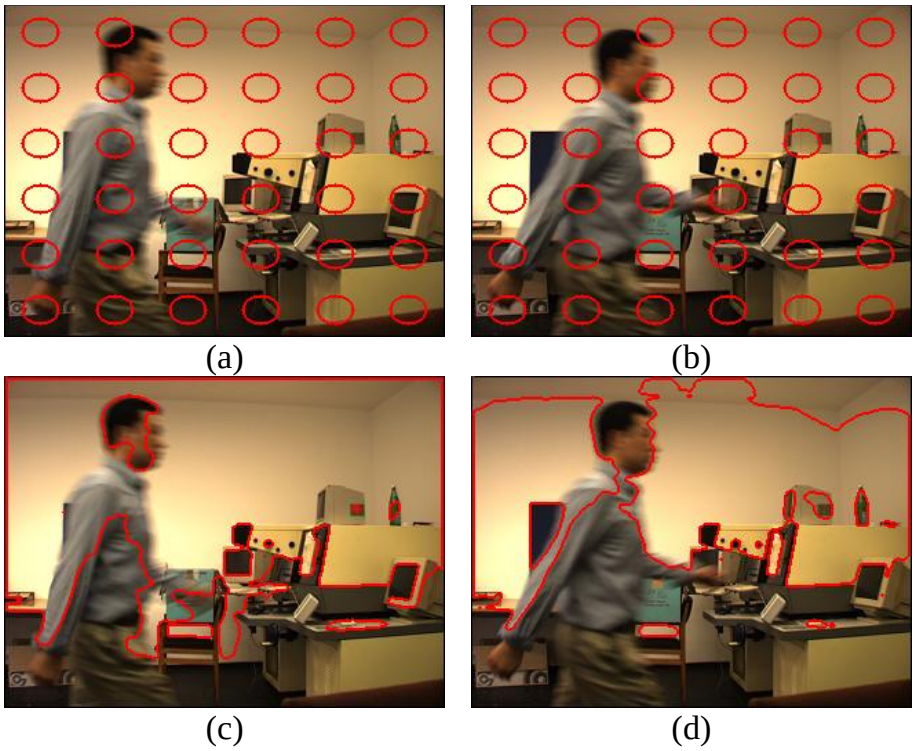


Figure 6.2: Level set method for identifying and segmenting blurred regions and unblurred regions. (Here the method is performed automatically without judging the parameters). (a)(b) initial images. (c)(d) Related results. The better segmentation results are closely related to the selection of reasonable parameters and how to fix a desired contour corresponding to a local energy minimum.

Supervised Image Segmentation

Supervised image partition and segmentation methods typically are based on one or two paradigms [96]: (a) Labeling of pieces of boundary includes the desired boundary for the desired object. (b) Labeling some sets pixels belongs to the desired object or background. Many current automatic segmentation methods can be considered as directly supervised or indirectly supervised and towards the target of unsupervised perceptual image segmentation.

In continuous spaces, partial differential equations based variational segmentation methods have been intensively investigated by researchers. For example, the level set method is inspired by the classical geodesic snake method but it has a lot of advantages than the snake method and is different from the snake method. Normally, the level set method evolves the boundary to a local energy minimum. For this purpose, an initial closed contour is generally needed near the desired boundary. However, some main difficulties of the level set method are the selection of reasonable parameters and how to fix a desired contour corresponding to a local energy minimum, especially on cluttered images, e.g., blurred images or partially-blurred natural images, shown in Fig. 6.2. The reason is that the level set methods are defined in the continuum and achieve a local energy minimum, leading difficulties to achieve a global solution for cluttered images. To avoid such difficulties, Cremers et al. [55], [56] integrate the level set method and statistical shape knowledge based on an energy functional, e.g., the Mumford-Shah functional. The statistical coded prior knowledge can be considered as constraints for guiding the curve evolution process and obtain stability and noise robustness results.

In the discrete spaces, there is also a lot of state-of-the-art methods developed, e.g., Markov random field based segmentation methods [273], [112], [143], [247] and so on. Recently, the graph cuts method has been developed as an interactive, seeded (labeling) optimization method for segmentation. The foreground and background of the image are labeled with some seeds so that the max-flow/min-cut computation can be performed to estimate the minimum-weight cut between the "source" and the "sink" region. Current work on the graph cut method are mostly focused on several aspects. First, since the method returns the cut that often separates the seeds from the rest of the image, a user needs to mark the seeds continuously to avoid the small cut problem. Second, since the K-way graph-cut problem is NP-Hard, the optimal cut becomes more difficult. Third, the graph-cuts segmentation algorithm has been extended in two different directions in order to address several issues. The issue of speed is addressed by applying a multi-level approach [149], by applying a watershed basin as "supernode" in a coarse graph [144]. The iterative estimation of a color model with some user interaction such as the graph cuts algorithm [32], the "Grabcut" algorithm [210], the closed form algorithm and optimization [141], [142] for image matting and segmentation.

Different Spectral Clustering Criteria for Segmentation

Spectral graph theory is well developed and gradually investigated for image processing and computer vision problems. Zahn [284] introduced graph-theoretical methods for detecting and describing Gestalt clusters. Wu and Leahy [275] firstly introduced a general approach of segmenting images by way of optimally partitioning an undirected graph using a global cost function. According to a cost function including boundary-cost metric, the sum of the edge weights along a cut boundary is minimized in a polynomial-time algorithm for finding optimal bisection partitioning results.

Since the bisection-partitioning problem is NP-complete, we need to approximate this intractable problem by some relaxing constraints. Likewise, to avoid unnatural bias of partitioning, a general strategy has to scale the cut weight. The crucial kernel of segmentation is how to use eigenvectors to achieve the possibly normalized "affinity matrix". Through the literature, several optimized cutting criteria such as normalized cuts [226], ratio cuts [53], average cuts [216], or "affinity factorization" [190] are used to measure the disassociation between two groups by efficient eigenvector calculations.

Shi and Malik [226] showed that for bi-partitioning, an approximate solution may be obtained by scaling and thresholding the eigenvector corresponding to the second smallest eigenvalue of the normalized Laplacian. Cox et al. [53] normalize the boundary-cost metric to avoid this bias using a ratio regions using a polynomial-time algorithm for finding bisection partitions in an undirected graph. The minimum ratio cut of an arbitrary graph is NP-hard. The ratio cut is limited to connected planar graphs, consisting of three reductions: minimum ratio cut to minimum ratio cycle, minimum ratio cycle to negative-cost cycle, and negative-cost cycle to minimum-cost perfect matching. The above reductions all operate on undirected graphs. Based on Cox's work, Wang et al. [254] propose a cut ratio cost function in a undirected graph. The cut ratio is defined as the ratio of the corresponding sums of two different weights of edges along the cut boundary, and the mean affinity is modeled between the segments separated by the boundary per unit boundary length. This cost function does not introduce a size, shape, smoothness, or boundary-length bias so that this method allows efficient iterated region-based segmentation as well as pixel-based segmentation.

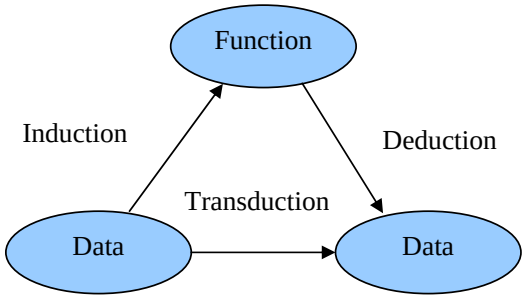


Figure 6.3: The underlying relationship among induction, deduction and transduction.

Statistic methods are incorporated with more information for image segmentation. Swendsen-Wang Cuts [20] use Bayesian based Markov Chain Monto Carlo to split and merge the sub-regions. Berkeley nature boundary detector [159], was recently successfully applied to object recognition. Region cues are computed as the similarity in brightness, color, and texture between image patches. Boundary cues are incorporated by looking for the presence of "intervening contour". The self-tuning clustering method [286] suggests that local adaption of the scaling parameter improves the image segmentation results.

Spectral Clustering for Segmentation Given Partial Constraints

Researchers have tried to give some interactive constraints to guide the segmentation. Yu et al. [282] enforce grouping smoothness and fairness on labeled data points so that sparse partial grouping information can be effectively propagated to the unlabeled data. The given partial grouping prior as constraints can often be derived based on a crude spatial attentional map that places common salient features and focuses on expected object locations. By generalizing the Rayleigh-Ritz theorem to project matrices, the global optimum in the relaxed continuous domain by eigen-decomposition, from which a near-global optimum to the discrete labeling problem can be obtained effectively.

Since the publication of Karmarkar's famous paper [122] in 1984, the area of interior-point polynomial-time methods for convex programming have been intensively developed by many researchers, focusing on linear and quadratic programming. Problems of special interest covered by the approach are those with positive semidefinite matrices as variables. These problems include numerous applications in modern control theory, combinatorial optimization, graph theory and computer sciences. Keuchel et al. [130] apply the semidefinite programming relaxations to the combinatorial problem of minimizing quadratic functions in binary decision variables subject to linear constraints. They introduce an interior-point methods (convex programming) and a random hyperplane to achieve parameter-free and high-quality combinatorial solutions based on spectral graph theory. Recently, the random walking algorithm [96] has used a similar affinity function for the segmentation problem, but the affinity value is computed after applying a linear transformation to the distance measure with human interactive interface.

6.1.3 Our Approach: Perceptual Image Segmentation and Restoration

Our target is to perceptually restore the nonuniform blurred (partially-blurred) images. Therefore, we describe our approach in two steps. The first step is how to automatically and per-

ceptually identify and segment blurred regions or objects. The second step is how to identify blur kernels and perceptually restore these identified and segmented regions or objects without influencing unblurred regions or objects.

Regularized Spectral Graph Clustering

We present a novel approach for perceptual image segmentation in this chapter. Here we present some closely related theory and work. Any supervised learning algorithm can be applied an inference problem, e.g., by training a classifier based on a certain data set, and then using the trained classifier to predict the labels of the unlabeled objects. Following this approach, one will have estimated a classification function defined on the whole domain of data set before predicting the labels of the unlabeled objects. According to Vapnik, [248], Zhou and Schölkopf [300], [301](see also page 221-232) estimating a classification function defined on the while domain is more complex than the original problem which only requires predicting the labels of the given unlabeled objects, and a better approach is to directly predict the labels of the given unlabeled objects. Therefore, we consider estimating a discrete classification function which is defined on the given objects only. Such estimation problem is called transductive inference [248], [300]. In psychology, transductive reasoning means linking particular to particular with no consideration of the general principles. It is generally used by young children. In contrast, deductive reasoning, which is used by adults and older children, means the ability to come to a specific conclusion based on a general premise. The diagram is shown in Fig. 6.3. It is well known that many meaningful inductive methods such as support vector machines (SVMs) can be derived from a regularization framework based on a empirical cost and a regularization term. Inspired by this work [248], [300], we consider to construct an approach by integrating regularization theory and spectral graph theory. Much existing work including spectral clustering, transductive inference and dimensionality reduction can be understood in this framework.

We formulate the problem of partially-blurred image restoration including identification, partition and restoration of blurred regions or objects. To motivate the algorithm, different characteristic properties [80], [153] (gradient, frequency, entropy, etc.) [67], [191] between blurred and unblurred regions or objects endowed with pairwise relationships can be naturally considered as a graph. We treat blind image restoration (BIR) of partially-blurred images as a combinatorial optimization problem [130], [85], [80] based on regularization theory [241], and spectral clustering theory on discrete graph spaces [51] and its related algebraic graph transformation [65], [80], [226], [282]. Some connections between some of these interpretations are also observed in [300], [282], [142], [51] based on transductive inferences and differential geometry. More important, this integration brings crucial insights to the understanding of these theories, underlying relationships and their potential roles.

As we know, segmentation is only a computing process not a final target. A meaningful segmentation needs to be integrated with a specific task. Discrete regularization can achieve meaningful segmentation from intrinsic ambiguities of a given image in that this approach induces and stores high-level knowledge (top-down: identify and segment partially-blurred regions or object) to control low-level image processing (bottom-up: pairwise measure between blurred and unblurred pixels and regions) via regularization. For example, the penalty term in regularization becomes a carrier of learned priors with certain smoothing weights and scales. Also, the concepts of using non-negative physical constraints are well matched and integrated into the discrete regularization. For example, blur kernels and images are non-negative. Therefore, the resulting

simplicity of this approach differs in an interesting way from those algorithms generated without the non-negativity constraint and generated descriptive priors.

A more fundamental problem that arises in ill-posed inverse problems is the scale problem. In other words, which scale is the right resolution to operate on? Scale-space theory [191], [267], [67] considers the behavior of the result across a continuum of scales. On the other hand, scale-space theory is an asymptotic formulation of the Tikhonov regularization [241]. Based on the regularization theory, the concept of scale is related quite directly to the regularization parameter. The discrete regularization can obtain an optimal regularization parameter as the optimal scale for the associated instance. The global optimization solution is guaranteed to directly relate to the energy function rather than to a numerical problem during the minimization. Therefore, the consistency of multiple levels of local distributions and the global convergence can be achieved in a reliable and robust manner.

In a summary, the main objective of the standard regularization techniques is to obtain a reasonable reconstruction which is resistant to noise in inverse problems. Based on these inheriting advantages, discrete regularization is about converting high-level targets (human demands), guiding low-level image processing and learning the optimal scale for achieving the global convergence with multi-levels of local distributions. Conceptually, the discrete regularization paradigm also reveals the roles of some well-known optimization algorithms. Algorithms such as graph-cuts [132], and variational regularization [184], [173], [267] can be viewed as either discrete regularization [24] with energy in binary discrete spaces or in continuous bounded variation spaces. Compared to Markov random fields based stochastic optimization approaches [85], [80], this paradigm in the discrete graph space is optimized in a deterministic way.

Natural Image Statistics and Variational Bayesian Leaning based Image Restoration

On the second step, these identified and segmented blurred regions or objects are mostly non-stationary blurred. It means that we cannot directly to represent these real blur kernels using some simple parametric blur kernels. Therefore, based on previous work, we reformulate and extend our previous double regularized Bayesian estimation approach to a more tractable variational Bayesian learning approach based on natural image statistics.

Our work relates to statistical approximation inference [181], variational free energy [183], variational Bayesian learning [14], ensemble learning [114], [27], [166], [167], natural image statistics based image restoration [227], [209], [73], [109] and variational methods in graphical models [121]. In the Bayesian estimation, in general, we may consider two approaches to determining the posterior distribution of the weights. The first is to find the maximum of the posterior distribution, and then to fit a Gaussian function centered on this maximization. The second approach is to express the posterior distribution in terms of a sample of representative vectors, generated using Monte Carlo techniques. The third method is called Bayesian ensemble learning which has been firstly introduced by Hinton [114], [27] and further developed by Miskin and Mackay [166], [167]. Although the Bayesian estimation provides a structured way to include prior knowledge concerning the quantities to be estimated. However, it is often intractable to perform inferences using the true posterior density over the unknown variables, especially for ill-posed inverse problems. Ensemble learning allows the true posterior to be approximated by a simpler approximate distribution for which the required inference are tractable.

This approach allows the true posterior to be approximated by a simpler approximate distribution for which the required inference are tractable. Moreover, reasonable and effect prior

probability is important in Bayesian learning. Natural image learning can help us find translation and scale-invariant spatial prior distribution. In particular, the approach makes effective use of the natural image statistics through the whole variational learning scheme. Our experiments show that the results derived from the algorithm are superior to this type of blurred images. The scheme can be further extended to other types blurred image restoration in real environments.

6.2 Regularization on Discrete Graph Spaces

In this section, we present an overview about the integration of regularization and spectral graph clustering methods in discrete graph space. Our goal is to design a practical regularization algorithm that is adopted to the structure of graphs for partially-blurred image segmentation, identification and restoration. The discrete regularization has the flexibility to derive a family of transductive [248], [300], [51] algorithms based on the integration of spectral graphs, regularization and image formation, composition theory in combinatorial optimization.

6.2.1 Discrete Regularization on Graphs

A general weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a finite set $\mathcal{V}$ with two subsets $\mathcal{A}$ and $\mathcal{B}$, together with a set $\mathcal{E} \subseteq \mathcal{A} \times \mathcal{B}$. The elements of $\mathcal{V} = \{v_i\}_{i=1}^{n}$ are the vertices of the graph $\mathcal{G}$, and the element of $\mathcal{E} \subseteq \{(i, j)\}$ are the edges of the graph, i.e., edges with one endpoint in $\mathcal{A}$ and the other in $\mathcal{B}$. A *self-loop* is an edge which starts and ends at the same vertex. A graph is connected when there is a path between any two vertices. A graph is *undirected* when the set of edges is symmetric, i.e., for each edge $(i, j) = (j, i) \in \mathcal{E}$. A undirected graph is shown in Fig. 6.4.

For a given blurred image $g = hf + \eta$, we approximate a regularization functional on a lattice-pixel based graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with the weight matrix w_{ij}. To restore the observed image g to the ideal image f by deconvolving the unknown blur kernel h, the direct solution of the least squares problem $J(f) = \arg\min\{(h * f - g)^2\}$ may lead to a vector f that is severely contaminate with noise. Therefore, Tikhonov regularization [241] is employed to get a more meaningful solution. The objective function is to minimize the square loss function with a smoothing penalty term $S_p(f)$. Thus, we have

$$\mathcal{J}(f) = \arg\min\{\lambda(g - h * f)^2 + S_p(f)\} \tag{6.1}$$

where $\mathcal{J}(f)$ represent the total energy need to be minimized. The first term on the right side is a squared fidelity term. The regularization parameter λ controls the trade-off between the fidelity and the smoothness term $S_p(f)$,

$$S_p(f) = \sum w_{ij}(f_i - f_j)^p \tag{6.2}$$

where the sum is taken over all the adjacent vertices $\mathcal{V}$. The $S_p(f)$ term can be seen as an discrete analogue of its continuous case. The gradient, divergence, Laplacian and curvature operators between these vertices and edges can be thought of as discrete analogous of their counterparts in

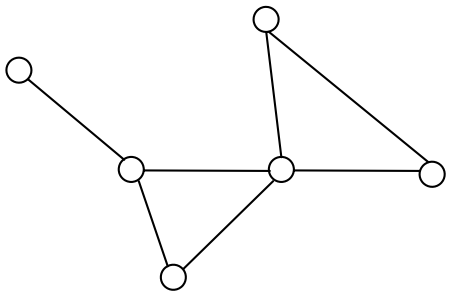


Figure 6.4: An undirected graph with vertices and edges

the continuous case [300], [51]. For example, L^p-norm Laplace operators, $p \in \{1, 2\} \in \mathcal{N}$ are also possible, e.g, $p = 2$ is a Tikhonov regularization form, $p = 1$ becomes a total variation functional. Both these regularization functionals are strictly convex with some nonnegative constraints.

To solve the optimization problem of the discrete regularization, we can either use some unconstrained optimization methods like conjugate gradient descent, Gauss-Seidel, etc. in an iterative approach, or in direct factorization methods, e.g., using Laplacian in spectral graph spaces. Laplacian provides a unifying framework for regression, classification, data representation and clustering in a regularization framework. It also allows to replace difficult optimization problems with standard linear algebra. The optimization can also be achieved with an existing, unique and stable solution in a convex manner. Furthermore, some related smoothing operators in graph spaces can be deduced in a transductive manner [300], [51]. In the following, we describe these smoothing operators in discrete graph spaces.

6.2.2 Discrete Operators on Weighted Graphs

The weighted undirected graph $\mathcal{G}$ has associated with it a weight function $w : \mathcal{V} \times \mathcal{V} \rightarrow \mathcal{R}$ satisfying $w(j, i) = w(i, j)$ and $w(i, j) \geq 0$. $W = \{w_{ij}\}$ is the $n \times n$ symmetrical adjacency matrix with rows and columns indexed by $\mathcal{V}$, and entries is equal to the number of edges between vertices i and j. The degree d_i of a vertex $i \in \mathcal{V}$ defined to be $d_i = \sum_j w(i, j)$ that represents the total connection from vertex i to all other vertices. D is the $n \times n$ diagonal matrix indexed by $\mathcal{V}$ with vertex degrees d on the diagonal. The un-weighted graph is just a special case where all the weights are 0 and 1.

The gradient, divergence, Laplacian and curvature operators between these vertices and edges can be thought of as discrete analogous of their counterparts in the continuous case. These operators are defined in the following.

Definition 6.2.2.1 *Let $\mathcal{H}(\mathcal{V})$ and $\mathcal{H}(\mathcal{E})$ denote the Hilbert space of real-valued functions for the set of vertices and edges, respectively. The graph gradient is an operator $\nabla : \mathcal{H}(\mathcal{V}) \rightarrow \mathcal{H}(\mathcal{E})$ given lattice-pixel based vertices image f defined at a vertex i*

$$(\nabla f)(i, j) := \sqrt{\frac{w(i, j)}{\varphi(i)}} f(i) - \sqrt{\frac{w(i, j)}{\varphi(j)}} f(j), (i, j) \in \mathcal{E}, \varphi \in \mathcal{H}(\mathcal{V})$$

where the gradient measures the variations on each edge, i.e., $(\nabla f)(i, j) = -(\nabla f)(j, i)$ means

that ∇f is skew-symmetric. While the graph gradient is defined on each vertex, the norm of the graph gradient $\|\nabla f\|$ is defined by

$$\|\nabla f\| = \left(\sum\nolimits_{i\sim j}(\nabla f)^2(i,j)\right)^{1/2}$$

Intuitively, the norm of the graph gradient measures the roughness of a function around a vertex (i.e., in lattice pixel image).

Definition 6.2.2.2 *The graph divergence is an operator* $div : \mathcal{H}(\mathcal{E}) \to \mathcal{H}(\mathcal{V})$ *where the inner product satisfies*

$$\langle \nabla f, \psi\rangle_{\mathcal{H}(\mathcal{E})} = \langle f, -div\psi\rangle_{\mathcal{H}(\mathcal{V})}, f \in \mathcal{H}(\mathcal{V}), \psi \in \mathcal{H}(\mathcal{E})$$

The negative gradient -div is defined to be the adjoint of the graph gradient. The graph divergence can be computed by

$$(\text{div } \psi)(i) = \sum_{i\sim j}\sqrt{\frac{w(i,j)}{\varphi(i)}}\left(\psi(i,j) - \psi(j,i)\right), \psi \in \mathcal{H}(\mathcal{E})$$

Intuitively, we can observe that the divergence measures the net data flow of function ψ at each vertex. Note that if ψ is symmetric, then $\text{div}(i) = 0$ for all $i \in \mathcal{V}$.

Definition 6.2.2.3 *The Laplace-Beltrami operator on differentiable functions on a manifold is intimately related to the heat flow. The Laplacian L can be thought of as discrete analogue of the Laplace-Beltrami operator on a manifolds,e.g., Riemannian manifold. The Laplacian operator* $\Delta : \mathcal{H}(\mathcal{V}) \to \mathcal{H}(\mathcal{V})$ *defined by* $\Delta := -div(\nabla f)$. *Substitute the gradient and divergence into this definition, we have*

$$L(i,j) = (\Delta f)(i) = f(i) - \sum_{i,j}\frac{w(i,j)}{\sqrt{(\varphi(i)\varphi(j))}}f(j)$$

The Laplacian is a linear operator because both the gradient and divergence operators are linear. Furthermore, the Laplacian is self-adjoint. It is easy to verify that the Laplacian L is symmetric and has row and column sums equal to zero. It can also be expressed in

$$L(i,j) = \begin{cases} d_i - w(i,j), & \text{if } i = j \\ -w(i,j), & \text{if } i \neq j,\ (i,j) \in \mathcal{E} \text{ are adjacent} \\ 0, & \text{if } i \neq j,\ (i,j) \notin \mathcal{E}, \text{otherwise} \end{cases} \tag{6.3}$$

where the term of two matrices associated with a graph as $L = D - W$ is positive semidefinite [199]. The eigenvalues of L are discrete $0 = \lambda_0 \leq \lambda_1 \leq ...\lambda_n \leq, ...$ corresponding eigenfunctions.

Definition 6.2.2.4 *The graph curvature as discrete analogue of the curvature of a surface is measured by the change in the unit normal. The graph curvature is an operator* $\mathcal{K} : \mathcal{H}(\mathcal{V}) \to \mathcal{H}(\mathcal{V})$

defined by $\mathcal{K}f := -\frac{1}{2} div(\frac{\nabla f}{\|\nabla f\|})$. *Substituting the gradient and divergence operator into this form, we obtain,*

$$(\mathcal{K}f)(i) = \frac{1}{2}\sum_{i\sim j}\frac{w(i,j)}{\sqrt{\varphi(i)}}\left(\frac{1}{\|\nabla_i f\|}+\frac{1}{\|\nabla_j f\|}\right)\left(\frac{f(i)}{\sqrt{\varphi(i)}}-\frac{f(j)}{\sqrt{\varphi(j)}}\right)$$

The curvature operator is a nonlinear operator and the Laplacian operator is a linear operator. In image processing, the Laplace-Beltrami operator is used as a linear isotropic diffusion operator. Recently, nonlinear operators [191], nonlinear anisotropic operators [259] and hyperbolic conservation laws based curvature operators [186], [184] are intensively studied in continuous regularization for improving the image restoration and visual perception in early vision.

6.2.3 Spectral Graph Clustering

As we know, physically, the original Laplace Beltrami operator on differentiable functions on manifold $\mathcal{M}$ is intimately related to the heat flow. The definition and utilization of Laplacian is very important for spectral clustering methods.

Spectral Clustering using Laplacian

Let $D = \text{diag}(\sum_i w_{1i}, ..., \sum_i w_{ni})$ be the diagonal matrix with $d_{ii} = \deg[i]$. The matrix is called the degree matrix of the graph $\mathcal{G}$ with adjacency matrix W. As we have discussed previously, the *unnormalized* graph Laplacian $L = D - W$ is defined. L is the main object in spectral graph theory [51], [168].

Given a vector $x = (x_1, ..., x_n) \in \mathcal{R}^n$, we get the following key identity in a quadratic objective function by means of the *unnormalized* Laplacian matrix $L = L(\mathcal{G})$ of the graph $\mathcal{G}$,

$$x^\top Lx = \frac{1}{2}\sum_{i,j} w_{ij}(x_i - x_j)^2 \tag{6.4}$$

This equation also shows that L is positive semi-definite, and $D^{1/2}$ is positive definite. The bisection problem can be formulated as the minimization of this identity, where as before, $L = D - W$. To see this, notice that W is symmetric and $D_{ii} = \sum_i w_{ij}$. Thus

$$\sum_{i,j}(x_i - x_j)^2 w_{ij} = \sum_{i,j}(x_i^2 + x_j^2 - 2x_i x_j)w_{ij} \tag{6.5}$$

$$= \sum_i x_i^2 D_{ii} + \sum_j x_j^2 D_{jj} - 2\sum_{i,j} x_i x_j w_{ij} = 2x^\top Lx \tag{6.6}$$

Let x be an n-vector with component $x_i = 1$ if $i \in \mathcal{A}$ and $x_i = -1$ if $x \in \mathcal{B}$, then

$$x^\top Lx = \sum_{(i,j)\in\mathcal{E}} w_{ij}(x_i - x_j)^2 = 4\text{vol}|\delta(\mathcal{A},\mathcal{B})| \tag{6.7}$$

where $i \in \mathcal{A}$, $j \in \mathcal{B}$. On the other hand,

$$x^\top Lx = x^\top Dx - x^\top Wx = \sum\nolimits_{i=1}^{n} d_i x_i^2 - 2 \sum_{(i,j)\in\mathcal{E}} x_i x_j = \sum_{(i,j)\in\mathcal{E}} (x_i - x_j)^2$$

Thus the bisection problem is equivalent to the problem of minimizing the quadratic form $x^\top Lx$ over n-vectors with components $x_i = \pm 1$ and $\sum_{i=1}^{n} x_i = 0$. Formally,

$$|\delta(\mathcal{A}, \mathcal{B})| = \min(x^\top Lx), \qquad x_i = \pm 1, \sum\nolimits_{i=1}^{n} = 0 \tag{6.8}$$

It is equal to maximizing similarity of the objects within each cluster, or, finding a *cut edge* through the graph $\mathcal{G}$ with minimal weight in the formulation of

$$\max(x^\top Wx) \Longleftrightarrow \min(x^\top Lx) \tag{6.9}$$

Since the bisection-partitioning problem is NP-complete, we cannot expect to solve this problem exactly. However, we can approximate this intractable problem by a tractable one if we relax the constraint that $x_i = \pm 1$ and let each component x_i vary continuously in value between $+\sqrt{n}$ and $-\sqrt{n}$. Thus we obtain the relax problem and its solution given by:

$$\begin{aligned} \min(x^\top Lx)_{x_i=\pm 1, \sum_{i=1}^{n} x_i = 0} \quad &\geq \min(x^\top Lx)_{\sum_{i=1}^{n} x_i^2 = n, \sum_{i=1}^{n} x_i = 0} \\ &= x_2^\top Lx_2 = \lambda_2(L) x_2^\top x_2 = n\lambda_2(L) \end{aligned} \tag{6.10}$$

where x_2 is the eigenvector corresponding to the smallest positive eigenvalue of the Laplacian matrix $L(\mathcal{G})$. The minimizer of the relaxed problem is the second eigenvector of the Laplacian. The closest partition vector to the second eigenvector is obtained by rounding the most positive $n/2$ components of the latter to $+1$, and the remaining components to -1.

The discrete optimization problem has a simple relaxation by letting x to take real values instead of $\{-1, 1\}$. A standard linear algebra argument using $L_1 = 0$ shows

$$\lambda_2 = \min_{x\in\mathcal{R}^n, xD1=0} \frac{x^\top Lx}{x^\top Dx} \tag{6.11}$$

where λ_2 is the second smallest eigenvalue of the generalized eigenvector problem $Lx = \lambda Dx$. It is clear that the smallest eigenvalue λ_1 of L is 0 and the corresponding eigenvector is 1. Moreover, the second eigenvector satisfies $\lambda_2 > 0$ in a connected graph. Thus, the eigenvector x corresponding to $\lambda_2 > 0$ is obtained by minimizing this equation. The line of reasoning is that a "cut edge" leads directly to the bipartitioning as relaxation of the weighted balanced cut.

6.2.4 Analysis of Eigenvectors

Recently, graph spectral methods have proved highly effective for image segmentation. The advantage of graph spectral methods is that they can be approximated or relaxed without the need for parallel iterative updates at the pixel level and sites. The method can also avoids the

complexity of searching. The graph spectral method is in fact the energy minimization in a cost function since the eigenvectors can be shown to be minimizers of a quadratic form. Another advantage of graph spectral methods is their stability with respect to noise.

Image is considered as a pixel-level lattice-grid graph and each pixel is a graph vertex. Since the image graph is not sparse, the computation of multiple eigenvectors is required. Therefore, a lot of authors like Shi and Malik [226], Scott and Longuet-Higgines [224], Peronan and Freeman [190], Weiss [269], Sarkar and Boyer [215], Jacobs, Weinshall and Gdalyahu [117] have suggested spectral clustering methods that are based on eigenvectors of the "affinity matrix".

In the following, we study and extend the normalized criterion to discrete regularization to achieve a global optimization. Several main reasons are summarized. First, the normalized term is modified in the optimization measure. The cost of a cut is normalized by the sum of the internal weights of a segment rather than by its area. Second, the graph is initialized with weights that directly reflect the intensity difference between neighboring regions. Third, the underlying connections between Laplacian and regularization can be unified into an integrated optimization framework.

Spectral clustering using Normalized Cuts Criterion

Shi and Malik [226] proposed a new measure of the disassociation between two groups. Instead of looking at the value of total edge weight connecting the two partitions, the cut cost is computed as a fraction of the total edge connections to all the nodes in the graph. This disassociation measure is called the normalized cut (Ncut): $Ncut(A,B) = \frac{cut(A,B)}{asso(A,V)} + \frac{cut(A,B)}{asso(A,V)}$. A and B are two initial sets. The similar objects grouping algorithm is fully exploited by an eigensolver called the Lanczos method which speeds up the running time. The degree of dissimilarity between two pieces can be computed as total weight of the edges that have been removed. The two partition criteria in the grouping algorithm is to minimize the disassociation between the groups and maximize the association within the group. They also showed an efficient computational technique based on a generalized eigenvalue problem that can be used to optimize this criterion. The minimization of this criterion can be formulated as a generalized eigenvalue problem; the eigenvectors of this problem can be used to construct good partitions of the image

For unnormalized spectral clustering, a similar argument shows that $Lx = \lambda x$ is used for unnormalized spectral clustering. In the normalized case, the second eigenvector of the generalized eigen-problem $Lx = \lambda Dx$ is equivalent to the second eigenvector of the normalized Laplacian,

$$\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{6.12}$$

The algorithm of spectral clustering using normalized cuts criterion is described in the following:

1. Given a set of features, set up a undirected weight graph $G = (V, E)$.
   Compute the weight, and summarize the information into W, L and D.
2. Find eigenvectors to the $2nd$ smallest eigenvalue of: $Lx = (D - W)x = \lambda Dx$
3. Obtain the partition: $A = [i] : x_i \geq 0, B = [i] : x_i \geq 0$ using the $Ncut$.
4. Decide if the current partition should be subdivided by checking
   the stability of the cut, and make sure Ncut is below pre-specified value.
5. Recursively repartition the segmented parts if necessary.

In image segmentation algorithms based on normalized cuts [226], one attempt to find the second smallest eigenvector of the matrix $D - W$ where W is a $n \times n$ pixels matrix whose elements are the pairwise affinities between pixels (i.e., the i, j entry of the matrix is w_{ij}) and D is a diagonal matrix whose diagonal elements are the sum of the affinities(i.e., equals 1). The second smallest eigenvector of any symmetric matrix A is a unit norm vector x that minimizes $x^\top Ax$ and is orthogonal to the first eigenvector. By direct inspection, the quadratic form minimized by normalized cuts is exactly the cost function J, that is $x^\top(D - W)x = J(x)$.

Thus, the algorithm minimizes the same cost function but under different weight constraints.

$$W = w_{(i,j)} = \exp^{d_{ij}/(2\sigma^2)} \tag{6.13}$$

where σ is a free parameter. d_{ij} can be represented in different affinity structures which measure the similarity between image features. Here, we use $d_{ij} = \|x_i - x_j\|^2$ to measure some vector data set in vector spaces, shown in Fig. 6.5. Several synthetic data clusters are clustered using this criterion. We also compared eigen-vectors for segmentation of the blurred and the unblurred image using the normalized cut criterion, shown in Fig. 6.6

6.3 Regularized Spectral Graph Clustering for Perceptual Image Segmentation

Our goal is to design a practical regularization approach that is adopted to the structure of graphs for nonuniform(partially-blurred) blurred image restoration. The detection and partition of blurred regions or objects is the first crucial step. In this approach, we consider the partially-blurred image as the composite of blurred regions and unblurred regions in linear blending of radiance values based on pixel-level information. The sum of blurred regions αF and unblurred regions $(1 - \alpha)B$ is equal to the entire image I.

$$I = \alpha F + (1 - \alpha)B \tag{6.14}$$

where α is the opacity of blurred regions or objects. The blurred regions can be formulated as $\alpha F = h * f + \eta$,

$$I = (h * f + \eta) + (1 - \alpha)B \tag{6.15}$$

where η is additive white Gaussian noise η, $(1 - \alpha)B$ is the rest part of unblurred regions or objects. This equation brings us several meaningful interpretations. First, one interpretation of this form is to reduce the dimensionality, either by extracting blurred regions, by extracting unblurred regions or by combining two distributions linearly into an entire image. Second, such interpretation of combination can directly avoid overlapping or over-fitting problem [63] for searching different classes of local distributions. There are some related similar assumptions in image processing and vision. Förstner [79] has used similar model for the detection of feature operators (inlier and outlier feature operators), where the distribution of outlier (here F_i) is simulated the Laplace-distribution or the Cauchy-distribution. Image matting methods [214], [50] assume the image based on foreground matting objects and background scene based on earlier proposed image matting techniques [76], and some extended methods [142]. Graph cuts methods measure the energy distance between the source and the target using maximum-flow and minimum cut theorem. Keuchel et al. [130] apply binary decision subject to linear constraints to a combinatorial problem. Spectral bi-section is to classify objects in two classes.

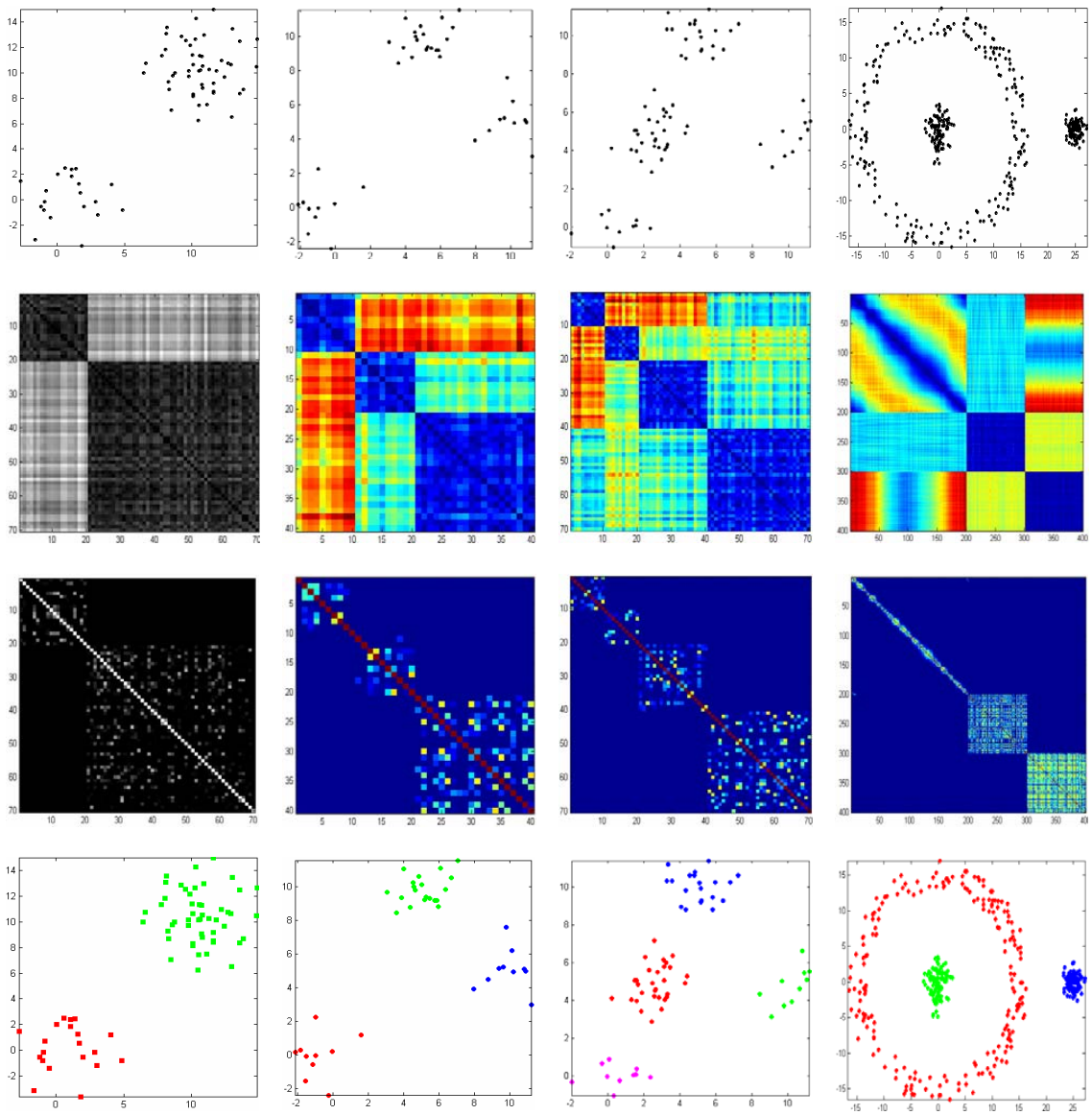


Figure 6.5: $a|b|c|d$ columns. Segmentation using the second generalized eigenvector with normalized cut criterion $(D-W)x = \lambda Dx, Wx = (1-\lambda)Dx$. In (a)(b)(c)(d) columns (from top to down): A simple clustering problem, the affinity matrix, the corresponding graph weight matrix W, and the clustering results.

6.3.1 Regularized Spectral Graph Clustering

Our goal is to design a discrete regularization approach that is adopted to the structure of graphs and prior information (labeling) for perceptual and optimal image segmentation. We consider the image as the composite of foreground regions F and background regions B in linear blending of radiance values based on i-th pixel-level information. The sum of foreground αF and background $(1-\alpha)B$ is equal to the entire image I,

$$I = \alpha F + (1-\alpha)B \tag{6.16}$$

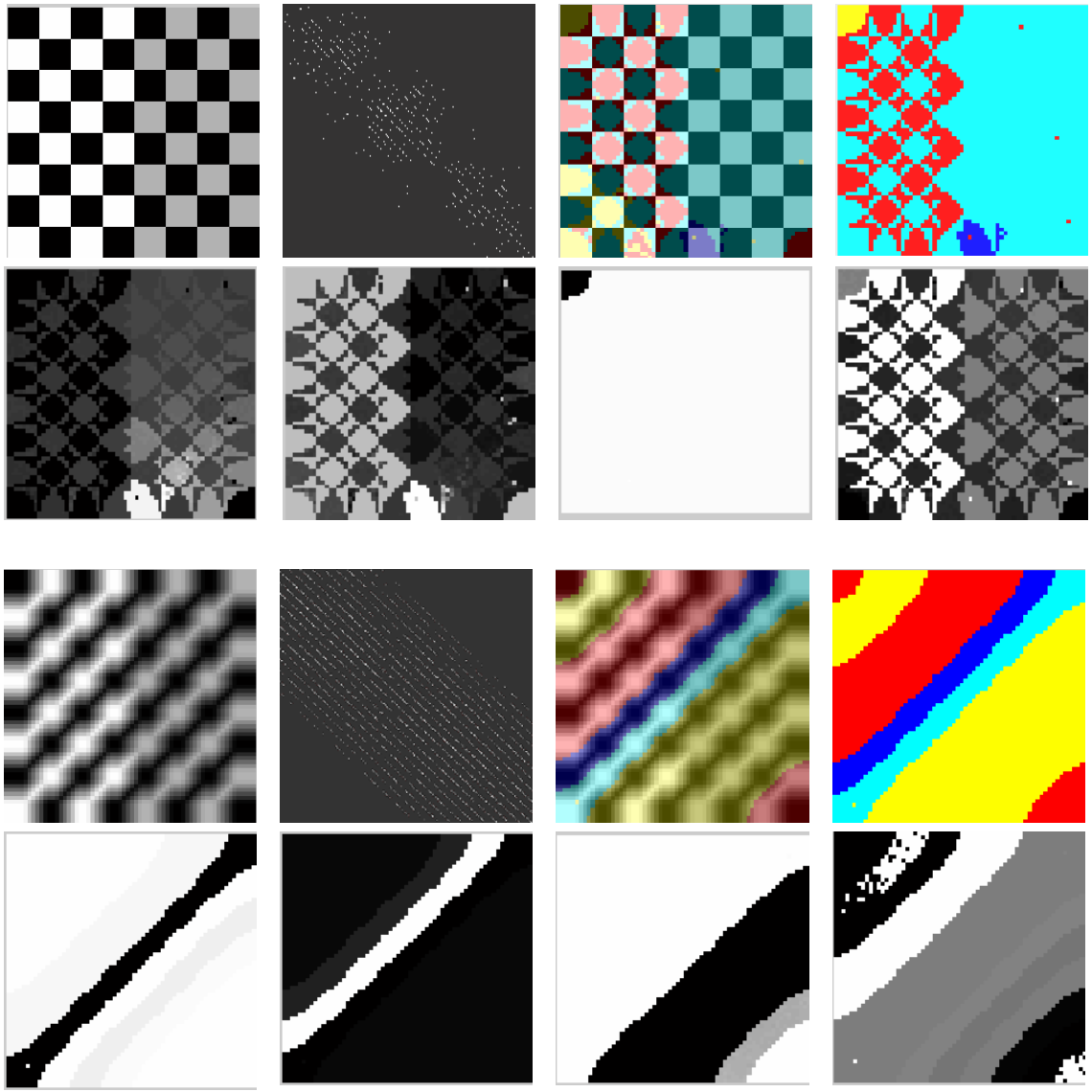


Figure 6.6: $\frac{a}{b}$ groups. Comparison of synthetic unblurred (checkerboard(8)) and blurred image (motion blur with angle at 45 degrees and 8 pixels strength) in group (a) and group (b). **In group(a) and group (b):** the first row (from left to right) is the test image (1st), the corresponding graph weight matrix W (2nd), semi-transparency marked clustering regions (3rd), color marked clustering regions (4th). The second row (from left to right) shows the eigenvectors corresponding to the second smallest to fifth smallest eigenvalues of the system. The eigenvectors are reshaped to have the size of the image.

where α is the opacity of foreground regions or objects. $(1-\alpha)B$ is the rest part of background regions or objects. This equation brings us several meaningful interpretations, e.g, reducing the computation complexity and avoid over segmentation, etc. The computation of α is crucial to segment foreground regions. We use a transform to simplify the formulas by allowing $u = 1/(F-B)$, $v = -B/(F-B)$, the Eq. 6.16 becomes $\alpha = uI + v$, where I is the input image, and output parameters α, u and v. For an entire image, the cost function on discrete image spaces with respect to i-th pixel-vertex becomes,

$$J(\alpha, u, v) = \arg\min\{\sum_{k\in I}(\sum_{i\in w_k} \|I_i u_k + v_k - \alpha_i\|^2 + \varepsilon u_k^2)\}$$

where εu^2 is a penalty smoothing term with parameter ε, and w_k is a small window around the pixel k.

For each small window w_k in the image, the solution can be formed in a least squares form,

$$J(\alpha_k, u_k, v_k) = \sum_k \left\| \begin{bmatrix} u_k \\ v_k \end{bmatrix} \Psi_k - \bar{\alpha}_k \right\|^2 \tag{6.17}$$

where k is a pixel vertex. Ψ_k is defined as a matrix $(|w_k| + 1) \times 2$ and contains a row of the form $[I_i, 1]$ for each window $i \in w_k$ and the last row of Ψ_k is $[\sqrt{\varepsilon}, 0]$. The partition region $\bar{\alpha}_k$ is a $(|w_k| + 1)$ vector with elements $\alpha_i, (i \in w_k)$ and the last element is 0, $|w_k|$ is the number of pixels in this window. To solve the segmented regions α_k, the optimal $\hat{u}_k$,$\hat{v}_k$ is the solution to the minimization of least squares (LSQ) problem.

$$(\hat{u}_k, \hat{v}_k) = \min \left\| \begin{bmatrix} u_k \\ v_k \end{bmatrix} \Psi_k - \bar{\alpha}_k \right\|^2 = (\Psi_k^\top \Psi_k)^{-1} \Psi_k^\top \bar{\alpha}_k$$

Substituting this solution into the energy minimization in Eq. 6.17, we get a quadratic cost function with unknown α.

$$J(\alpha, u, v) = \sum_k \left\| \Psi_k (\Psi_k^\top \Psi_k)^{-1} \Psi_k^\top \bar{\alpha}_k - \bar{\alpha}_k \right\|^2 \tag{6.18}$$

where we denote $\bar{\Psi}_k = I - \Psi_k (\Psi_k^\top \Psi_k)^{-1} \Psi_k^\top$, then we get

$$J(\alpha) = \sum_k \left\| \bar{\Psi}_k \bar{\alpha}_k \right\|^2 = \sum_k \bar{\alpha}_k^\top \bar{\Psi}_k^\top \bar{\Psi}_k \bar{\alpha}_k = \alpha^\top L \alpha \tag{6.19}$$

where $L = \bar{\Psi}_k^\top \bar{\Psi}_k$ can be expressed in the following,

$$K_{ij} - \frac{1}{|w_k|} \left(1 + (I_i - \delta_k)(I_j - \delta_k)(\sigma_k^2 + \frac{\varepsilon}{|w_k|})^{-1} \right) \tag{6.20}$$

where K_{ij} is the kronecker delta, δ_k and σ_k^2 are the mean and variation of the intensities in the window w_k around k. We refer the semidefinite matrix L to an affinity Laplacian matrix. The matrix L can also be explained in $L = D - W$ in spectral graph theory, with $D(i,i) = \sum_j W(i,j)$ is a diagonal matrix. The W is a symmetric matrix and its off-diagonal matrix are defined by the definition of weights.

6.3.2 Semi-supervised Learning and Labeling: From Local Patches to Global Image Understanding

Our target is to identify and segregate blurred and unblurred regions in an unsupervised manner. Pairwise difference between blurred and unblurred regions or objects (gradient, frequency, entropy, etc.)[67], [80], [153] are one kind of useful empirical image statistics. Therefore, to enhance pairwise differences and attenuate the difference inside of both regions can be a reasonable

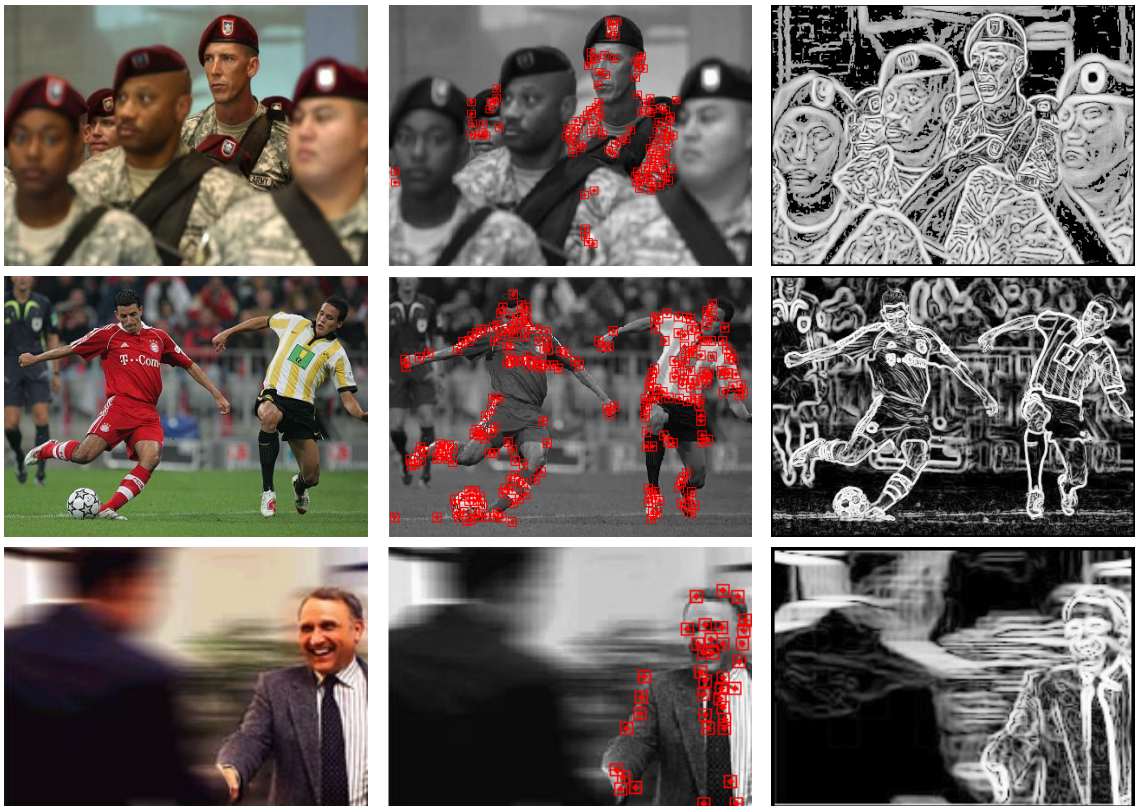


Figure 6.7: $a|b|c$. Unsupervised feature operators and gradients. (a) Partially-blurred images. (b) Unsupervised labeling using feature corners on unblurred regions is prior for partition. (c) Pairwise differences of edge gradients between blurred and unblurred regions.

way to improve the partition results. We transfer the empirical knowledge (high-level) as prior labeling to guide low-level segmentation processing. We combine the edge gradient difference (edge prior) and unsupervised feature operators labeling (feature patches) [79], to collect the affinities and segregate the dissimilarities, shown in Fig. 6.7.

To extract the optimal opacity of blurred regions α, we construct a regularization energy function which can use these unsupervised patch prior and gradient prior,

$$\alpha = \arg\min\{\alpha^\top L\alpha + \xi(\alpha^\top - d_l^\top)D_l(\alpha - d_l)\} \tag{6.21}$$

where ξ is a regularization parameter and denotes the strength of smoothing penalty term. This smoothing penalty term is adjusted by prior labeling patches d_l. d_l is the vector containing the unsupervised patch-values of labeling and 0 for all other pixels. D_l is a diagonal matrix with diagonal value 1 for detected feature patches and 0 for all other pixels. Since this energy function follows quadratic regularization, we can differentiate the equation and set the derivatives to 0. This equation can be well adapted to use patch labeling (detected feature patches) and pairwise difference of edge gradients for partitioning the blurred and unblurred regions or objects. It also allows a globally optimal partition of α using these sparsely distributed priors via this equation.

Figure 6.8: $a|b|c$ columns. The detection of feature corners in background and foreground are controlled by a nonlinear image diffusion filter. After nonlinear image diffusion, some stronger edges or corner are enhanced, while some weak edges or corners are eliminated, e.g., boundary of car, power line and the tree behind the house. (a) column: test images. (b) column: The detection of feature corners is performed on original partially-blurred images. (c) column: The detection of feature corners is performed after nonlinear image diffusion, K = 10. For such methods it can be shown that small scales are smoothed faster than large ones, so if the method is stopped at a suitable final time and given a suitable K, we may expect that noise is smoothed while large-scale features are preserved to some extent. By this way, we can judge the distribution of feature corners using nonlinear image filtering.

6.3.3 Maintenance of Foreground and Background

Nonlinear Filtering Adjusting Feature Detection

As we have discussed in the 2nd chapter, Perona and Malik's nonlinear diffusion filters has some different roles on image diffusion which can be further developed for different targets.

$$\mathcal{C}(\nabla I) = \exp^{\frac{\|\nabla I\|^2}{K^2}}, \text{ and } \mathcal{C}(\nabla I) = \frac{1}{1 + \frac{\|\nabla I\|^2}{K^2}} \tag{6.22}$$

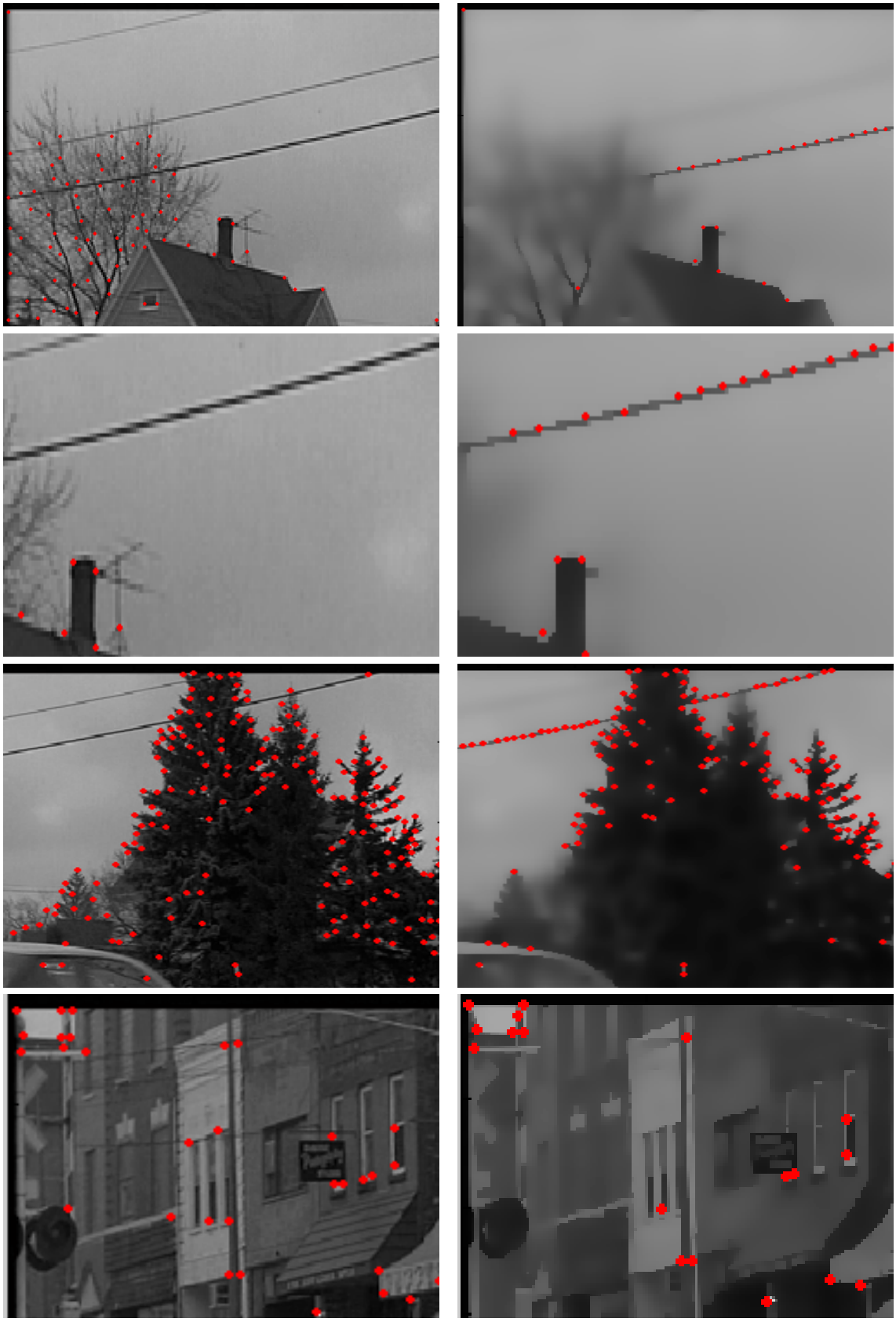


Figure 6.9: *a|b.* Comparison of feature corners in original and diffused images (Zoom in images) in unblurred regions. (a) column: Zoom in original images. (b) column: Zoom in diffused images with feature detection.

These two functions generated by the scale-spaces are different: the first privileges high-contrast edges over low-contrast ones, the second privileges wide regions over smaller ones. These characteristic properties can be further extended to control the distribution of feature detection and strength of affinity weights. Recently, these two diffusion functions have been also considered as diffusion kernels for controlling the optimal classification on graph spaces in machine learning community.

Inspired by these difference in these filters, we use the $\mathcal{C}(\nabla I) = 1/\left(1 + \frac{\|\nabla I\|^2}{K^2}\right)$ for improving the unsupervised detection of feature corners and label the unblurred regions with sparse feature corners. Through the experiment, it can be shown that small scales are smoothed faster than large ones, so if the method is stopped at a suitable final time and given a suitable K, we may expect that noise is smoothed while large-scale features (most are unblurred discontinuities) are preserved to some extent. By this way, we can judge the feature corner detection using nonlinear image filtering. Fig. 6.8 and Fig. 6.10 show that partially-blurred images can get more accurate labeling of unblurred regions after diffusion. Fig. 6.9 shows that weak strength image discontinuities (small scale, tree behind house, one power line) are eliminated, while large scale objects (one power line can get more feature corners after diffusion) are enhanced.

Therefore, the unblurred regions and blurred regions can get more accurate labeling using optimal controlled nonlinear diffusion filters.

Maintenance using Adaptive Nonlinear Diffusion

To achieve high-quality partition, an idea is to enhance the difference between blurred and unblurred regions or object. That is to say, intraclass differences need to be decreased and interclass difference need to be increased. Based on this idea, we attenuate the contrast inside one blurred or unblurred regions (intraclass) so that the boundary between these two kinds of regions (interclass) can be enhanced and increased. Simultaneously, other inner edges or inner discontinuities in one region are eliminated.

Motivated by nonlinear inhomogeneous diffusion [191], [235], we extend the nonlinear diffusion method to attenuate the difference inside the background region while preserving the contrast Z_{ij} across the boundaries between blurred and unblurred regions.

$$Z_{ij} = ||I_i - I_j||^2 \cdot \frac{1}{1 + |\nabla I_B|^2/K^2 \cdot \exp(-\frac{d_{ij}^2}{\sigma_d})} \tag{6.23}$$

where K is a contrast parameter to be tuned for a particular application. This diffusion filter is scalar-valued, decreasing with isotropic but non-homogeneous effects. d_{ij} measures the dissimilarity between pair (I_i, I_j) in the image I and $\nabla I_B = I_i^B - I_j^B$ measures pairwise pixel difference inside background I_B (Here we only measure the background). The $d_{ij} = \max\left\{|I_i - I_i^B|, |I_j - I_j^B|\right\}$ is a Hausdorff distance-like definition. The Hausdorff distance has been widely used for image matching, object recognition due to its underlying properties [116]. If d_{ij} is small, the attenuation strength should be large ($\exp(-d_{ij}^2/\sigma_d) \to 1$), and the pixel pair (I_i, I_j) might belong to the same region(blurred or unblurred). Otherwise, if d_{ij} is large ($\exp(-d_{ij}^2/\sigma_d) \to 0$), the attenuation strength is small, and it probably belongs to the boundary contrast between blurred and unblurred regions. In our experiments, we take $K = [5, 10]$ and $\sigma_d = [10, 50]$.

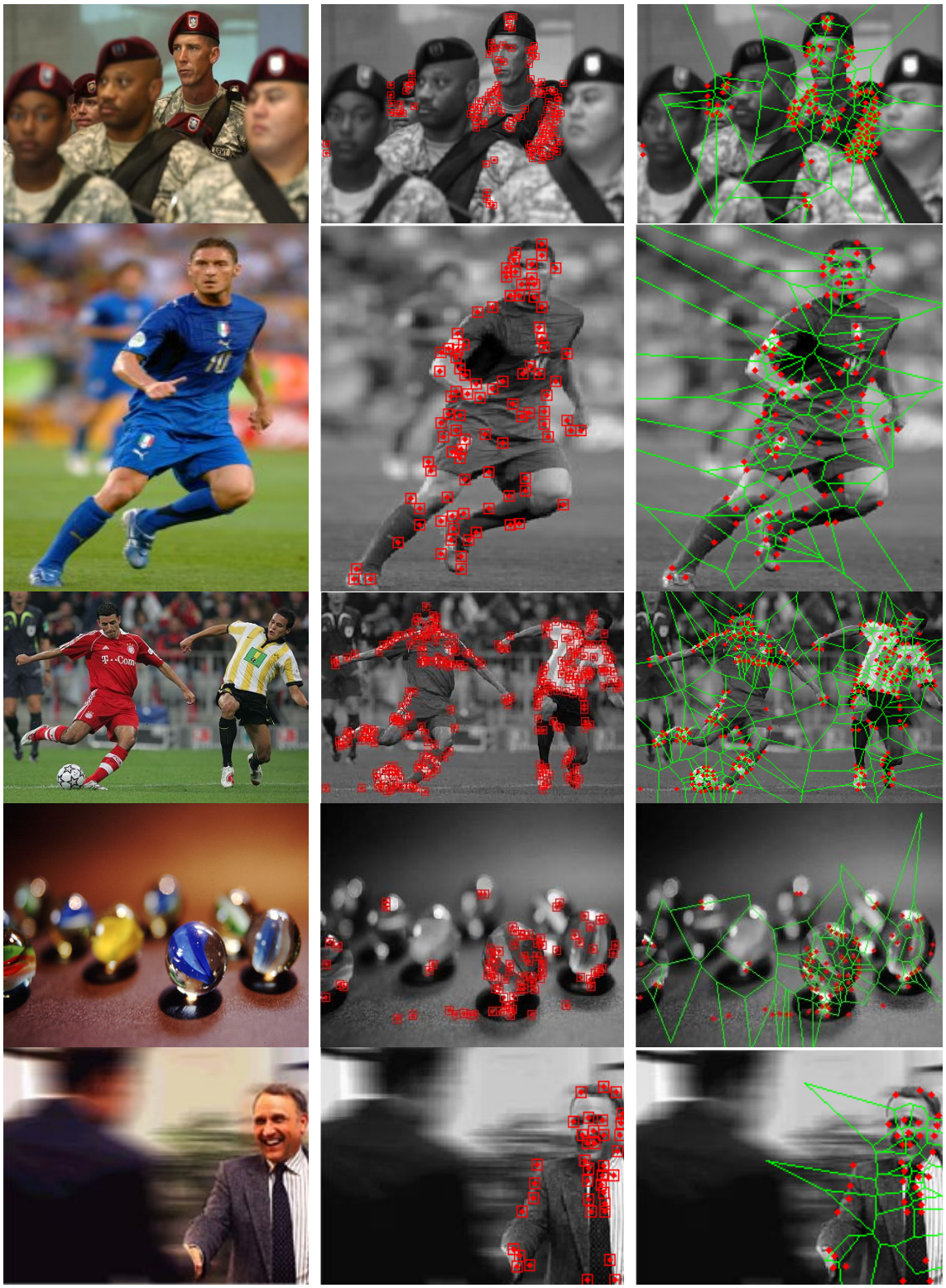


Figure 6.10: (a) Partially-blurred images (b) Unsupervise detected feature corners are a natural prior for labeling unblurred regions in the suggested method. (c) Unblurred objects or regions have highest feature density in Voronoi

6.4 Variational Bayesian Learning for Nonuniform Blurred Image Reconstruction

The Bayesian approach is, in fact, the framework in which the most recent blur kernel estimation methods have been introduced, e.g, simultaneous kernel estimation and image restoration [289], estimating Bayesian hyperparameters [169], factorizing kernels into parametric models [166], [167], [289], and measuring the strength of discontinuities in Gaussian scale space [67], etc. However, these methods are limited in certain parametric models to stationary blurred images.

This section presents the method for restoring uniform and nonstationary blurred images in a variational Bayesian ensemble learning framework. First, through large observations and experiments, we classify natural blurred images into three main blurred groups so that we can design an efficient methods. As a result, we obtain an approach, which can compute and use the translation and scale-invariant marginal probability distribution of image gradients as *a priori* through the Bayesian learning scheme. In a sense, the distribution can be shared by most similar type of blurred images and therefore requires relatively few training images.

Based on variational Bayesian approaches [114], [14], Miskin and Mackay [166], [167] have firstly applied this method to deal with blind deconvolution using a prior on raw pixel intensities. Results are shown on synthesized image blur. Using image statistical prior, Fergus et al. [73] have extended this method for removing camera shaking blur from a single blurred image. The blur kernel is estimated and interpolated in high-accuracy using a multi-scale approach [227]. Although the ringing effects has been observed by Fergus et al., the image deblurring is directly using an extended Richardson-Lucy (RL) method without using image statistical prior and local spatial conditions for deblurring. Inspired by Fergus's et al, in our approach, we use image statistical prior not only for kernel estimation but also for weighted space-adaptive image deblurring with ringing reduction.

For image deblurring, ringing effects and amplified noises influence the results due to Gibbs phenomena in Fourier transformation. One type of ringing effects often happens around edges and discontinuities due to the high frequency loss during blurring. The other type of ringing effects is due to the mismatch between nonstationary real blurred images and stationarity assumption. Such phenomena have been observed by [289], [73], [135]. Since most original scenes are without ringing, such restoration results are usually undesirable. Therefore, the deblurring approach needs to be designed for both two types of ringing reduction.

Furthermore, in Bayesian estimation, a generic prior model needs to represent common descriptive or generative information from an observed image. Such prior distribution can be translation and scale-invariant for representing a global image. Natural image statistics based prior learning has such properties to represent image structure, textures [109], discontinuities and blurred edges [73]. On the other hand, natural images are often inhomogeneous with piecewise uniform regions separated by edges and discontinuities. Therefore, the measure of distributions of local edges, textures as well as the pixel intensity values can be used as local spatial conditions for ringing reduction in image reconstruction.

Different from Fergus's work [73], [166], [167], our approach has several effects. First, through some observations and experiments, we classify natural blurred images into three main groups so that we can design an efficient method. Second, in contrast to previous work [166], [167], [73], [197], natural image statistics is used not only for kernel estimation in a global image but also for piecewise image reconstruction in a newly designed regularization function. Therefore,

we obtain an approach, which can use the scale-invariant statistical prior for kernel estimation and integrate with local spatial conditions for deblurring with ringing reduction.

6.4.1 Natural Image Statistics for Prior Learning

The objective of learning a generic prior model is to look for common descriptive or generative information from observed natural images. Such information are then incorporated into a probability distribution as a prior model which will bias learning algorithms. For this objective, find a translation and scale-invariant prior distribution is expected. Natural images statistics has some properties to represent image structure, textures, and discontinuities so that it has great potentials to find such prior distribution.

From a combination of psychophysical and computational approaches, Field [74], [75] has presented that real cluttered images obey heavy-tailed distributions in their gradients. The distribution of gradients has most of its mass on small values but gives significant probability to large values than a Gaussian distribution but rather a Student's t-distribution. Later, Olshausen and Field [179] have proposed an approach to understanding such response properties of visual neurons and their relationship to the statistical structure of natural images in terms of efficient coding.

From signal and image processing approaches, Mallat [156], and Simoncelli [227] have described that non-Gaussian nature of the statistical distribution, e.g., high kurtosis, heavy tails, it is a similar distribution as an exponential density with exponent less than 1. These heavy-tailed natural image priors have shown the usefulness in state-of-the-art methods, e.g., image segmentation [109], denoising [227], [209], removing camera shake [73], Gibbs-reaction diffusion [303] and so on.

To compute such distributions, one way is compute the joint statistics of derivative filters at different locations, size or orientations [197]. The other way is to observe marginal statistics of more complicated feature detectors [305]. In this paper, we extend these methods to yielding translation and scale invariant prior. For example, Fig. 6.11 illustrates this fact and shows several natural images and their histogram of gradient magnitudes. Similar histogram are observed for vertical derivative filters and the for the gradient magnitude ∇I_x and ∇I_y.

6.4.2 Construction of Variational Bayesian Estimation Model

Following a Bayesian paradigm, the true f and the PSF h will be estimated by using only a given observed g. However, how to efficiently build the Bayesian estimation which can optimally use the information from the observed image? During the blurring, the changes of image discontinuities and edge gradients are larger and more representative than the changes in the homogeneous regions. Therefore, we construct a probabilistic model based on marginal distribution of image gradients. We process the gradients of f and g and construct the new convolution equation using the original equation $g = h * f + \eta$. Suppose we have a model which tells how a number sequence $\nabla f = \nabla f(1), ..., \nabla f(t)$ transforms into sequence $\nabla g = \nabla g(1), ..., \nabla g(t)$, we then have $\nabla g(t) = \nabla f(t) * h + \eta(0, \sigma^2)$ with zero-mean identical and independently distributed Gaussian noise.

Based on this model, the Bayesian MAP estimation utilizes an input $P(\nabla g)$ to achieve two

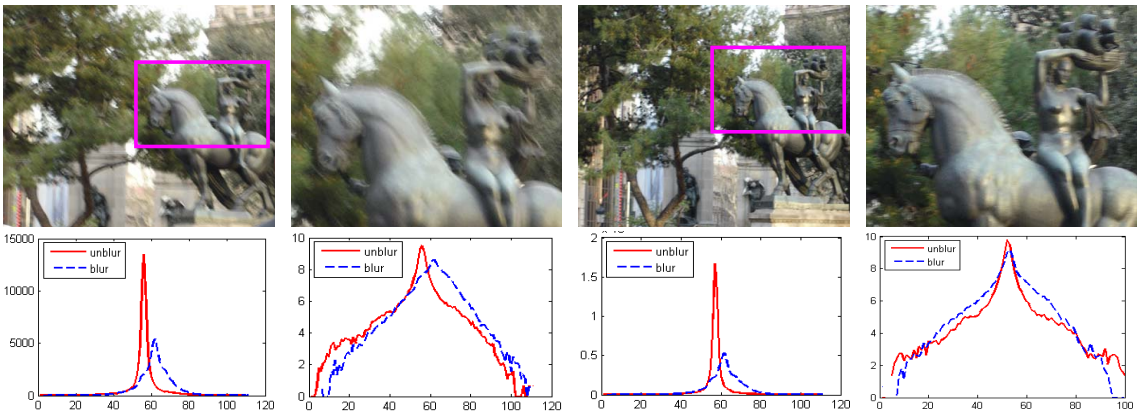


Figure 6.11: $\frac{a|b|c|d}{e|f|g|h}$. Comparison of marginal distribution of blurred and unblurred gradients. (a)(b) Blurred image. (c)(d)Unblurred image. (e)(f) Histogram and Log-histogram (y) of gradients $\nabla_x I$. (g)(h) Histogram and Log-histogram (y)of gradients $\nabla_y I$.

convergent posteriors $p(h)$and $p(\nabla f)$, and is formulated in

$$p(\nabla f, h|\nabla g) = \frac{p(\nabla g|\nabla f, h)P(\nabla f, h)}{p(\nabla g)} \propto p(\nabla g|\nabla f, h)P(\nabla f)P(h) \tag{6.24}$$

In $p(\nabla f, h|\nabla g)$, we can easily have more stable prior distribution, e.g., log-histogram of image gradients. However, computing the full posterior distribution $p(\nabla f, h|\nabla g)$ is more difficult than computing $p(f, h|g)$ in direct Bayesian MAP estimation using normal gradient descent methods. The minimization of all gradients ∇g in $p(\nabla f, h|\nabla g)$ is not the right output value, while we always want to have high image gradients for the restored images. Moreover, marginalizing the posterior distribution is difficult. We cannot take a point estimate (e.g.,the MAP estimate) because this leads to overfitting. This is because the MAP estimate does not guarantee a high probability mass in the peak of the posterior distribution and so the posterior distribution may be sharp around the MAP estimate. Therefore it is necessary to approximate the posterior density by a more tractable form for which it is possible to perform any necessary probability mass of the posterior.

In order to apply the Bayesian approach for modeling, the model needs to be given in probabilistic terms, which means stating the joint distribution of all the variables in the model. In principle, any joint distribution can be regarded as a model, but in practice, the joint distribution will have a simple form.

Joint Posterior Distribution

According to $\nabla g(t) = \nabla f(t) * h + \eta(0, \sigma^2)$, the prior $P(\nabla f)$ on the restored image gradients is a Gaussian mixture model with variance v_i and weight w_i for the i-th Gaussian ($i \in N$). The blur kernel prior $P(h)$ is a mixture of K blur kernel parametric models with exponential distributions and the size factors s_k and weights w_k for the k-th distribution component. Therefore, the joint density of all the variables in Eq. **??** can be formulated for posterior distribution given the image gradient distribution $P(\nabla g)$,

$$p(\nabla f, h|\nabla g) \propto p(\nabla g|\nabla f, h)P(\nabla f)P(h) = \tag{6.25}$$

$$\prod_t (\nabla g(t)|h * \nabla f(t), \sigma^2) \prod_t \sum_{i=1}^{N} w_i \mathbb{G}(\nabla f(t)|0, v_i) \prod_{t'} \sum_{k=1}^{K} w_k \mathbb{E}(h_k|s_k)$$

where t indexes over image pixels and t' denotes blur kernel pixels. $\mathbb{G}$ and $\mathbb{E}$ denote Gaussian and Exponential distributions respectively. For the application of these equations, some constraints of the PSF and the image are assumed due to the fact that the image pixels are independent identically distributed and does not influence the pixel correlations.

In the field of statistical approximate inferences, mean field methods [181], [182], variational free energy [183] have also been investigated intensively. According to the variational methods for graphic models [121], Attias [14], the posterior distribution in Bayesian estimation can be simplified in variational transformations based on convex duality. The original full posterior $p(\nabla f, h|\nabla g)$ is then approximated by a tractable distribution $q(\nabla f, h)$ by minimizing the Kullback-Leibler information which acts as a distance measure between the two distributions. The tractable distribution can be further processed in an ensemble learning approach.

6.4.3 Variational Ensemble Learning for Blurred Regions Reconstruction

Ensemble learning [114], [166] is a method for parametric approximations of the posterior distributions. It assumes a Gaussian distribution or other parametric distribution, but in which the mean and the variance are allowed to evolve during the learning process. Based on Miskin and Mackay's ensemble learning method [166], [73], the distributions for each estimated gradients and blur kernel element are represented by their mean and variance. The variational ensemble learning can be expressed in terms of a minimization of the Kullback-Leibler distance between the model distribution and the true posterior. It is formulated as,

$$\begin{aligned} KL\{q(\nabla f, h)||p(\nabla f, h|\nabla g)\} &= \int q(\nabla f, h) \ln \frac{q(\nabla f, h)}{p(\nabla f, h|\nabla g)} d\nabla f dh \\ &= \int q(\nabla f, h) \ln \frac{q(\nabla f, h)}{p(\nabla g|\nabla f, h) P(\nabla f, h)} d\nabla f dh + \ln p(\nabla g) \end{aligned} \tag{6.26}$$

The Kullback-Leibler information is greater than or equal to zero, with equality if and only if the two distributions, $p(\nabla f, h|\nabla g)$ and $q(\nabla f, h)$ are equivalent.

Training and learning the approximating ensemble can be done by assuming a fixed parametric form for the ensemble (for instance assuming a product of Gaussians). As a consequence, the parameters of the distributions can be set to minimize the cost function. Therefore, the $q(\nabla f, h) \rightarrow q(\nabla f, h, \sigma^2)$ can be further approximated by adding a noise prior σ^{-2}(inverse variance) in the form of a Gamma distribution according to [166]. Thus, we have hyper-parameters $x, y : p(\sigma^2|x, y) = \Gamma(\sigma^{-2}|x, y)$. The variational posterior is $q(\sigma^{-2})$ in a Gamma distribution. If we note that the term $p(\nabla g)$ is a constant over all the models, we can define a cost function C_{KL} which we are required to obtain the optimum approximating distribution,

$$\begin{aligned} C_{KL} &= KL\{q(\nabla f, h, \sigma^2)||p(\nabla f, h|\nabla g)\} - \langle \ln p(\nabla g) \rangle \\ &= \int q(\nabla f) \ln \frac{q(\nabla f)}{p(\nabla f)} d\nabla f + \int q(h) \ln \frac{q(h)}{p(h)} dh + \int q(-\sigma^2) \ln \frac{q(-\sigma^2)}{p(-\sigma^2)} d(-\sigma^2) \end{aligned} \tag{6.27}$$

Where the subindex of C_{KL} denotes the variables that are marginalized over in the cost function, $\langle \ln p(\nabla g) \rangle$ is the average over all variables. In general, they are the unknown variables of the model. Because of the product from of the true posterior density, the cost function C_{KL} can be factorized into a sum of simpler terms.

On the other hand, the Kullback-Leibler information is a global measure, providing that the approximating distribution is a global distribution. Therefore, the measure will be sensitive to probability mass in the true posterior distribution rather than the absolute value of the distribution itself.

According to the cost function C_{KL}, the parameters of the distributions are minimized alternately using the coordinate descent method. The most crucial part is the initial value that we choose the means of the distributions $q(h)$ and $q(\nabla f)$ (a trained prior distribution from other similar type of blurred images). The variance σ^2 is given high value due to the uncertainty of the initial value. The minimization are repeated until the change in C_{KL} becomes negligible.

According to the cost function C_{KL} in Eq. 6.26, the parameters of the distributions are minimized alternately using the coordinate descent method. The most crucial part is the initial value that we choose the means of the distributions $q(h)$ and $q(\nabla f)$ (a trained prior distribution from other similar type of blurred images). The variance σ^2 is given high value due to the uncertainty of the initial value. The minimization are repeated until the change in C_{KL} becomes negligible. The ensemble learning algorithm is provided online by Miskin and Mackay [166]. Furthermore, multi-scale [227], [73] and multigrid [37] methods have been proven to be very useful in computer vision. These methods can avoid local minima. Following Fergus et al. [73] and Simoncelli [227], we implement our algorithm using multi-scale based coarse-to-fine refinements. At the coarsest level, the blur kernel is initialized at very coarse level. The initial estimation for ideal image gradients is then adapted to the blur kernel till the edge gradients distribution is well adjusted. At the finest resolution, the blur kernel is full interpolated.

6.4.4 Image Deblurring and Reconstruction without Ringing Effects

Analysis of Ringing Effects

According to $g = h * f + \eta$, using the Tikhonov-Miller regularized solution, the restored image $\hat{F}$ in the frequency domain is,

$$\hat{F}(u,v) = \frac{H^*(u,v)}{|H(u,v)|^2 + \alpha |L(u,v)|^2} G(u,v) = T(u,v)G(u,v) \tag{6.28}$$

where G, H, F are the DFT of g, h, f, respectively, (u, v) are the spatial frequency variables, $L(u, v)$ represents a regularizing operator with a regularization parameter α. $T(u, v)$ deviates from the inverse of the blur kernel $H^{-1}(u, v)$. The deviation is expressed by the error spectrum $E(u, v; \alpha) = 1 - T(u, v; \alpha)H(u, v)$. The restored image $\hat{F}$ in the frequency domain is given by,

$$\begin{aligned} \hat{F}(u,v) &= T(u,v;\alpha)[H(u,v)F(u,v) + \eta(u,v)] & (6.29) \\ &= F(u,v) - E(u,v;\alpha)F(u,v) + (1 - E(u,v;\alpha))H^{-1}\eta(u,v) & (6.30) \end{aligned}$$

where the restoration error is $\|\hat{F}(u,v) - F(u,v)\|$. On the right side, the second term denotes the error due to the use of filter T, i.e., a regularization error; the third term presents the noise

η magnification error. There exists an optimal value α between two types of errors. The noise magnification error has a global degrading effects resulting from the observed noise. Also, the regularization error is a function of F, and its effect will therefore be related strongly to the local spatial structures encountered within the image. Ringing effects can be seen as a structure dependent phenomenon and can be classified as a regularization error.

Iterative Reweighted Regularization for Deblurring and Reconstruction

Therefore, we propose an iterative reweighted regularization function which can use the measure distributions of local edges, textures as well as the pixel intensity values for image deblurring. Similar to Eq. 6.24, $p(g|f,h)$ follows a Gaussian distribution and $p(f)$ is prior with some constraint conditions,

$$\mathcal{J}(f|g,h) \propto \arg\min\{\frac{1}{2}\sum w_1(g(x) - h(x) * f(x))^2 + \frac{1}{2}\lambda \sum w_2(c_1(x) * f(x))^2\}$$

where $\mathcal{J}(f|h,g) = -\log\{p(g|f,h)p(f)\}$ express that the energy cost $\mathcal{J}$ is equivalent to the negative log-likelihood of the data [181], [289], [292]. λ is a regularization parameter that controls the trade-off between the fidelity to the observation and smoothness of the restored image. The smoothness constraint $c_1(x)$ is an regularization operator and usually is a high-pass filter. The energy function achieves an optimal result by searching for f minimizing the reconstruction error $(g - h * f)^2$ and the weights prior w_2 controlling f to be satisfactorily smooth.

The weights w_1 and w_2 reduce these ringing effects adaptively to achieve better visual evaluation. $w_1 = 1$, if data at x is reliable, otherwise $w_1 = 0$; the image weight $w_2 = 1/[1 + k\hat{\sigma}_f^2(x)]$, $\hat{\sigma}_f^2(x)$ is local variance of the observed image $g(x)$ at x in a $P \times Q$ window, k is a contrast parameter. However, it is difficult to directly compute such local variances in a small moving window for a single blurred image and its unknown ideally restored image. In contrast to most existing approaches [289], [73], we use the distributions of statistical edge gradients as the local prior weights, which can bias the results. We use a $w_2' = \exp^{k\hat{\sigma}_f^2(x)}$ from a general exponential function family and has similar effects as w_2 [191]. The heavy-tailed curve of w_2' is directly controlled by using the image statistical prior distribution. The cost function of this equation is minimized in an iterative reweighted optimization approach [178] via conjugate gradient descent.

6.5 Experimental Results

Experiments on synthetic and real data are carried out to demonstrate the effectiveness of our algorithm. At the first step, we propose to use edge gradient prior based on spectral graph clustering methods for identifying and segmenting blurred regions or objects, shown in Fig. 6.12. However, this approach can not achieve optimal segmentation results. The detailed explanation will be presented on experiments parts. Second, we propose a novel method based on the integration of sparse (unsupervised or semi-supervised learning and labeling) labeling prior and regularization on graph spaces. This approach can achieve high quality and perceptual image segmentation for nonuniform (partially-blurred) blurred images via optimal control, shown in Fig. 6.16. The approach is summarized in the following steps. Finally, some experiments on natural images show the robustness of our method.

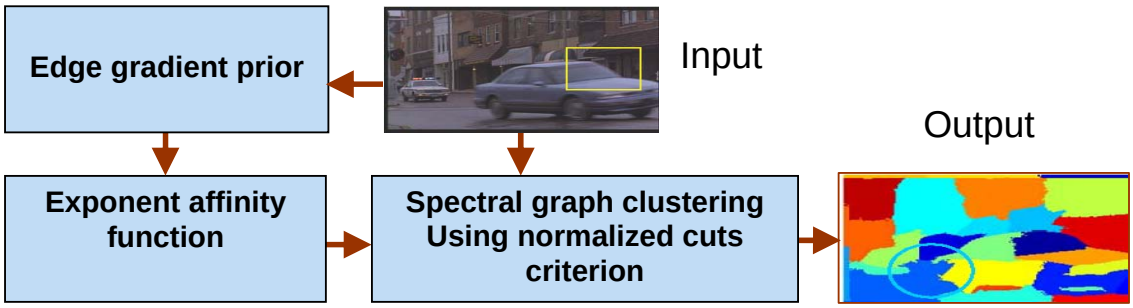


Figure 6.12: Diagram of segmentation using edge gradient prior and the normalized cut criterion. Edge gradient prior in spectral clustering using normalized cut criterion for image segmentation.

6.5.1 Segmentation Using Different Affinity Functions

Partitioning Using Exponent Affinity Function

We extend two different affinity functions [226], [282], [142] in this discrete regularization approach and compare the results. We measure the degree of dissimilarity between pairwise blurred and unblurred regions based on the special characteristic properties, i.e., stronger difference of edge gradients, pairwise blur and unblur. The edge weight w_{ij} between node i and j as the product of a feature similarity term and spatial proximity term:

$$w_{(i,j)} = exp\frac{-\|Q(i)-Q(j)\|_2^2}{\sigma_I} * exp\frac{-\|X(i)-X(j)\|_2^2}{\sigma_X} \tag{6.31}$$

if $(\|X(i) - X(j)\|_2) < r$, and $w_{ij} = 0$, otherwise. $Q(i) = \nabla G_\sigma * I(x)$ is the edge gradients and large differences between pairwise blurred and unblurred regions. G_σ is a Gaussian filter, I(x) is an input image, $X(i)$ is the spatial location of node i. The partitioning of blurred and unblurred regions or objects is guided dynamically by computed edge gradient prior values without any supervision. For some natural images like Fig. 6.13, we can get accurate segmentation results, while the number of segmentation is given manually. However, for some complex partially-blurred and cluttered images, shown in Fig. 6.14 and Fig. 6.15, we can easily find there are some small errors on some regions of segmentation, e.g., the cutting edges, and some regions are misclassified. The reason is that the σ_I and σ_X are global constant. The affinity is large and rough so that those nearby pixels with relatively similar intensity values are misclassified.

Partitioning Using Window-Based Affinity Function

Firstly, using the affinity weight can be improved by given more descriptive prior to guide the cut edges at the first step. Secondly, the affinity weight can be small and window-based so that the affinity can reasonably and accurately represent pixels in small windows. Here, we extend a window-based weight function from [142] into discrete regularization for measuring the affinity.

$$w_{(i,j)} = \sum_{k|(i,j)\in w_k} \frac{1}{|w_k|}\left(1 + \frac{(I_i - \delta_k)(I_j - \delta_k)}{\sigma_k^2 + \frac{\varepsilon}{|w_k|}}\right) \tag{6.32}$$

where δ_k and σ^2 are mean and variance of the intensities in the window w_k around k, and $|w_k|$ is the number of pixels in this window. The main difference of this affinity function is that it uses

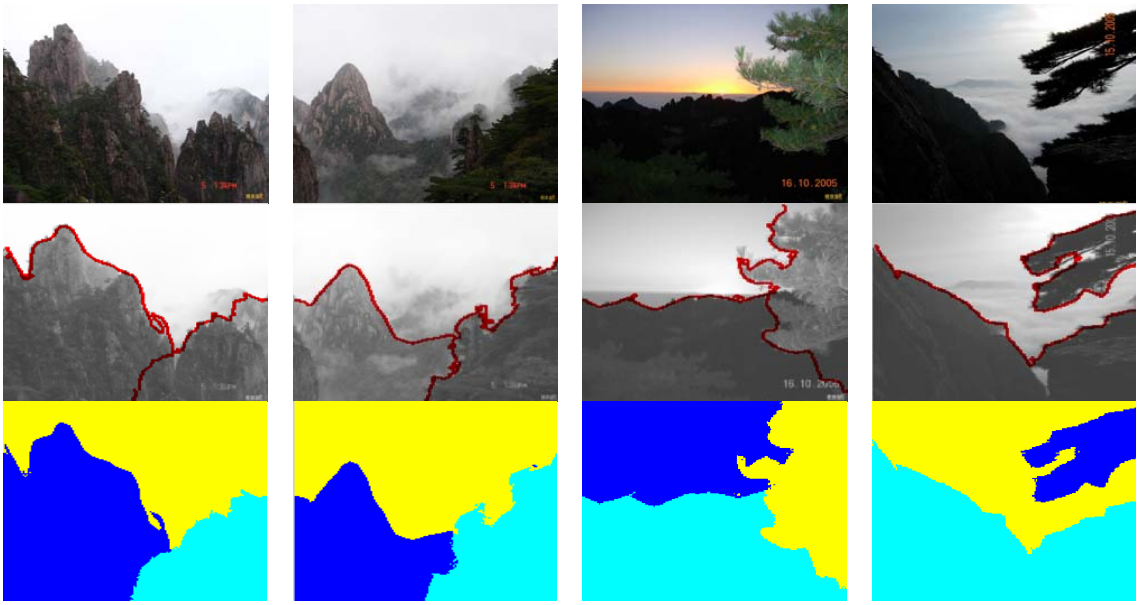


Figure 6.13: $a|b|c|d$. Performance of the spectral clustering using normalized-cuts criterion on natural scene images. Clustering number is manually defined. The first row is original images. The second row is marked clustering regions. The third row is color marked regions.

local estimates of means of variances instead of the global derivation σ_I and σ_X. The affinity between two pixels of the intensity decreases with distance, while the affinity between pixels of different intensity is zero. Neighbor pixels with similar intensity have high affinity, otherwise, the affinity is small.

As we know, a meaningful segmentation needs to be integrated with a specific task, i.e., integration of top-down and bottom-up processing. The flexibility of the smoothing term in the discrete regularization allows to integrate edge-based prior, and patch-based descriptive prior or constraints. Simultaneously, different from most image matting and segmentation methods, pairwise differences between blurred and unblurred regions support many types of unsupervised descriptive and generative priors to guide the Laplacian "cut edges". Föstner [79] feature operator (except Lowe's SIFT descriptor, it is based on different principle.) finds most intensity cross corners as prior labeling, shown in Fig. 6.17 (b) and Fig. 6.18 (a). The Laplacian partitions tends to be piecewise constant in the same region where the smallest eigenvectors are piecewise constant. If the values inside a partition in the eigenvector image are coherent, a simple seeds or patch labeling within such a partition is sufficient to attenuate the difference and find the right cut edges to the entire segment.

In the experiment, we note that these sparse feature corners and edge gradient prior are sufficient to segment the blurred and unblurred regions, shown in Fig. 6.17 (b) and Fig. 6.18 (b)(c). Fig. 6.17 shows the segmentation of gray value partially-blurred images. The video frames are captured from films or video data. The unsupervised labeling is controlled using Perona-Malik image diffusion filter shown in Fig. 6.17 (b). The blurred region and unblurred foreground car are segmented in different layer, shown in Fig. 6.17 (c) and (d). The color images are separated into RGB colour channels and each channel is processed accordingly. However, there are still some parts that are not well segmented in that the intensity of gray values are too similar for a small sized window, e.g., 3×3. Therefore, by an optimal control of the weights, unsupervised labeling can achieve high-quality segmentation results.

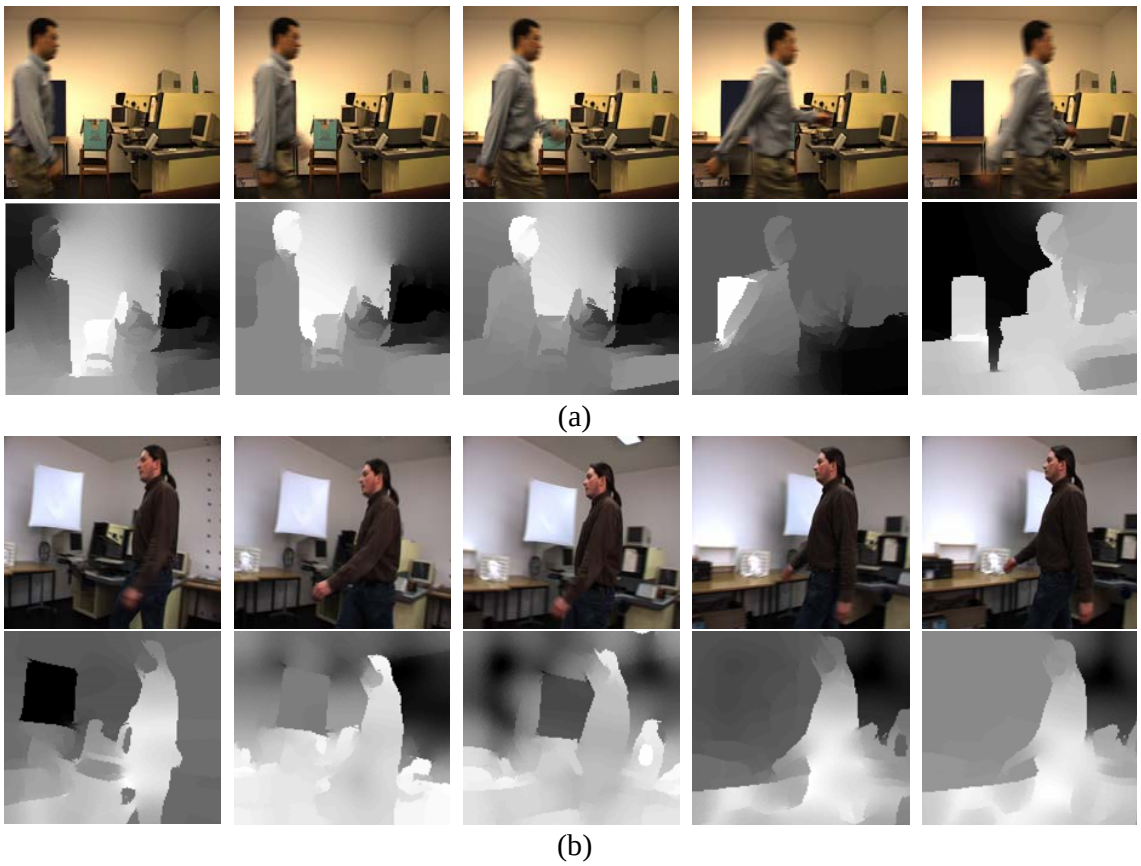


Figure 6.14: (a) Identified blurred foreground walking man in front of unblurred background. (b) Identified unblurred foreground walking man in front of blurred background. However, some small blurred regions are misclassified.

6.5.2 Restoration on Entirely Nonstationary Blurred Images

To evaluate this algorithm, the performance of the approach has been investigated by using different types of real images. In these experiments, first, we show that it is easy to get ringing effects in normal deblurring methods. Second, we reconstruct several types of blurred images and compare the results with other methods. Finally, we make a summarization for the suggested approach.

The first experiment is performed for an indoor image, shown in Fig. 6.19. Based on the estimated blur kernel in Fig. 6.21(a), we reconstruct this image using two methods. We can easily find that the classical Richardson-Lucy (RL) method can achieve sharp deblurring results but suffering stronger ringing effects. Fig. 6.19(c) is reconstructed using our suggested method with natural image statistical prior weights and space-adaptive smoothing. Compared to the RL method, the reconstructed result in our method is smoother and without ringing effects.

The second experiments present image restoration on blurred images to demonstrate the deblurring and restoration results of the proposed algorithm. The restored images are illustrated in Fig. 6.20 and their identified blur kernels are shown in Fig. 6.21, respectively. In this experiment, we compare our deblurring with a multi-scale based RL methods that was used by Fergus et al. [73]. From the results, we note that the multi-scale RL method can achieve sharp restoration results but the noise is also amplified, shown in Fig. 6.20(b) column. Our method can achieve the

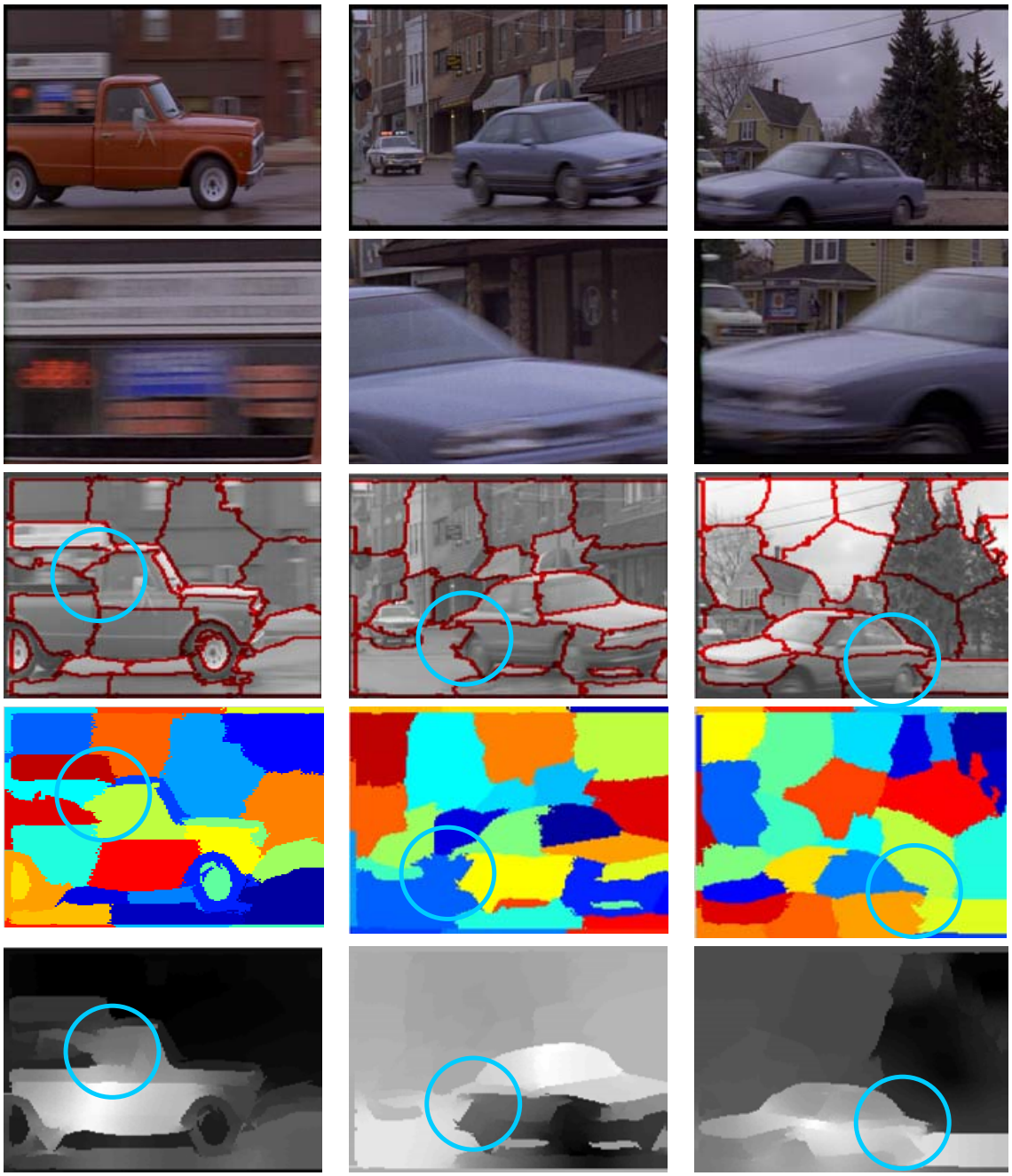


Figure 6.15: $a|b|c$. (a)(b)(c) are three partially-blurred images. The 3rd row and 4th show the cut edges with some errors in blue circles. Some parts are misclassified in that the affinity weight is too large an rough to represent these small corner regions.

sharp deblurring results with more smoothing surfaces due to different reconstruct mechanism, shown in Fig. 6.20(c). In this experiment, we show three blurred images with different illumination, contrast and environments. The first image is an indoor image of a person, the second image of a copper sculpture has some reflections, and the third one has cluttered movements in the evening. The results show the robustness of image deblurring and reconstruction of the suggested approach for different types of nonstationary real blurred images.

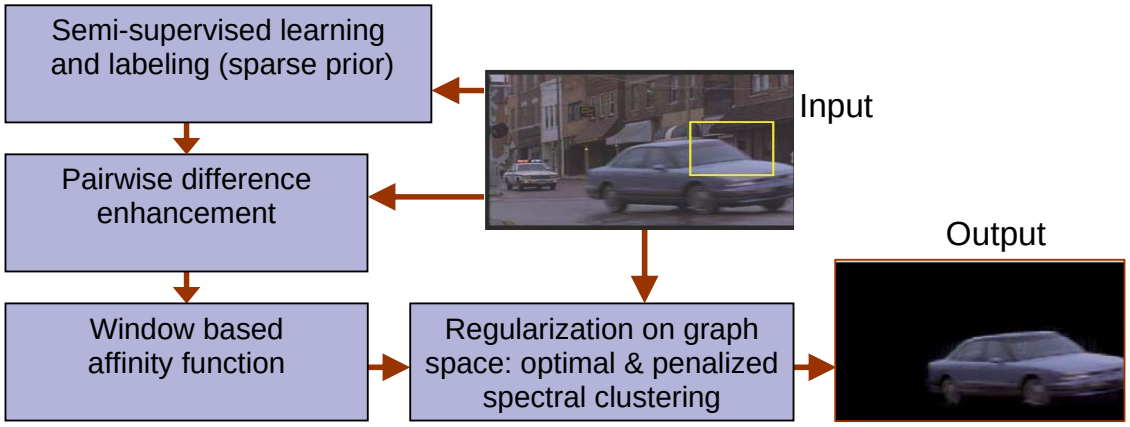


Figure 6.16: Diagram of regularization on discrete graph space for segmentation via semi-supervised learning and optimal control.

Figure 6.17: $\frac{a|b}{c|d}$ The performance of segmentation for partially-blurred image in pure gray value. (a) Original video data. unblurred foreground car with blurred background. (b) Feature detection after Perona-Malik nonlinear image filtering, K = 10. (c) Segmented and identified blurred background in gray value. (d) Segmented and identified unblurred car.

From these experiments, we note that these estimated blur kernels cannot be simply represented by some parametric models. The reason is that the random movements and different noise influences (illumination, projective distortion based blur changing, reflections etc.) during the image formation period. In a sense, based on natural image statistics, we can estimate blur kernels in

Figure 6.18: $a|b|c$. Partition and identification of partially-blurred images. (a) Detected feature corners (Föstner operator [79]) as unsupervised prior labeling. These labeling corners can be high-level seeds indicating regions of the image belongs to one regions or object. (b) Segmented unblurred regions or objects. (c) Segmented and identified blurry regions or objects.

Figure 6.19: $a|b|c$. (a) Blur degraded images. (b) Restored image using the normal RL method with ringing effects. (c) Restored images using the suggested method.

a variational Bayesian learning method and restore and reconstruct the images sequentially in an iterative reweighted energy function. On the other hand, we also note that image noise is smoothed during image deconvolution in our approach.

Figure 6.20: $a|b|c$ columns. (a) Blur degraded images. (b) Restored image using a similar method in [73]: multi-scale based RL method. (c) Restored images using the suggested method with natural image statistical prior weights and space-adaptive smoothing.

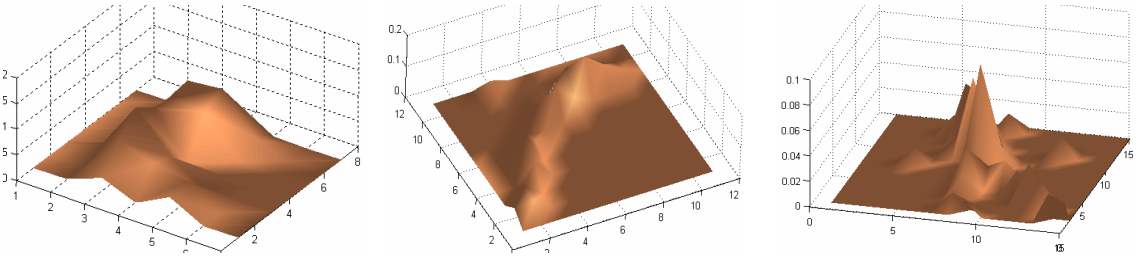


Figure 6.21: $a|b|c$. Identified blur kernels with respect to the image of people, street and horse, respectively. (a) for people. (b) for horse. (c) for street.

6.5.3 Discussion of Image Priors and Probability Models

This method is robust regarding different types of noises and blur, because it searches the differences of regions or objects based a global optimization and natural related descriptive prior

information. These prior information can be as seeds growing, merging to find the "biggest" similarity affinity matrix. The descriptive information here means that more representative eigenvectors can achieve better segmentation results. The search for eigenvectors is based on the guidance of descriptive prior information. Furthermore, these descriptive information can be simply based on image contrast, feature descriptors [150], [207], gray scale, gradient, or can be computed through statistical learning.

Identification and segmentation of blurred and unblurred regions or objects in partially-blurred images is a relatively specific task. We utilize the potential differences between blurred and unblurred regions to find the descriptive prior information, so that we can achieve high-quality segmentation results in a discrete regularization approach via unsupervised learning and optimal control. Although these two different types of prior information are computed from different view points, the underlying principle is still to find the descriptive information for the guidance of clustering and merging. Therefore, a number of difficulties of identification and segmentation can be solved and avoided via the descriptive prior information which is also a penalty term in the discrete regularization. This method can be applied in general segmentation cases. In such cases, different descriptive information and its related labeling may achieve different task-driven results. Another underlying idea in the suggested discrete regularization is to make progress on scale problems and find some invariant information for segmentation.

Recently, Zhu [302] classifies the probability model into *descriptive* and *generative* model based on the study of natural image statistics, the analysis of natural image components, the grouping of natural image elements, and the modeling of visual patterns. The descriptive model is constructed based on statistical descriptions of the image ensembles via natural image statistics and natural image patterns. Thus, the descriptive model is attractive in that a single descriptive model can integrate all statistical measures of different image features. For example, Gibbs model of texture, Gibbs model of shapes (2D simple curves).

Different from the projection of pursuit method (a simple product of the likelihoods or marginals on different features), the descriptive model uses sophisticated energy functions to account for the dependency of these features. Also, the descriptive model are all built on certain lattices-graph structures in homogeneous models (statistics are assumed to be same for all vertices of the graph) and inhomogeneous model (the vertices of the graph are labeled, and different statistics of features are used for different regions or sites). To tackle the computational complexity of descriptive models, generative models are used to reduce dimensions based on some vocabulary of visual descriptions. The elements in the vocabulary specify how images are generated from hidden variables.

In the Bayesian framework, descriptive and generative models are used as prior probabilities and likelihoods, while discriminative models approximate the posterior probabilities of hidden variables based on local features. The underlying of our approach belongs in a Bayesian estimation based descriptive model for classifying and identifying blurred and unblurred regions for partially-blurred images.

Firstly, our approach can be considered as a combination of these two models, i.e., the interaction between unsupervised learning and labeling (generative models) and partially-blurred image segmentation and identification. Secondly, the probability models of stationary blurred images and non-stationary blurred images can be interpreted as homogeneous and inhomogeneous models. It leads to the choice of probability models and prior knowledge. Zhu and Mumford [303], [304] call the generic prior models *the first-level prior*. A more sophisticated prior model should

incorporate concepts like object geometry; such prior models are called the *second-level prior*. For example, diffusion functions derived from this second-level prior are studied in image segmentation [306], and in scale space of shapes [131]. In our approach, diffusion functions are used to adjust the semi-supervised labeling and segmentation.

6.5.4 Noise Robustness

The theoretical results describe how the expected segmentation should behave in response to different weights of i.i.d. random noise. The behavior of the partition considers the segmentation globally, even in a small window. From these experiments on synthetic and natural images, we can find that different amounts of noise and different levels of natural noise do not influence image segmentation. It is also one of the most distinguished advantages of spectral clustering methods in comparison to other state-of-the-art segmentation methods. Window-based weight affinity function can achieve more accurate segmentation results in natural partially-blurred images.

6.6 Conclusions

We have proposed a unifying discrete regularization approach to achieve high quality partially-blurred image partition, identification and blind restoration. This approach integrates and shares the advantages from both spectral graph theory and regularization theory for solving ill-posed inverse problems. Different from existing off-line and supervised labeling methods, this approach allows on-line unsupervised learning and labeling so that we can achieve a meaningful task-driven segmentation. Perceptual blind image restoration can be achieved in a different identified layer via optimal scale control. This approach has robust performance on different types of partially-blurred natural images. The integrated approach also demonstrates that the mutual support between natural prior knowledge and low-level image processing has great potential to improve the results in early vision.

we have introduced a new approach to enable variational Bayesian ensemble learning for restoring real nonstationary blurred images. The experiments suggest that the approach enables an efficient trade-off between intractable inferences and tractable solution for difficult inverse problems. In particular, the approach makes effective use of the natural image statistics prior through the learning scheme. By alternating the radius of the natural image statistics, we are able change the approach to restore more types of blurred images including nonuniform blurred images. A thorough evaluation has shown that the proposed approach has more flexibilities for identifying and restoring blurred images in real environments. The new approach outperforms state-of-the art methods on challenging real blurred data, underlining the effectiveness of the approach.

7 Summary and Future Work

Wir müssen wissen, Wir werden wissen. - David Hilbert

7.1 Summary

In this thesis, we have proposed three main approaches for blur identification, image restoration and partially-blurred identification, segmentation and restoration in an integrated Bayesian estimation and regularization framework based on continuous Hilbert, bounded variation spaces and discrete graph spaces.

The observation and experiments have guided us to seek alternative approaches of scoring candidate statistical learning based optimization methods. Furthermore, convex optimization criteria are employed to achieve existing, unique, stable solution for blind image reconstruction or restoration and segmentation. These approaches are an integration of statistics and regularization, and transductive inference of regularization on discrete graph spaces. The soundness of these approaches is demonstrated by numerical experiments. The main contributions of this thesis to the computer vision, image processing and pattern recognition community are summarized in the following.

The first part of our work focuses on the strategy of global nonparametric estimation to local parametric optimization for high-accuracy blur identification. The nonparametric estimation used to adapt adaptation of the parametric methods to the data when the parametric structural assumption is not fulfilled. This approach is based on statistical learning priors and deterministic regularization for blur identification and image restoration for stationary-blurred images. The proposed double regularized Bayesian estimation is strictly convex so that the approach can achieve the global convergence. The accurate initial value can also speedup the convergence for the estimation of point spread function and image restoration. An early work of this approach has been published in [295], [289]. Simultaneously, a family of variational functionals like Mumford-Shah, and total variation have been investigated and implemented for image restoration and segmentation, published in [294], [293], [291] and natural image statistics and variational Bayesian learning in [287], [298].

The second part of this work focuses on high-fidelity and perceptual image restoration. Variational regularization in the BV space has been extended in a Bayesian framework to achieve simultaneously blur identification and image restoration. Based on a family of general and more general linear-growth functional in the BV space, we propose a Bayesian based double variational blind image restoration functional which can be optimal controlled via self-adjusting diffusion operators, self-adaptive regularization parameters, and the optimal time of stopping the process. The underlying mathematic principles and practical roles are embodied in an energy optimization approach. Related works have been published in [294], [290], [292], [296] and the submitted journal paper [297].

The third part of this work focuses on partially-blurred image segmentation, identification and restoration [292],[299]. The original idea of this work is to find underlying mathematic relationship between regularization theory and spectral graph theory, since these two theories can be individually used for image segmentation based on the same strategy of global optimization. Moreover, both theory based approaches can use Laplacian for controlling and smoothing the convergence results. The proposed discrete regularization approach integrates spectral graph theory and regularization theory in graph spaces based on the underlying mathematic connections and generalization. First, this novel approach can efficiently utilize, covert, and store the high-level knowledge to guide low-level image segmentation and restoration. Second, the segmentation is also optimized by modifying weight affinity function. Furthermore, the flexibility and generality make this approach easily extendable to solve many related image processing and vision tasks. Several related papers have been submitted recently.

In this thesis, we have introduced several new approaches for low-level vision problems based on an integrated statistical learning, Bayesian estimation and regularization framework. These experiments suggest that our strategy and our suggested approaches enables efficient and tractable solutions for difficult inverse problems. In particular, these approaches makes effective use of the natural image statistics based generative and discriminative prior information through their related learning scheme. By alternating the radius of the image statistics and learning in Bayesian estimation, we are able change our approaches to solve other inverse problems in pattern recognition and computer vision. These approaches outperforms state-of-the art methods on challenging real vision problems, underlining the effectiveness of our strategy and these introduced approaches.

7.2 Future Work

7.2.1 Theoretical Aspects

The target of this research can be formulated as the systemic conception of feasible unsupervised signal and image restoration, and segmentation methods. These methods make use of all available information, i.e., with respect to the data acquisition, perturbation models and natural generative priors which can largely improve the performance. We feel there are several directions that can be further and fruitfully explored at the next step in theoretical and practical aspects. Also, the underlying mathematic concept is widely open for future research and arouses a lot of new questions.

1. Minimization of various cost-functions. The optimization is important for several reasons. Primarily, it allows obtaining rigorous solutions with respect to parameter selection and optimization techniques, and an opportunity to find solutions in a robust convex manner.

2. Construction of a variational statistical framework combining discrete regularization. An important extension of the analytical results regarding the properties of optimization is the integration into a statistical estimation framework. Statistical description of these properties as a function of the randomness of the data will be further explored.

3. The computational aspects are crucial for the success of a signal and image restoration method. The properties of convex optimization schemes lead us to reduce the search space

and thus simplify the optimization. Also, the integration of knowledge in form of partial differential equations into convex optimization can be further explored for achieving more robust automatic visual perception results.

4. The regularization is integrated in discrete graph spaces through spectral graph theory. Discrete regularization operators are simultaneously used to smooth the image and smooth the spectral clustering information. The encouraging results inspire us to further explore the potential properties of this unified discrete regularization approach.

5. The utilization of bottom-up and top-down segmentation and recognition strategy. We can combine low level image processing and mid or high level knowledge for repartitioning or grouping blurred and unblurred objects or regions.

7.2.2 Practical Applications

1. Control of static and dynamic convergence behavior of isotropic and anisotropic regions. The cost function, or average convergence performance of blind identification and image restoration can be classified as static convergence analysis and dynamic convergence analysis of the the stochastic dynamics of equalization algorithms.

2. Nonstationary recursive image estimation. The identification and restoration of non-stationary blur is more useful for real life data. Currently, we have built a robust statistic estimation and convex optimization framework so that we can investigate the deblurring and denoising of non-stationary, partially-blurred and entirely-blurred images.

3. Simultaneous blur identification, image restoration and segmentation for more complicated blurred images. A more general blurred image is partially-blurred, or nonstationary blurred. The proposed approach integrates and shares the advantages from both regularization, spectral graph theory and statistical learning theory. Also, these advantages of this approach can be directly or indirectly applied to the data coming from different sources, e.g., tomograph, synthetic aperture radar or electronic microscope data.

4. Apply the proposed regularization and Bayesian learning framework to other related pattern recognition, computer vision problems. The proposed strategy and approaches are demonstrated to be a more flexible and robust learning and optimization framework then most state-of-the-art methods for inverse problems, low-level vision problems etc.

5. Add new statistical strategies and approaches into current work. For example, statistical approximate inferences including mean field methods, variational methods and free energy are well developed and can be further applied in the research of computer vision.

A Methods Not Requiring Evaluation of Derivatives

All the gradient methods require calculation of at least the gradient $\nabla f(x_k)$ and possibly the Hessian matrix $\nabla^2 f(x_k)$ at each generated point (pixel level) x_k. It is possible to use the same algorithms as earlier with all unavailable derivatives approximated by finite differences. Thus, second derivatives may be approximated by the *forward difference formula*

$$\frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} \sim \frac{1}{h}[\frac{\partial f(x_k + he_j)}{\partial x^i} - \frac{\partial f(x_k)}{\partial x^i}] \tag{A.1}$$

or the *backward difference formula*

$$\frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} \sim \frac{1}{h}[\frac{\partial f(x_k)}{\partial x^i} - \frac{\partial f(x_k - he_j)}{\partial x^i}] \tag{A.2}$$

or the *central difference formula*

$$\frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} \sim \frac{1}{2h}\left[\frac{\partial f(x_k + he_j)}{\partial x^i} - \frac{\partial f(x_k - he_j)}{\partial x^i}\right] \tag{A.3}$$

In these relations, h is a small positive scalar and e_j is the jth unit vector (jth column of the identity matrix). The central difference formula has the disadvantage that it requires twice as much computation as the forward difference formula. However, it is much more accurate. Practical experience suggests that a good policy is to keep the scalar h for each derivative a fixed value which balances the truncation error against the cancelation error. A good practical rule is to use the forward and backward difference formulas until the absolute value of the corresponding approximate derivative becomes less than a certain tolerance; i.e.,

$$|(1/h[f(x_k + he_i) - f(x_k)])| \leq \varepsilon$$

where ε is some small pre-specified scalar. At that point a switch to the central difference formula is made, i.e., whenever the inequality above is satisfied. An extensive discussion of implementation of gradient methods based on finite difference approximations can be found in Gill et al. (1981) [90]. There are several other algorithms for minimizing differential functions without the explicit use of derivatives, the most interesting of which, at least from the theoretical point of view, are coordinate descent methods [152, 285].

We discretize (using a fixed point finite differences scheme) the Euler-Lagrange equation associated with the minimization of the total variation model of Rudin-Osher-Fatemi.

We would like to find the (unique) minimizer, u, of

$$\inf_u F(u) = \lambda \int_\Omega |f-u|^2 \, dxdy + \int_\Omega |\nabla u| \, dxdy, \tag{A.4}$$

where f is the noisy data and $\lambda > 0$ is a scaling parameter. The associated Euler-Lagrange equation of the Rudin-Osher-Fatemi model formally is

$$\begin{cases} u = f + \frac{1}{2\lambda}\mathrm{div}(\frac{\nabla u}{|\nabla u|}) \text{ in } \Omega, \\ \frac{\partial u}{\partial \vec{n}} = 0 \text{ on } \partial\Omega. \end{cases} \tag{A.5}$$

First, we remove the singularity when $|\nabla u| = 0$, by approximating $F(u)$ by $F_\varepsilon(u)$, where

$$F_\varepsilon(u) = \lambda \int_\Omega |f-u|^2 \, dxdy + \int_\Omega \sqrt{\varepsilon^2 + |\nabla u|^2} dxdy,$$

with $\varepsilon > 0$ being a small constant. Then, the Euler-Lagrange equation minimizing $F_\varepsilon(u)$ formally is:

$$u = f + \frac{1}{2\lambda}\mathrm{div}(\frac{\nabla u}{\sqrt{\varepsilon^2 + |\nabla u|^2}}) \text{ in } \Omega \tag{A.6}$$

$$\frac{\partial u}{\partial \vec{n}} = 0 \text{ on } \partial\Omega. \tag{A.7}$$

Assume for simplicity $\Omega = (0,1)^2$, $h > 0$ and let $x_i = ih$, $y_j = jh$, $h = 1/M$, for $0 \le i, j \le M$, be the discrete points (in our numerical calculations, we have $h = 1$). We recall the following notions:

$u_{i,j} \approx u(x_i, y_j)$,
$f_{i,j} \approx f(x_i, y_j)$,
$\Delta^x_\pm u_{i,j} = \pm(u_{i\pm1,j} - u_{i,j})$,
$\Delta^y_\pm u_{i,j} = \pm(u_{i,j\pm1} - u_{i,j})$,
$\Delta^x_0 u_{i,j} = (u_{i+1,j} - u_{i-1,j})/2$, and
$\Delta^y_0 u_{i,j} = (u_{i,j+1} - u_{i,j-1})/2$.

A discrete form of the Euler-Lagrange equation is:

$$
\begin{aligned}
u_{i,j} = f_{i,j} &+ \frac{1}{2\lambda h}\Delta_-^x\left[\frac{1}{\sqrt{\varepsilon^2+\left(\frac{\Delta_+^x u_{i,j}}{h}\right)^2+\left(\frac{\Delta_0^y u_{i,j}}{h}\right)^2}}\frac{\Delta_+^x u_{i,j}}{h}\right] \\
&+\frac{1}{2\lambda h}\Delta_-^y\left[\frac{1}{\sqrt{\varepsilon^2+\left(\frac{\Delta_0^x u_{i,j}}{h}\right)^2+\left(\frac{\Delta_+^y u_{i,j}}{h}\right)^2}}\frac{\Delta_+^y u_{i,j}}{h}\right] \\
= f_{i,j} &+ \frac{1}{2\lambda h^2}\frac{u_{i+1,j}-u_{i,j}}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j}-u_{i,j}}{h}\right)^2+\left(\frac{u_{i,j+1}-u_{i,j-1}}{2h}\right)^2}} \\
&-\frac{1}{2\lambda h^2}\frac{u_{i,j}-u_{i-1,j}}{\sqrt{\varepsilon^2+\left(\frac{u_{i,j}-u_{i-1,j}}{h}\right)^2+\left(\frac{u_{i-1,j+1}-u_{i-1,j-1}}{2h}\right)^2}} \\
&+\frac{1}{2\lambda h^2}\frac{u_{i,j+1}-u_{i,j}}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j}-u_{i-1,j}}{2h}\right)^2+\left(\frac{u_{i,j+1}-u_{i,j}}{h}\right)^2}} \\
&-\frac{1}{2\lambda h^2}\frac{u_{i,j}-u_{i,j-1}}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j-1}-u_{i-1,j-1}}{2h}\right)^2+\left(\frac{u_{i,j}-u_{i,j-1}}{h}\right)^2}}
\end{aligned}
$$

We use a fixed point Gauss-Seidel iteration method for the above equation and so we now introduce the following linearized equation:

$$
\begin{aligned}
u_{i,j}^{n+1} = f_{i,j} &+ \frac{1}{2\lambda h^2}\frac{u_{i+1,j}^n-u_{i,j}^{n+1}}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j}^n-u_{i,j}^n}{h}\right)^2+\left(\frac{u_{i,j+1}^n-u_{i,j-1}^n}{2h}\right)^2}} \\
&-\frac{1}{2\lambda h^2}\frac{u_{i,j}^{n+1}-u_{i-1,j}^n}{\sqrt{\varepsilon^2+\left(\frac{u_{i,j}^n-u_{i-1,j}^n}{h}\right)^2+\left(\frac{u_{i-1,j+1}^n-u_{i-1,j-1}^n}{2h}\right)^2}} \\
&+\frac{1}{2\lambda h^2}\frac{u_{i,j+1}^n-u_{i,j}^{n+1}}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j}^n-u_{i-1,j}^n}{2h}\right)^2+\left(\frac{u_{i,j+1}^n-u_{i,j}^n}{h}\right)^2}} \\
&-\frac{1}{2\lambda h^2}\frac{u_{i,j}^{n+1}-u_{i,j-1}^n}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j-1}^n-u_{i-1,j-1}^n}{2h}\right)^2+\left(\frac{u_{i,j}^n-u_{i,j-1}^n}{h}\right)^2}}
\end{aligned}
$$

Introducing the notations:

$$
\begin{aligned}
c_1 &= \frac{1}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j}^n-u_{i,j}^n}{h}\right)^2+\left(\frac{u_{i,j+1}^n-u_{i,j-1}^n}{2h}\right)^2}}, \\
c_2 &= \frac{1}{\sqrt{\varepsilon^2+\left(\frac{u_{i,j}^n-u_{i-1,j}^n}{h}\right)^2+\left(\frac{u_{i-1,j+1}^n-u_{i-1,j-1}^n}{2h}\right)^2}} \\
c_3 &= \frac{1}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j}^n-u_{i-1,j}^n}{2h}\right)^2+\left(\frac{u_{i,j+1}^n-u_{i,j}^n}{h}\right)^2}}, \\
c_4 &= \frac{1}{\sqrt{\varepsilon^2+\left(\frac{u_{i+1,j-1}^n-u_{i-1,j-1}^n}{2h}\right)^2+\left(\frac{u_{i,j}^n-u_{i,j-1}^n}{h}\right)^2}}
\end{aligned}
$$

and solving for $u_{i,j}^n + 1$, we obtain:

$$u_{i,j}^{n+1} = \left(\frac{1}{1+\frac{1}{2\lambda h^2}(c_1+c_2+c_3+c_4)} \right) \cdot \left[f_{i,j} + \frac{1}{2\lambda h^2}(c_1 u_{i+1,j}^n + c_2 u_{i-1,j}^n + c_3 u_{i,j+1}^n + c_4 u_{i,j-1}^n) \right]$$

We let $u_{i,j}^0 = f_{i,j}$. Then, we note that if $m_1 \leq f_{i,j} \leq m_2$, for any $0 \leq i,\, j \leq M$, we have $m_1 \leq u_{i,j}^n \leq m_2$, for any $n \geq 0$. We use the above equation for $u_{i,j}^{n+1}$ for all interior points (x_i, y_i) such that $1 \leq i,\, j \leq M-1$.

The boundary condition can be implemented in the following way: if $u_{i,j}^n$ has been computed using the above numerical scheme for $1 \leq i,\, j \leq M-1$, then we let $u_{0,j}^n = u_{1,j}^n$, $u_{M,j}^n = u_{M-1,j}^n$, $u_{i,0}^n = u_{i,1}^n$, $u_{i,M}^n = u_{i,M-1}^n$, and $u_{0,0}^n = u_{1,1}^n$, $u_{0,M}^n = u_{1,M-1}^n$, $u_{M,0}^n = u_{M-1,1}^n$, $u_{M,M}^n = u_{M-1,M-1}^n$.

- The coefficient λ has to be optimized for each image. Too small λ will introduce too much smoothing in the recovered image u. However, too large λ will keep noise in the solution u.
- Note that this scheme may introduce some asymmetry, but not visible in general. Other schemes can be proposed, for instance alternating at each iteration the discretization of the div operator, with all four (schematic) choices

$$\begin{array}{l} \Delta_-^x(\Delta_+^x), \Delta_-^y(\Delta_+^y) \\ \Delta_+^x(\Delta_-^x), \Delta_+^y(\Delta_-^y) \\ \Delta_+^x(\Delta_-^x), \Delta_-^y(\Delta_+^y) \\ \Delta_-^x(\Delta_+^x), \Delta_+^y(\Delta_-^y) \end{array}$$

Upwind Differences

Once f and $\overrightarrow{V}$ are defined at every grid point on the image pixel grid, we can apply numerical methods to evolve f forward in time moving the diffusion across the grid. Updating f in time consists of finding new values of f at every grid point after some time increment Δt. We denote these new values of f by $f^{n+1} = f(t^{n+1})$, where $t^{n+1} = t^n + \Delta t$.

The first-order accurate methods for the time discritization of $f_t + \nabla f \cdot \overrightarrow{V} = 0$ is the forward Euler method given by

$$\frac{f^{n+1} - f^n}{\Delta t} + \nabla f^n \cdot \overrightarrow{V}^n = 0, \tag{A.8}$$

where $\overrightarrow{V}^n$ is the given external velocity field at time t^n, and ∇f^n evaluates the gradient operator using the values of f at time t^n.

Naively, one might evalute the spatial derivative of f in a straightforward manner using equation: first-order accurate forward difference $\frac{\partial f}{\partial x} \approx \frac{f_{i+1} - fi}{\Delta x}$ abbreviated as $D^+ f$, first-order accurate backward difference $\frac{\partial f}{\partial x} \approx \frac{f_i - fi - i}{\Delta x}$ abbreviated as $D^- f$, or a second-order accurate central

difference $\frac{\partial f}{\partial x} \approx \frac{f_{i+1}-f_{i-1}}{a\Delta x}$ abbreviated as $D^0 f$. However, this straightforward manner approach will fail.

One generally needs to exercise great care when numerically discretizing partial differential equations. The Eq. A.8 in expanded form is,

$$\frac{f^{n+1} - f^n}{\Delta t} + u^n f_x^n + v^n f_y^n + w^n f_z^n = 0, \tag{A.9}$$

and address the evaluation of the $u^n f_x^n$ term first. The techniques can be applied in a dimension-by-dimension manner.

For simplicity, consider the one-dimensional version of Eq. A.9,

$$\frac{f^{n+1} - f^n}{\Delta t} + u^n f_x^n = 0, \tag{A.10}$$

when the sign of u^n indicates whether the values of f are moving to the right or to the left. Since u^n can be spatially varying, we focus on a specific gird point x_i, when we write,

$$\frac{f^{n+1} - f^n}{\Delta t} + u_i^n (f_x)_i^n = 0, \tag{A.11}$$

when $(f_x)_i$ denotes the spatial derivative of f at the point x_i. If $u_i > 0$, the values of f are moving from left to right, and the method of characteristics tells us to look to the left of x_i to determine what value to f will land on the point x_i at the end of a time step. Similarly, if $u_i < 0$, the values of f are moving from right to left, and the method of characteristics implied that we should look to the right to determine an appropriate value of f_i at time t^{n+1}. Clearly, $D^- f$ should be used to approximate f_x when $u_i > 0$. In contrast, $D^+ f$ can not possibly give a good approximation, since it fails to contain the information to the left of x_i that dictates the new value of f_i. Similar reasoning indicates that $D^+ f$ should be used as approximate f_x when $u_i < 0$. This method of choosing an approximation to the spatial derivatives based on the sign of u is known as upwind differencing or upwinding. Generally, upwind methods approximate derivatives by biasing the finite difference stencil in the direction where the characteristic information is coming from.

The upwind discretization is summarized as follows.

1. At each grid point, define f_x^- as $D^- f$ and f_x^+ as $D^+ f$.
2. If $u_i > 0$, approximate f_x with f_x^-. If $u_i < 0$, approximate f_x with f_x^+.
3. When $u_i = 0$, the $u_i (f_x)_i$ term vanishes, and f_x does not need to be approximated.

This is a first-order accurate discretization of the spatial operator, since $D^- f$ and $D^+ f$ are first-order accurate approximations of the derivatives; i.e., the error are $O(\Delta x)$.

The combination of the forward Euler time discretization with the upwind difference scheme is a constant finite difference approximation to the partial differential equation, since the approximation error converges to zero as $\Delta t \to 0$ and $\Delta x \to 0$. According to the Lax-Richtmeyer equivalence theorem a finite difference approximation to a linear partial differential equation is convergent, i.e., the correct solution is obtained as $\Delta t \to 0$ and $\Delta t \to 0$, if and only if it is both consistent and stable. Stability guarantees that small errors in the approximation are not amplified as the solution is marched forward in time.

B Proof of Data-Driven Image Diffusion Functional

Given a blurred noise image formation model $I = u + noise$, we get an energy optimization functional.

$$\mathcal{J}(\hat{f}_{(g)}) = \frac{\alpha_1}{2}\int_\Omega (u-I)^2 dA + \underset{f\in BV\cap L^2(\Omega)}{\lambda \min} \int_\Omega \phi(x, Du)dA \tag{B.1}$$

where $u \in BV(\Omega)$, $Du = \nabla u.L^n + D^s u$ is a Radon measure, ∇u is the density of the absolutely continuous part of Du with respect to the n-dimensional Lebesgue measure, L^n and $D^s u$ is the singular part referred to [69].

A general convex linear-growth functional $\phi = \phi(Du)$ is proposed by [107]. A more general functional has a variable exponent and $\phi = \phi(x, Du)$ [48]. More related work on linear growth functionals and their flows are refer to [21], and alternate variational approach [39] reduces staircasing by minimizing second order functionals.

Definition. For $\Omega \subset^n$, define

$$\int_\Omega \phi\left(x, Dv\right) := \int_\Omega \phi\left(x, \nabla v\right)dx + \int_\Omega |D^s v| \tag{B.2}$$

where ϕ is defined in the following,

$$\phi(x,r)dA = \begin{cases} \frac{1}{q(x)}|r|^{q(x)}, & |\nabla \hat{f}| < \beta \\ |r| - \frac{\beta q(x) - \beta^{q(x)}}{q(x)}, & |\nabla \hat{f}| \geq \beta \end{cases} \tag{B.3}$$

Furthermore, denote

$$\Phi_\lambda\left(v\right) := \int_\Omega \phi\left(x, Dv\right) + \frac{\lambda}{2}\int_\Omega |v-I|^2 dx \tag{B.4}$$

$$\Phi_g\left(v\right) := \int_\Omega \phi\left(x, Dv\right) + \int_{\partial\Omega} |v-g|\, d\mathrm{H}^{n-1} \tag{B.5}$$

and

$$\Phi_{\lambda,g}\left(v\right) := \int_\Omega \phi\left(x, Dv\right) + \frac{\lambda}{2}\int_\Omega |v-I|^2 dx + \int_{\partial\Omega} |v-g|\, d\mathrm{H}^{n-1} \tag{B.6}$$

Remark. For *simplicity*, *we assume* that the *threshold* $\beta = 1$ in (1.4) for all of our *theoretical results*. We can establish lower semi-continuity of the functional Φ_g [34]. The proof is presented in the following,

LEMMA. *Using the notation in definition,*

$$\Phi_g(u) = \tilde{\Phi}_g(v) \tag{B.7}$$

for all $u \in BV(\Omega)$ *where furthermore,* $\Phi_g(u)$ *is lower semi-continuous on* $L^1(\Omega)$; *that is, if* u_j, $u \in BV(\Omega)$ *satisfy* $u_j \to u$ in $L^1(\Omega)$ as $j \to \infty$, *then* $\Phi_g(u) \leq \liminf_{j\to\infty} \Phi_g(u_j)$.

Proof. For each $\psi \in C^1(\Omega, \mathbb{R})$, $u \to \int_\Omega -u\mathrm{div}\psi - \frac{q(x)-1}{q(x)} |\psi|^{\frac{q(x)}{q(x)-1}} dx + \int_{\partial\Omega} \psi \cdot n g d\mathcal{H}^{n-1}$ is continuous and affine on $L^1(\Omega)$. Therefore, $\tilde{\Phi}_g(u)$ is convex and lower semi-continuous on $\tilde{\Phi}_g(u)$ and the domain of $\tilde{\Phi}_g(u)$, $\tilde{\Phi}_g(u)$, is precisely $BV(\Omega)$.

We now show that $\Phi_g(u) = \tilde{\Phi}_g(u)$. For $u \in BV(\Omega)$, we have that for each $\psi \in C^1\left(\overline{\Omega}, \mathbb{R}^n\right)$,

$$-\int_\Omega u\mathrm{div}\psi dx = \int_\Omega \nabla u.\psi dx + \int_\Omega D^S u.\psi - \int_{\partial\Omega} u\psi.\mathrm{ngdH}^{n-1}$$

Therefore, since the measures dx, $D^s u$, and $d\mathcal{H}^{n-1}$ are mutually singular, standard arguments show that

$$\int_\Omega \nabla u \cdot \psi - \frac{\mathrm{q(x)} - 1}{\mathrm{q(x)}} |\psi|^{\frac{\mathrm{q(x)}}{\mathrm{q(x)}-1}} dx + \int_\Omega \left|D^S u\right| + \int_{\partial\Omega} |u - g| \, d\mathrm{H}^{n-1}$$

The proof is then complete once we establish that

$$\int \phi(x, \nabla u)\, dx = \sup_{\psi \in C^1\left(\overline{\Omega}, \mathbb{R}^n\right)\ |\psi| \leqslant 1} \int \nabla u.\psi - \frac{\mathrm{q(x)} - 1}{\mathrm{q(x)}} |\psi|^{\frac{\mathrm{q(x)}}{\mathrm{q(x)}-1}} dx \tag{B.8}$$

Since any $\rho \in L^\infty(\Omega, \mathbb{R}^n)$ can be approximated in measure by $\psi \in C^1\left(\overline{\Omega}, \mathbb{R}^n\right)$, we have that

$$\begin{aligned} & \sup_{\psi \in C^1\left(\overline{\Omega}, \mathbb{R}^n\right)\ |\psi| \leqslant 1} \int_\Omega \nabla u.\psi - \frac{\mathrm{q(x)} - 1}{\mathrm{q(x)}} |\psi|^{\frac{\mathrm{q(x)}}{\mathrm{q(x)}-1}} dx \\ = & \sup_{\rho \in L^\infty\left(\overline{\Omega}, \mathbb{R}^n\right)\ |\rho| \leqslant 1} \int_\Omega \nabla u.\rho - \frac{\mathrm{q(x)} - 1}{\mathrm{q(x)}} |\rho|^{\frac{\mathrm{q(x)}}{\mathrm{q(x)}-1}} dx \end{aligned} \tag{B.9}$$

Choosing $\rho(x) = 1_{\{|\nabla u| \leq 1\}} |\nabla u|^{q(x)-1} \frac{\nabla u}{|\nabla u|} + 1_{\{|\nabla u| > 1\}} \frac{\nabla u}{|\nabla u|}$, where 1_E is the indicator function on E, we see that the right hand side of Eq. B.10 is

$$\geq \int_\Omega \frac{1}{q(x)} |\nabla u|^{q(x)} 1_{\{|\nabla u| > 1\}} + \left[|\nabla u| - \frac{q(x) - 1}{q(x)}\right] 1_{\{|\nabla u| > 1\}} dx = \int_\Omega \phi(x, \nabla v)\, dx \tag{B.10}$$

To show equality in Eq. B.8, we proceed as follows. For any $\rho \in L^\infty\left(\overline{\Omega}, ^n\right)$, since $q(x) > 1$ we have that for almost all x, $\nabla u(x) \cdot \rho(x) \leq \frac{1}{q(x)} |\nabla u|^{q(x)} + \frac{q(x)-1}{q(x)} |\rho(x)|^{\frac{q(x)}{q(x)-1}} \leq \frac{1}{q(x)} |\nabla u|^{q(x)}$.

In particular, if $|\nabla u| \leq 1$,

$$\nabla u\left(x\right) \cdot \rho\left(x\right) - \frac{q\left(x\right)-1}{q\left(x\right)} \left|\rho\left(x\right)\right|^{\frac{q(x)}{q(x)-1}} \leq \frac{1}{q\left(x\right)} \left|\nabla u\right|^{q(x)} \tag{B.11}$$

On the other hand, if $|\nabla u| > 1$ and $|\rho| \leq 1$, then since $q\left(x\right) > 1$ for almost all x we have that $\nabla u \cdot \rho = |\nabla u| \frac{\nabla u}{|\nabla u|} \cdot \rho \leq |\nabla u| \left[\frac{1}{q(x)} + \frac{q(x)-1}{q(x)} |\rho|^{\frac{q(x)}{q(x)-1}}\right]$ and so

$$\begin{aligned} \nabla u \cdot \rho - \frac{q\left(x\right)-1}{q\left(x\right)} \left|\rho\right|^{\frac{q(x)}{q(x)-1}} &\leqslant \frac{1}{q\left(x\right)} \left|\nabla u\right| + \left(\left|\nabla u\right| - 1\right) \frac{q\left(x\right)-1}{q\left(x\right)} \left|\rho\right|^{\frac{q(x)}{q(x)-1}} \\ &\leqslant \left|\nabla u\right| - \frac{q\left(x\right)-1}{q\left(x\right)} \end{aligned} \tag{B.12}$$

Combining Eq. B.10, Eq. B.11, and Eq. B.12, we have that

$$\sup_{\psi \in C^1\left(\overline{\Omega}, \mathbb{R}^n\right)\ |\psi| \leqslant 1} \int_\Omega \nabla u.\psi - \frac{\mathrm{q\left(x\right)}-1}{\mathrm{q\left(x\right)}} \left|\psi\right|^{\frac{\mathrm{q(x)}}{\mathrm{q(x)}-1}} dx = \int_\Omega \phi\left(x, \nabla v\right) dx$$

and so for all $u \in BV\left(\Omega\right)$, $\tilde{\Phi}_g\left(u\right) = \Phi_g\left(u\right)$, where Φ_g is defined in Eq. B.5.

C Proof of Fully Discrete Image Formation Model

A stochastic model for the data recorded by the ijth pixel of a CCD array is given

$$D_{ij} \propto Poisson(g_{ij}) + Normal(0, \sigma^2) \tag{C.1}$$

The Poisson component models the photon count, while the additive Gaussian term accounts for background noise in the recording electronics.

We consider the continuous image formation equation in a discrete manner. We denote a realization of the random variable D_{ij} by d_{ij}. The $n_x \times n_y$ array d, whose components are the d_{ij}'s, is called the noisy blurred discrete image. For each index pair (i, j), $d(i, j)$ is a realization of a Gaussian random variable with zero mean and variance σ^2 added to a realization of a Poisson random variable with mean and variance g_{ij}. These random variables are assumed to be independent of each other and independent of the random variables corresponding to the other pixels.

Thus, the continuous image formation model can be denoted in a discrete form.

A fully discrete model may be obtained truncating the region of integration to be the discrete union Ω_{ij} and each union has area $\Delta x \times \Delta y$, and let (x_i, y_j) denote the midpoint. we get a discrete image formation model

$$\mathrm{g_{ij}} = \sum\nolimits_{\mu=0}^{n_x-1} \sum\nolimits_{\nu=0}^{n_y-1} h(x_i - x_\mu, y_j - y_\nu) f(x_\mu, y_\nu) + \epsilon_{ij}^{quad} \tag{C.2}$$

The blurring process is sometimes assumed to be invariant under spatial translation. This means that the PSF is linear invariant under spatial translation, we simplifies the computation. Since the integral in

$$g(x, y) = \iint h(x - x', y - y') f\left(x', y'\right) dx' dy' + \eta(x, y) \tag{C.3}$$

can in principle compute the continuous image g using the convolution theorem $g = F^{-1}F(h)F(f)$. Here the continuous Fourier transform of a (possibly complex-valued) function f defined $\mathbb{R}$(d =2 for two-dimensional imaging) is given in a Fourier form. From $g = F^{-1}F(h)F(f)$, one can derive the Fourier inversion formula $f = F^{-1}[F(g)/F(h)]$. In this case, if $F(h)$ takes on zero values, the formula is not valid. If it takes on small non zero values, this reconstructed f is unstable with respect to perturbations in the data g.

Discrete convolution product can then be given by

$$\mathrm{d_{ij}} = \sum\nolimits_{\mu=0}^{n_x-1} \sum\nolimits_{\nu=0}^{n_y-1} t_{i-\mu, j-\nu, f_{\mu,\nu}} + \eta_{ij} \tag{C.4}$$

with PSF $t_{ij} = h(i\Delta x, j\Delta y)\Delta x \Delta y$. The Discrete convolution product defines a linear operator.

D Hausdorff Measure and Hausdorff Dimension

Given a real number $\alpha > 0$ we are going to define a Borel external measure $\mathcal{H}^\alpha$ on $\mathbb{R}^n$ with values in $[0, +\infty]$ which will comprehend and generalize the concepts of length (for $\alpha = 1$), area ($\alpha = 2$) and volume ($\alpha = 3$) of sets in $\mathbb{R}^n$. In particular if $M \subset \mathbb{R}^n$ is an m-dimensional regular surface then one will show that $\mathcal{H}^m(M)$ is the m-dimensional area of M. However, being an external measure, $\mathcal{H}^m$ is defined not only on regular surfaces but on every subset of $\mathbb{R}^n$ thus generalizing the concepts of length, area and volume. In particular, for $m = n$, it turns out that the Hausdorff measure $\mathcal{H}^n$ is nothing else than the Lebesgue measure of $\mathbb{R}^n$.

Given any fixed set $E \in \mathbb{R}^n$ one can consider the measures $\mathcal{H}^\alpha(E)$ with α varying in $[0, +\infty]$. We will see that for a fixed set E there exists at most one value α such that $\mathcal{H}^\alpha(E)$ is finite and positive; while for every other value β one will have $\mathcal{H}^\beta(E) = 0$ if $\beta > \alpha$ and $\mathcal{H}^\beta(E) = +\infty$ if $\beta < \alpha$. For example, if E is a regular 2-dimensional surface then only $\mathcal{H}^2(E)$ (which is the area of the surface) may possibly be finite and different from 0 while, for example, the volume of E will be 0 and the length of E will be infinite.

This can be used to define the dimension of a set (this is called the Hausdorff dimension). A very interesting fact is the existence of sets with dimension which is not integer, as happens for most fractals. Also, the measure $\mathcal{H}^\alpha$ is naturally defined on every metric space (X, d), not only on $\mathbb{R}^n$.

Definition D.0.2.0.1 *Let (X, d) be a metric space. Given $E \subset X$, we define the diameter of E as $diam(E) := \sup_{x,y \in E} d(x, y)$.*

Given a real number α, we consider the conventional constant

$$w_\alpha = \frac{\pi^{\alpha/2}}{\Gamma(\alpha/2 + 1)}$$

where $\Gamma(x)$ is the gamma function, which can be thought of as the natural way to generalize the concept of the factorial to non-integer arguments. The first gamma function was by Euler(1729).

For all $\delta > 0, \alpha \geq 0$ and $E \subset X$, let us define

$$\mathcal{H}^\alpha_\delta(E) := \inf \left\{ \sum_{j=0}^{\infty} w_\alpha (\frac{diam(B_j)}{2})^\alpha : B_j \in X, \bigcup_{j=0}^{\infty} B_j \supset E, diam(B_j) \leq \delta, \forall j = 0, 1, \cdots \right\}$$

The infimum is taken over all possible enumerable families of sets $B_0, B_1, \ldots, B_j, \ldots$ which are sufficiently small ($diam(B_j) \leq \delta$) and which cover E.

Notice that the function $\mathcal{H}^\alpha_\delta(E)$ is decreasing in δ. In fact given $\delta' > \delta$ the family of sequences B_j considered in the definition of $\mathcal{H}^\alpha_{\delta'}$ contains the family of sequence considered in the definition of $\mathcal{H}^\alpha_\delta$ and hence the infimum is smaller. So the limit in the following definition exists:

$$\mathcal{H}^\alpha_\delta(E) := \lim_{\delta \to 0^+} \mathcal{H}^\alpha_\delta(E) \tag{D.1}$$

The number $\mathcal{H}^\alpha_\delta(E) \in [0, +\infty]$ is called α-dimensional Hausdorff measure of the set $E \in X$.

E Bibliography

[1] F. Abramovich and B. W. Silverman. Wavelet decomposition approaches to statistical inverse problems. *Biometrika*, 85:115–129, 1998.

[2] R. Acar and C. R. Vogel. Analysis of bounded varaition penalty methods for ill-posed problems. *Inverse problems*, 10(6):1217–1229, 1994.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Control*, AC-19:716723, 1974.

[4] S. Alliney and S. A. Ruzingsky. An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation. *IEEE Trans. on Signal Procesing*, 42(3):618–627, 1994.

[5] L. Alvarez and Y. Gousseau. Scales in natural images and a consequence on their bounded variation norm. In *Scale-Space*, volume Lectures Notes on Computer Science, 1682, 1999.

[6] L. Alvarez and L. Mazorra. Signal and image restoration using shock filters and anisotropic diffusion. *SIAM J. Numer. Anal.*, 31(2):590–605, 1994.

[7] L. Alvarez, L. P.L., and J. Morel. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.*, 29:845–866, 1992.

[8] S. Amari. *Differential-Geometrical Methods in Statistics*. Lecture Notes in Satistics, Springer-Verlag, 1985.

[9] S. Amari. Neutral learning in structured parameter spaces- natural Riemannian gradient. *NIPS'96*, 9:MIT Press, 1996.

[10] S. Amari and A. Cichocki. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, 8, 1996.

[11] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Science Publcations, 2000.

[12] H. Andrews and B. Hunt. *Digital image restoration*. Upper Saddle River, NJ: Prentice-Hall, 1977.

[13] C. Andrieu and A. Doucet. Joint Bayesian model selection and estimation of noisy sinusoidsvia reversible jump MCMC. *IEEE Transactions on Signal Processing*, 47(10):2667–2676, 1999.

[14] H. Attias. A variational Bayesian framework for graphical models. In T. e. a. Leen, editor, *Advances in NIPS*, pages 209–215, 2000.

[15] G. Aubert, R. Deriche, and P. Kornprobst. Computing optical flow via variational techniques. *SIAM Journal Numer. Anal.*, 34(5):1948–1979, 1999.

[16] l. Aubert and P. Kornprobst. *Mathematical problems in image processing: partical differential equations and the Calculus of Variations*. Springer, 2002.

[17] l. Aubert and L. Vese. A variational method in image recovery. *SIAM Journal Numer. Anal.*, 34(5):1948–1979, 1997.

[18] G. Ayers and J. Dainty. Iterative blind deconvolution method and its application. *Optic Letters*, 13:547–549, 1988.

[19] M. Banham and A. Katsaggelos. Digital image restoration. *IEEE S. P.*, 14:24–41, 1997.

[20] A. Barbu and S. Zhu. Graph partition by Swendsen-Wang cut. In *In Proc. of International Conf. on Computer Vision*, Nice, France, 2003.

[21] G. Bellettini, V. Caselles, and M. Novaga. The total variation flow in R^N. *Journal of Differential Equations*, (184):475–525, 2002.

[22] A. Ben-Tal, A.and Nemirovski. *Lectures on Modern Convex Optimization*. SIAM, 2004.

[23] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. New York: Wiley, Series in Applied Probability and Statistics, 1994.

[24] M. Bertero, C. D. Mol, and E. R. Pike. Linear inverse problems with discrete data: II. stability and regularization. *Inverse Problems*, 4:573–594, 1988.

[25] M. Bertero, T. A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, 1988.

[26] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, 48:259 – 302, 1986.

[27] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[28] C. M. Bishop and M. E. Tipping. Bayesian regression and classification. In *Advances in Learning Theory: Methods, Models and Applications*, pages 267–285, 2003.

[29] M. Black, G. Sapiro, D. Marimont, and D. Heeger. Robust anisotropic diffusion. *IEEE Trans. on Image Processing*, 7:421–432, 1998.

[30] M. J. Black. *Robust Incremental Optical Flow*. PhD thesis, Yale University, Department of Computer Science, 1992.

[31] A. Blake. Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. *IEEE Trans. PAMI*, 11:2–12, 1989.

[32] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In T. Pajdla and J. Matas, editors, *ECCV*, LNCS 3021, pages 428–441. Springer, May 2004.

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

[34] G. Bouchitte and M. Valadier. Integral repsentation of convex functionals on a space of measures. *Journal of Functional Analysis*, 80:398–420, 1988.

[35] C. Bouman and K. Sauer. A generalized gaussian image model for edge-preserving MAP estimation. *IEEE Transactions of Image Processing*, 2:296– 310, 1993.

[36] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[37] A. Bruhn, J. Weickert, T. Kohlberger, and C. Schnörr. A multigrid platform for real-time motion computation with discontinuity-preserving variational methods. *IJCV*, 70(3):257–277, 2006.

[38] F. Catte, P. L. Lions, J. M. Morel, and T. Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM, J. Num. Anal.*, 29(1):182–193, 1992.

[39] A. Chambolle and P. L. Lions. Image recovery via total variation minimization and related problems. *Numer. Math.*, 76(2):167–188, 1997.

[40] T. Chan, Osher, J. Shen, and L. Vese. Variational PDE models in image processing. *Notice of Am. Math. Soc.*, 50:14–26, 2003.

[41] T. F. Chan, S. H. Kang, and J. Shen. Euler's elastica and curvature based inpaintings. *SIAM J. Appli. Math*, pages 564–592, 2002.

[42] T. F. Chan and J. Shen. Theory and computation of variational image deblurring. *Lecture Notes on "Mathematics and Computation in Imaging Science and Information Processing"*, 2006.

[43] T. F. Chan and J. H. Shen. *Image Processing and Analysis - Variational, PDE, wavelet, and stochastic methods*. SIAM Publisher, Philadelphia, 2005.

[44] T. F. Chan and C. K. Wong. Total variation blind deconvolution. *IEEE Trans.on Image Processing*, 7(3):370–375, 1998.

[45] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Trans.on Image Processing*, 6(2):298–311, 1997.

[46] L. Chen and K. H. Yap. Efficient discrete techniques for blur support identification in blind image deconvolution. *IEEE Tran. Sig. Pro.*, 54:1557–1562, 2006.

[47] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[48] Y. Chen, S. Levine, and M. Rao. Variable exponent, linear growth functionals in image restoration. *SIAM Journal of Applied Mathematics*, 66(4):1383–1406, 2006.

[49] Y. Chen and M. Rao. Minimization problems and associated flows related to weighted p energy and total variation. *SIAM Journal of Applied Mathematics*, 34:1084–1104, 2003.

[50] Y. Chuang, B. Curless, D. Salesin, and R. Szeliski. A Bayesian approach to digital matting. In *CVPR*, 2001.

[51] F. Chung. *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics, American Mathematical Society, 1997.

[52] O. Coulon and S. Arridge. Dual echo MR image processing using multi-spectral probabilistic diffusion coupled with shock filters. In *British Conference on Medical Image Understanding and Analysis*, 2000.

[53] I. Cox, S. Rao, and Y. Zhong. "Ratio regions": a technique for image segmentation. volume 2, pages 557 – 564, 1996.

[54] P. Craven and G. Wahba. Smoothing noisy data with spline functions-estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

[55] D. Cremers. Dynamical statistical shape priors for level set based tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(8):1262–1273, August 2006.

[56] D. Cremers, S. J. Osher, and S. Soatto. Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. *International Journal of Computer Vision*, 69(3):335–351, September 2006.

[57] L. Csato, M. Opper, and O. Winther. TAP Gibbs free energy, belief propagation and sparsity. *Advances in Neural Information Processing Systems*, 14, 2002.

[58] K. Deimling. *Nonlinear Functional Analysis*. Springer-Verlag, Berlin, 1985.

[59] F. Demengel and R. Teman. Convex functions of a measure and applications. *Indiana University Mathematics Journal*, 33:673–709, 1984.

[60] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal Royal. Statistical Society B*, 39:1–38, 1977.

[61] S. Didas and J. Weickert. Integrodifferential equations for continuous multiscale wavelet shrinkage. *Inverse Problems and Imaging*, 1:47–62, 2007.

[62] R. L. Dobrushin. The description of a random filed by means of conditional probabilities and conditions of its regularity. *Theory Probability Application*, 13:197–224, 1968.

[63] R. Duda and P. Hart. *Pattern Classification*. Wiley Interscience, 2 edition, 2000.

[64] B. Efron and C. Morris. Data analysis using Stein's estimator and its generalization. *Journal of the American Statistical Association*, 70:350, 1975.

[65] H. Ehrig, K. Ehrig, U. Prange, and G. Taentzer. *Fundamentals of Algebraic Graph Transformation*. EATCS Monographs in Theoretical Computer Science, Springer, 2006.

[66] I. Ekeland and R. Temam. *Convex Analysis and Variational Problems*. North-Holland Publ. Comp., 1976.

[67] J. H. Elder and S. W. Zucker. Local scale control for edge detection and blur estimation. *IEEE Trans. on PAMI*, 20(7):699–716, 1998.

[68] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

[69] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, FL, 1992.

[70] M. P. Evgeniou, T. and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

[71] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19:1257–1272, 1991.

[72] L. Fei-Fei and P. Perona. A Bayesian hierarcical model for learning natural scene categories. *Proc. CVPR*, 2005.

[73] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. Freeman. Removing camera shake from a single photograph. *SIGGRAPH 2006 Conference Proceedings, Boston, MA*, 25:787–794, 2006.

[74] D. J. Field. Relations between the statistics and natural images and the responses properties of cortical cells. *J. Optical Soc. Am. A.*, 4:2379–2394, 1987.

[75] D. J. Field. What is the goal of sensory coding? *Neural Comput.*, 6:559–601, 1994.

[76] R. Fielding. *The Technique of Special Effects Cinematography*. Focal/Hastings House, 3rd edition edition, 1972.

[77] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. on PAMI*, 24(3):381–396, 2002.

[78] M. A. T. Figueiredo and J. Leitao. Unsupervised image restoration and edge location using compound Gauss-Markov random fields and the MDL principle. *IEEE Trans. Image Processing*, 6:1089–1102, 1997.

[79] W. Förstner. A feature based correspondence algorithm for image matching. *Int. Arch. Photogrammetry Remote Sensing*, 26(3):150–166, 1986.

[80] W. Freeman and E. Pasztor. Learning low-level vision. In K. Academic, editor, *International Journal of Computer Vision*, volume 40, pages 24–57, 2000.

[81] D. Geiger, Girosi, and Federico. Parallel and deterministic algorithms from MRF's surface reconstruction. *IEEE Trans. Pattern Anal. Machine Intell.*, 13(5):401–412, 1991.

[82] S. Gelfand and S. Mitter. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 29:999–1018, 1991.

[83] S. Gelfand and S. Mitter. Metropolis-type annealing for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 31:111–131, 1993.

[84] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Trans. Image Processing*, 4:932–946, 1995.

[85] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. on PAMI*, 6:721–741, 1984.

[86] S. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Trans. PAMI.*, 14:367–383, 1992.

[87] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* 1992.

[88] W. Gibbs. *Elementary Principles of Statistical Mechanics.* Yale University Press, 1902.

[89] G. Gilboa, N. A. Sochen, and Y. Y. Zeevi. Regularized shock filters and complex diffusion. *A. Heyden et al. (Eds.): ECCV 2002, LNCS 2350*, pages 399–413, 2002.

[90] P. E. Gill, W. Murray, and W. H. Wright. *Practical optimization.* Academic Press, New York, 1981.

[91] F. Girosi, J. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

[92] E. Giusti. *Minimal Surfaces and Functions of Bounded Variation.* Birkhäuser, 1984.

[93] C. Goffman and J. Serrin. Sublinear functions of measures and variational integrals. *Duke Math. J.*, 31:159–178, 1964.

[94] H. M. Golub, G. and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:205, 1979.

[95] Y. Gousseau and J.-M. Morel. Are natural images of bounded variation? *SIAM Journal on Mathematical Analysis*, 33:634–648, 2001.

[96] L. Grady. Random walks for image segmentation. *IEEE Trans on. PAMI*, 28(11):1768–1783, 2006.

[97] P. Green. Bayesian reconstruction from emission tomography data using a modified EM algorithm. *IEEE Tr. Med. Imaging*, 9:84–92, 1990.

[98] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society, Series B*, 51(2):271, 1989.

[99] C. Groetsch and O. Scherzer. Nonstationary iterated Tikhonov-Morozov method and third order differential equations for the evaluation of unbounded operators. *Math. Methods Appl. Sci.*, 23:1287–1300, 2000.

[100] C. W. Groetsch. *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind.* Pitman, Boston, 1984.

[101] C. W. Groetsch. *Inverse Problem in the Mathematical Sciences.* Vieweg Verla, Wiesbaden, 1993.

[102] J. Hadamard. *Lectures on the Cauchy Problem in Linear Partial Differential Equations.* Yale University Press, 1923.

[103] M. Hahn. *Bildsequenzanlyse für die passive Navigation.* PhD thesis, Universität Stuttgart, 1995.

[104] P. Hansen. Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank. *SIAM J. Sci. Stat. Comput.*, 11:503–518, 1990.

[105] P. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion.* Phiadelphia, 1997.

[106] P. Hansen and D. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14:1487–1503, 1993.

[107] R. Hardt and X. Zhou. An evolution problem for linear growth functionals. *Comm. Partial Differential Equations*, 19:1879–1907, 1994.

[108] T. Hebert and R. Leahy. A generalized EM algorithm for 3D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans. Medical Imaging*, 8(2):194–202, 1989.

[109] M. Heiler and C. Schnörr. Natural image statistics for natural image segmentation. *IJCV*, 63(1):5–19, 2005.

[110] M. Heiler and C. Schnörr. Learning sparse representations by non-negative matrix factorization and sequential cone programming. *Journal of Machine Learning Research*, 7:1385–1407, 2006.

[111] O. Hellwich. *Linienextraktion aus SAR-Daten mit einem Markoff-Zufallsfeld-Modell*. PhD thesis, TU Muenchen, 1997.

[112] O. Hellwich. Geocoding SAR interferograms by least squares adjustment. *ISPRS Journal of Photogrametry and Remote Sensing*, 55(4):277–288, 2000.

[113] O. Hellwich, H. Mayer, and G. Winkler. Detection of lines in synthetic aperture radar (SAR) scenes. In *Int. Archives Photogrammetry Remote Sensing (ISPRS)*, volume 31, pages 312–320, Vienna, Austria, 1996.

[114] G. E. Hinton and D. v. Camp. Keeping neural networks simple by minimizing the description length of the weightss. In *Sixth Annual ACM Conference on Computational Learning Theory*, pages 5–13, 1993.

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

[116] D. Huttenlocher, G. A. Klanderman, and W. Rucklidge. Comparing images using the Haussdorff distance. *IEEE on PAMI*, 15:850–863, 1993.

[117] D. W. Jacobs, D. Weinshall, and Y. Gdalyahu. Class representation and image retrieval with non-metric distances. In *International Conference Computer Vision*, 1998.

[118] B. Jähne and Haussecker. *Computer Vision and Applications*. AP Academic Press, London, 2000.

[119] A. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Englewood Cliffs, 1989.

[120] A. Jain and R. Dubes. *Algorithms for Clustering Data*. New Jersy Prentice Hall, Englewood Cliffs, 1988.

[121] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *In Learning Graphical Models, MIT Press*, pages 105–161, 1999.

[122] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

[123] R. L. Kashyap and R. Chellappa. Estimation and choice of neighbors in spatial-interaction models of images. *IEEE Transactions on Information Theory*, 29(1):60–72, 1983.

[124] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[125] A. Katsaggelos, J. Biemond, R. Mersereau, and R. Schaefer. Nonstationary image restoration. In *in Proc. IEEE Int. Conf. Acoustics Speech, Signal Processing*, pages 696–699, 1985.

[126] A. Katsaggelos, J. Biemond, R. Schafer, and R. Mersereau. A regularized iterative image restoration algorithm. *IEEE Tr. on Signal Processing*, 39:914–929, 1991.

[127] A. k. Katsaggelos. *Digital Image Restoration*. Springer-Verlag, 1991.

[128] P.-K. P. Kazakos, D. *Detection and estimation*. Computer Science Press, New York, w. h. freeman and company, edition, 1990.

[129] J. Keller. Inverse problems. *American Math.*, 83:107–118, 1976.

[130] J. Keuchel, C. Schnörr, C. Schellewald, and D. Cremers. Binary partitioning, perceptual grouping, and restoration with semidefinite programming. *IEEE Trans. on PAMI*, 25:1364–1379, 2003.

[131] B. Kimia, A. Tannebaum, and S. Zucker. Shapes, shocks, and deformations I: The components of two-dimensional shape and the reaction-diffusion space. *Int'l J. Computer Vision*, 15:189–224, 1995.

[132] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004.

[133] P. Kornprobst, R. Deriche, and G. Aubert. Image coupling, restoration and enhancement via PDE. In *Proc. Int. Conf. on Image Processing*, page 458C461. Springer, 1997.

[134] D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE Signal Process. Mag.*, May(5):43–64, 1996.

[135] R. Lagendijk, J. Biemond, and D. Boekee. Regularized iterative image restoration with ringing reduction. *IEEE Tr. on Ac., Sp., and Sig. Proc.*, 36(12):1874–1888, 1988.

[136] R. Lane. Blind deconvolution of speckle images. *J. Opt. Soc. Amer.A*, 9:1508–1514, 1992.

[137] A. Lanterman. Schwarz, Wallace and Rissanen: Intertwining themes in theories of model order estimation. *Int'l Statistical Rev.*, 69:185–212, 2001.

[138] V. M. Lavrentev, V. Romanov, and S. Shishatskii. *Ill-posed Problems of Mathematical Physics and Analysis*. American Mathematical Society, Providence, Rhode Island, 1997.

[139] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 40(10):788–791, 2001.

[140] K. Leibovic. *Science of Vision*. Springer-Verlag, 1990.

[141] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *ACM Trans. Graph*, 2004.

[142] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. In *CVPR*, 2006.

[143] S. Li. *Markov Random Field Modeling in Computer Vision*. Springer Verlag, 1995.

[144] Y. Li, J. Sun, C. Tang, and H. Shum. Lazy snapping. *ACM SIGGRAPH*, pages 303–308, 2004.

[145] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Comm.*, 28:84–95, 1980.

[146] T. Lindeberg. Edge detection and ridge detection with automatic scale selection. *International Journal of Computer Vision*, pages 117–156, 1998.

[147] T. Linderberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[148] B. G. Lindsay. The geometry of mixture likelihoods: A general theory. *Annals of Statistics,*, 11(1):86–94, 1983.

[149] H. Lombaert, Y. Sun, L. Grady, and C. Xu. A multilevel banded graph cuts method for fast image segmentation. In *ICCV*, page 259265, 2005.

[150] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[151] D. G. Luenberger. *Optimization by vector space methods*.

[152] D. G. Luenberger. *Linear and nonlinear programming*. Addison-Wesley Publishing Company, 1984.

[153] M. Luxen and W. Förstner. Characterizing image quality: Blind estimation of the point spread function from a single image. In *PCV02*, page A: 205, 2002.

[154] D. J. Mackay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[155] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International Journal of Computer Vision*, 43(1):7C27, 2001.

[156] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE on PAMI*, 11:674–693, 1989.

[157] D. Marr. Early processing of visual information. *Phil. Trans. R. Soc. London*, 1976.

[158] D. Marr and E. Hildreth. Theory of edge detection. *Proc. Royal Soc. London*, 1980.

[159] D. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. PAMI*, 26:530–549, 2004.

[160] S. Martucci. Symmetric convolution and the discrete sine and cosine transforms. *IEEE Trans. Signal Process.*, 42:1038–1051, 1994.

[161] B. C. McCallum. Blind deconvolution by simulated annealing. *Optics Communication*, 75(2):101–105, 1990.

[162] G. J. McLachlan. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, 36(3):318–324, 1987.

[163] G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, 2000.

[164] Y. Meyer. Oscillating patterns in image processing and in some nonlinear evolution equations. *The 15th Dean Jacquelines B. Lewis Memorial Lectures*, 3, 2001.

[165] K. Miller. Least-squares method for ill-posed problems with a prescribed bound. *SIAM Journal of Math.*, 1:52–74, 1970.

[166] J. Miskin and D. J. C. MacKay. Ensemble Learning for Blind Image Separation and Deconvolution. In M. Girolani, editor, *Adv. in Independent Component Analysis*. Springer-Verlag, 2000.

[167] J. W. Miskin. *Ensemble Learning for Independent Component Analysis*. PhD thesis, University of Cambridge, December 2000.

[168] B. Mohar. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications.*, 2:871–898, 1991.

[169] R. Molina, A. Katsaggelos, and J. Mateos. Bayesian and regularization methods for hyperparameters estimate in image restoration. *IEEE Trans.on Sig. Pro.*, 8:231–246, 1999.

[170] R. Molina and B. Ripley. Using spatial models as priors in astronomical image analysis. *J. App. Stat.*, 16:193–206, 1989.

[171] V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7:414–417, 1966.

[172] V. A. Morozov. *Methods for Solving Incorrectly Posed Problems*. Springer, New York, 1984.

[173] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics*, 42:577–684, 1989.

[174] M. Nashed. Operator-theoretic and computational approaches to ill-posed problems with applications to antenna theory. *IEEE Transactions on Antennas and Propagation*, 29:220 –231, 1981.

[175] M. K. Ng, R. H. Chan, and W.-C. Tang. A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM Journal on Scientific Computing*, 21(3):851–866, 1999.

[176] M. Nikolova. Markovian reconstruction using a GNC approach. *IEEE Trans. on Image Processing.*, 8(9):1204–1220, 1999.

[177] M. Nikolova. Weakly constrained minimization: application to the estimation of images and signals involving constant regions. *J. Math. Image Vision*, 21(2):155–175, 2004.

[178] D. P. Oleary. Robust regression computation using iteratively reweighted least squares. *SIAM J. Matrix Anal. Appl.*, 11:466–480, 1990.

[179] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[180] M. Opper. On-line versus off-line learning random examples: Gerneral results. *Physical Review letters, The American Physical Society*, 77:4671–4674, 1996.

[181] M. Opper and D. Saad. *Advanced Mean Field Methods: Theory and Practice*. The MIT Press, Cambridge, Massachusetts, 2001.

[182] M. Opper and O. Winther. Gaussian processes and SVM: Mean field results and leave-one-out. In A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *In Advances in Large Margin Classifiers*, pages 311–326. MIT Press, 2000.

[183] M. Opper and O. Winther. Expectation consistent free energy for approximate inference. *The Journal of Machine Learning Research*, 6:2177–2204, 2005.

[184] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi equations. *Journal of Computational Physics*, 79:12–49, 1988.

[185] S. J. Osher and L. I. Rudin. Feature-oriented image enhancement using shock filters. *SIAM J. Numer. Anal.*, 27:919–940, 1990.

[186] S. J. Osher and L. I. Rudin. Shocks and other nonlinear filtering applied to image processing. *SPIE Appl. Dig. Image Proc.*, 1567:414–430, 1991.

[187] J.-S. Pan, Z.-M. Lu, and S.-H. Sun. An efficient encoding algorithm for vector quantization based on subvector technique. *IEEE Trans. Image Processing*, 3:265–270, 2003.

[188] A. Papoulis. *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 1965.

[189] A. Pentland. *Perceptual organization and the representation of natural form*. Morgan Kaufmann Publishers Inc., 1987.

[190] P. Perona and W. T. Freeman. A factorization approach to grouping. *ECCV, LNCS Springer*, 12:629–639, 1999.

[191] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. PAMI.*, 12:629–639, 1990.

[192] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.

[193] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

[194] T. Poggio, F. Girosi, and M. Jones. From regularization to radial, tensor and additive splines. *Proceedings of the 1993 International Joint Conference on Neural Networks*, pages 223–227, 1993.

[195] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[196] J. Polzehl and V. Spokoiny. Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society*, 2000.

[197] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *IJCV*, 40(1):49–72, 2000.

[198] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli. Image denoising using scale mixtures of Gaussians in the Wavelet domain. *IEEE Trans. on Image Processing*, 12(11):1338–1351, 2003.

[199] A. Pothen, H. Simon, and K. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM. J. Matrix Anal. App.*, 11:435–452, 1990.

[200] S. W. Ra and J. K. Kim. A fast mean-distance-ordered partial codebook search algorithm for image vector quantization. *IEEE Trans. Circuits and Systems II: Analog and Digital Signal Processing*, 40:576–579, 1993.

[201] A. Raftery, D. Madigan, and J. Hoeting. Bayesian model averaging for regression models. *Journal of the American Statistical Association*, pages 179–191, 1992.

[202] S. J. Reeves and R. M. Mersereau. Blur identification by the method of generalized cross-validation. *IEEE Trans. Image Processing*, 1(3):301–311, 1992.

[203] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[204] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, 1989.

[205] M. Rivera and J. Marroquin. Adaptive rest condition potentials: second order edge-preserving regularization. In *ECCV 2002*, LNCS 2350, pages 113–127, Springer, Berlin, 2002. A. Heyden et al. (Eds).

[206] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1972.

[207] V. Rodehorst and A. Koschan. Comparison and evaluation of feature point detectors. In *Proc. of 5th International Symposium Turkish-German Joint Geodetic Days*, Berlin, 2006.

[208] B. M. Romeny. *Gemetry-driven diffusion in computer vision*. Kluwer Academic Publishers, 1994.

[209] S. Roth and M. Black. Fields of experts: A framework for learning image priors. In *CVPR*, pages 860–867, San Diego, 2005.

[210] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

[211] S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11(2), 2004.

[212] L. Rudin and S. Osher. Total variation based image image restoration with free local constraints. *Proc. IEEE ICIP*, 1:31–35, 1994.

[213] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total varition based noise removal algorithm. *Physica D*, 60:259–268, 1992.

[214] M. Ruzon and C. Tomasi. Alpha estimation in natural images. In *CVPR*, 2000.

[215] S. Sarkar and K. Boyer. Quantitative measures of change based on feature organization: eigenvalues and eigenvectors. In *IEEE Conf. Computer Vision and Pattern Recognition*, 1996.

[216] S. Sarkar and P. Soundararajan. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. *IEEE Trans. PAMI.*, 22(5):504–525, 2000.

[217] O. Scherzer and C. Groetsch. Inverse scale space theory and inverse problems. pages 317–325. Springer, 2001.

[218] O. Scherzer and J. Weickert. Relations between regularization and diffusion filtering. *Journal of Mathematical Imaging and Vision*, 12:43–63, 2000.

[219] C. Schnörr. Unique reconstruction of piecewise smooth images by minimizing strictly convex nonquadratic functionals. *Journal of Math. Imaging Vision*, 4:189–198, 1994.

[220] C. Schnörr. A study of a convex variational diffusion approach for image segmentation and feature extraction. *Journal of Math. Imaging Vision*, 8:271–292, 1998.

[221] C. Schnörr and R. Sprengel. A nonlinear regularization approach to early vision. *Journal Biological Cybernetics*, 72(2):141–149, 1994.

[222] B. Schölkopf and A. J. Smola. *Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, Cambridge, 2002.

[223] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[224] G. Scott and H. C. Longuet-Higgins. Feature grouping by relocalisation of eigenvectors of the proxmity matrix. In *British Machine Vision Conference*, pages 103–108, 1990.

[225] C. E. Shannon. A mathematical theory of communication. *Bell System Technology Journal*, 27:379–423/623–656, 1948.

[226] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(3):888–905, 2000.

[227] E. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *In 31st Asilomar Conf. on Sig., Sys. and Computers*, Pacific Grove, CA, 1997.

[228] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. In *Proc. Int'l Conf. Computer Vision*, Beijing, 2005.

[229] W. Snyder. Image relaxation: Restoration and feature extraction. *IEEE Trans. Pattern Anal. Machine Intell.*, 17:620624, 1995.

[230] N. Sochen. Stochastic process in vision I: From Langevin to Beltrami. *Technion CCIT report 285*, 1999.

[231] V. Spokoiny. *Local Parametric Methods in Nonparametric Estimation*. Springer, 2006.

[232] L. Staganski and R. Carroll. Deconvoluting kernel density estimators. *Statistics*, 21:169–184, 1990.

[233] J. L. Starck and A. Bijaoui. Filtering and deconvolution by the wavelet transform. *Signal Processing,*, 35:195–211, 1994.

[234] R. Stevenson and E. Delp. Fitting curves with discontinuities. pages 127–136, 1990.

[235] J. Sun, W. Zhang, X. Tang, and H. Shum. Background cut. In *ECCV*, pages 628–641, 2006.

[236] P. Symth. Model selection for probabilistic clustering using cross-validated likelihiood. *Statistics and Computing*, 1:63–72, 2000.

[237] R. Szeliski. Bayesian modeling of uncertainty in low-level vision. *Internatinoal Journal of Computer Vision*, 5:271–301, 1990.

[238] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields. In *ECCV 2006*, 2006.

[239] G. Talenti. *Inverse Problems*. Lecture Notes in Mathematics 1225, Berlin, 1986.

[240] L. H. Thomas. Elliptic problems in linear difference equations over a network. Technical report, Watson Scientific Computing Laboratory, Columbia University, New York, NJ, 1949.

[241] A. Tikhonov and V. Arsenin. *Solution of Ill-Posed Problems*. Wiley, Winston, 1977.

[242] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester, U.K., 1985.

[243] H. J. Trussell and B. R. Hunt. dd. *IEEE Trans. on Image Processing*, 1979.

[244] D. Tschumperle and R. Deriche. Diffusion tensor regularization with contraints preservation. In *Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[245] D. Tschumperle and R. Deriche. Orthonormal vector sets regularization with PDE's and application. *International Journal of Computer Vision (IJCV, Specical Issue VLSM)*, 2002.

[246] D. Tschumperle and R. Deriche. Vector-valued image regularization with PDEs: a common framework for different applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4), 2005.

[247] Z. Tu, X. Chen, A. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and object recognition. *International Journal of Computer Vision*, 2005.

[248] V. Vapnik. *Statistical Learning Theory*. A Wiley-Interscience Publication, New-York, NY, USA, 1998.

[249] L. A. Vese. A study in the BV space of a denoising-deblurring variational problem. *Applied Methematics and Optimization*, 44:131–161, 2001.

[250] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM J. Scientific Computing*, 17(1):227–38, 1996.

[251] C. R. Vogel and M. E. Oman. Fast, robust total variation-based reconstruction of noisy, blurred images. *IEEE Trans. on Image Processing*, 7:813–824, 1998.

[252] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBNS-NSF Regional Conference series in applied mathematics*. SIAM, 1990.

[253] C. Wallace and P. Freeman. Estimation and inference via compact coding. *J. Royal Statistical Soc.(B)*, 49(3):241–252, 1987.

[254] S. Wang and J. M. Siskind. Image segmentation with Ratio Cut. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(6), 2003.

[255] L. Wasserman. Bayesian model selection and model averaging. In *Mathematical Psychology Symposium on "Methods for Model Selection"*, 1997.

[256] J. Weickert. Anisotropic diffusion filters for image processing based quality control. In A. Fasano and M. Primicero, editors, *Proceedings of Seventh European Conf. on Mathematics in Industry*, pages 355–362, 1994.

[257] J. Weickert. Nonlinear diffusion scale-spaces: From the continuous to the discrete setting. In *Proc. ICAOS: Images, Wavelets, PDE*, ICAOS, pages 111–118, Paris, 1996.

[258] J. Weickert. A review of nonlinear diffusion filtering. *Lecture Notes in Computer Sciences*, 1252:3–28, 1997.

[259] J. Weickert. *Anisotropic Diffusion in Image Processing*. Teubner-Verlag, Stuttgart, 1998.

[260] J. Weickert. On discontinuity-preserving optic flow. In J. C. N. K. E. S. Orphanoudakis, P. Trahanias, editor, *Computer Vision and Mobile Robotics Workshop*, CVMR'98, pages 115–122, Santorini, 1998.

[261] J. Weickert. Coherence-enhancing diffusion filtering. *International Journal of Computer Vision*, 31:111–127, 1999.

[262] J. Weickert and T. Brox. Diffusion and regularization of vector- and matrix-valued images. In M. Z. Nashed and O. Scherzer, editors, *Inverse Problems, Image Analysis, and Medical Imaging*, page 251C268, 2002.

[263] J. Weickert and H. Hagen. *Visualization and Processing of Tensor Fields*. Springer-Verlag Berlin Heidelberg, 2006.

[264] J. Weickert, S. Ishikawa, and A. Imiya. Linear scale-space has first been proposed in japan. *Journal of Mathematical Imaging and Vision*, 10:237252, 1999.

[265] J. Weickert, B. M. t. H. Romeny, and M. Viergever. Efficient and reliable schemes for nonlinear diffusion filtering. *IEEE Transactions on Image Processing*, 7(3):398–410, 1998.

[266] J. Weickert and C. Schnörr. Variational image motion computation: theoretical framework, problems and perspectives. In G. Sommer, N. Krüger, and C. Perwass, editors, *DAGM Invited Paper*, pages 476–487, 2000.

[267] J. Weickert and C. Schnörr. A theoretical framework for convex regularizers in PDE-based computation of image motion. *International Journal of Computer Vision*, 45(3):245–264, 2001.

[268] K. Weinberger and L. K. Saul. Unsupervised learning of image manifold by semidefinite programming. In *CVPR*, 2004.

[269] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *In Proceedings of the IEEE International Conference on Computer Vision*, volume 7, pages 975–982, Los Alamitos, CA, USA, 1999.

[270] M. Wertheimer. Untersuchungen zur Lehre von der Gestalt II. In *Psychologische Forschung*, pages 4:301–350, 1923.

[271] J. Westlake. *A Handbook of Numerical Matrix Inversion and Solution of Linear Equationa*. Wiley, New York, 1968.

[272] G. Winkler. Aspekte der kantenerhaltenden glättung. GSF Bericht 17/00, National Research Center for Environment and Health, Neuherberg-Munich, Germany, 2000.

[273] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*. Springer-Verlag Berlin, Heidelberg, New York, 2nd edition edition, 2002.

[274] P. H. Winston. *Artificial Intelligence*. Addison-Wesley, 2 edition edition, 1984.

[275] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993.

[276] J. Xu. *Iterative Regularization and Nonlinear Inverse Scale Space Methods in Image Restoration*. PhD thesis, Department of Mathematics, University of California, Los Angeles, 2006.

[277] J. Xu and S. Osher. Iterative regularization and nonlinear inverse scale space applied to Wavelet based denoising. *IEEE Transactions of Image Processing*, 2006.

[278] Y. Yang, N. Galatsanos, and H. Stark. Projection based blind deconvolution. *Jounral Optical Society America.*, 11:2401–2409, 1994.

[279] G. L. Yap, Kim-Hui and L. W.Q. A recursive soft-decision approach to blind image deconvolution. *IEEE Tr. Sig. Pro.*, 51:515–526, 2003.

[280] Y. You and M. Kaveh. A regularization approach to joint blur identification and image restoration. *IEEE Tr. on Image Processing*, 5(3):416–428, 1996.

[281] D. Yu and J. Fessler. Edge-preserving tomographic reconstruction with nonlocal regularization. *IEEE Trans. on Medical Imaging*, 21(2):159–173, 2002.

[282] S. X. Yu and J. Shi. Segmentation given partial grouping constraints. *IEEE Trans. on PAMI*, 26(2):173–183, 2004.

[283] M. Yu-Li You; Wenyuan Xu; Tannenbaum, A.; Kaveh. Behavioral analysis of anisotropic diffusion in image processing. *IEEE Trans. on Image Processing*, 5(11):1539–1553, 1996.

[284] C. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, C-20:68–86, 1971.

[285] W. I. Zangwill. *Nonlinear programming: a unified approach*. Prentice-Hall, Englewood Cliffs, N.J., 1969.

[286] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *In Advances in Neural Information Processing Systems 17*. Springer, 2005.

[287] H. Zheng and O. Hellwich. Image statistics for nonstationary blurred image reconstruction. In *Pattern Recognition, DAGM 2007, to appear*.

[288] H. Zheng and O. Hellwich. Bayesian estimation based Mumford-Shah regularization for blur identification and segmentation in video sequences. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 129–133, 2006.

[289] H. Zheng and O. Hellwich. Double regularized Bayesian estimation for blur identification in video sequences. In *P.J. Narayanan et al. (Eds.),the 7th Asian Conference on Computer Vision (ACCV) 2006*, volume 3852 of *LNCS*, pages 943–952. Springer, 2006.

[290] H. Zheng and O. Hellwich. An edge-driven total variation approach to image deblurring and denoising. pages 705–709. IEEE International Conference on Innovative Computing, Information and Control, ICICIC, 2006.

[291] H. Zheng and O. Hellwich. Extended Mumford-Shah regularization in Bayesian estimation for blind image deconvolution and segmentation. In *R. Reulke et. al. (Eds.), Proc. of the 11th International Workshop on Combinatorial Image Analysis (IWCIA) 2006*, volume 4040 of *LNCS*, pages 144–158. Springer, 2006.

[292] H. Zheng and O. Hellwich. Introducing dynamic prior knowledge to partially-blurred image restoration. In *K. Franke et al.(Eds), Pattern Recognition, DAGM 2006*, volume 4174 of *LNCS*, pages 111–121. Springer, 2006.

[293] H. Zheng and O. Hellwich. Joint prior models of Mumford-Shah regularization for blur identification and segmentation in video sequences. In *Proc. of the International Conference on Computer Vision Theory and Applications (VISAPP) 2006*, volume 1, pages 56–63, Feb. 2006.

[294] H. Zheng and O. Hellwich. Variational regularized Bayesian estimation for joint blur identification and edge-driven image restoration. In *S. Wilson et. al. (Eds.), Proc. of the Computation Intensive Methods for Computer Vision (CIMCV) in the 9th ECCV, 2006*, pages 131–142, May 2006.

[295] H. Zheng and O. Hellwich. VQ-based Bayesian estimation for blur identification and image selection in video sequences. *International Journal of Innovative Computing, Information and Control, IJICIC*, 2(2):1–11, 2006.

[296] H. Zheng and O. Hellwich. Adaptive data-driven regularization for variational image restoration in the BV space. In *Proc. of the International Conference on Computer Vision Theory and Applications (VISAPP) 2007*, pages 53–61, 2007.

[297] H. Zheng and O. Hellwich. Adaptive data-driven regularization for variational image restoration in the BV space. *Journal of Mathematical Imaging and Vision*, Prepared, 2007.

[298] H. Zheng and O. Hellwich. Bayesian modeling of natural images for various blurred image identification, segmentation and restoration. *International Journal of Computer Vision*, Prepared, 2007.

[299] H. Zheng and O. Hellwich. Discrete regularization for perceptual image segmentation via semi-supervised learning and optimal control. In *IEEE International Multimedia & Expo, ICME*, pages 1982–1985, 2007.

[300] D. Zhou and B. Schölkopf. Regularization on discrete spaces. In *DAGM,Wien*. Springer., 2005.

[301] D. Zhou and B. Schölkopf. *Discrete Regularization*. Number 221-232. MIT Press, Cambridge, 2006.

[302] S. Zhu. Statistical modeling and conceptualization of visual patterns. *IEEE Trans. on PAMI.*, pages 691–712, 2003.

[303] S. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Trans. on PAMI*, 19(11):1236–1249, 1997.

[304] S. Zhu and X. Wu. Learning in Gibbsian field: How accurate and how fast can it be? *IEEE Trans. on PAMI*, 24(7):1001–1006, 2002.

[305] S. Zhu, X. Wu, and D. Mumford. Minimax entropy principle and its to texture modeling. *Neural Computation*, 9(8):1627–1660, 1997.

[306] S. Zhu and A. L. Yuille. Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *In Processings of International Conference on Computer Vision*, pages 416–423, 1995.

[307] S.-C. Zhu, C.-E. Guo, Y. Wang, and Z. Xu. Statistical modeling and conceptualization of visual patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25:691–712, 2003.

[308] S.-C. Zhu, C.-E. Guo, Y. Wang, and Z. Xu. What are textons? *International Journal of Computer Vision*, 62:121–143, 2005.

[309] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian field Harmonic functions. In *International Confernce on Machine Learning ICML*, Washington DC, 2003.

[310] W. P. Ziemer. *Weakly Differentiable Functions: Sobolov Space and Functions of Bounded Variation*. Springer-Verlag, 1980.

F List of Figures

G List of Tables

H List of Symbols and Abbreviation

Basic Notation of Statistics and Image Restoration

g, f, η, h	Degraded image, original image, additive noise, and blur.
$\hat{f}, \hat{h}, v$	Estimates to the original image, the blur and edge-curves.
$h_i(\theta)$	The i-th PSF parametric model with unknown parameters θ.
$\mathcal{J}(\hat{f}\|g, \hat{h})$	Image-domain MAP cost function.
$\mathcal{J}(\hat{h}\|g, \hat{f})$	Blur-domain MAP cost function.
$\mathcal{J}_\varepsilon(\hat{f}, \hat{h}, v)$	Approximated energy functional including all the estimates.
α, β, γ	Regularization parameters of image, edge and blur term.
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	A undirected weight graph with edges between vertices.
Δf	Laplacian operator: $\nabla f = \sum_{i=1}^{N} \frac{\partial^2 f}{\partial x_i^2}$.
$D^2 f$	Hessian matrix of f (in the distributed sense).
∇f	Gradient of f in the classical sense.
$div(f)$	Divergence operator:$div(f) = \sum_{i=1}^{N} \frac{\partial f}{\partial x_i}$.
$\nabla^2 f$	Hessian matrix of f in the classical sense: $(\nabla^2 f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.
$\frac{1}{\Omega}\int_\Omega f dx$	Mean value of over Ω.
$SNR(I_1/I_2)$	Estimate the quality of an image I_2 with respect to a reference image I_1.
$\alpha \vee s \wedge \beta$	Truncates function equal to α if $s \leq \alpha, \beta$ if $s \geq \beta$, s otherwise.
$sign(s)$	Sign function equal to 1 if $s > 0$, 0 if $s = 0$, and -1, if $s < 0$.
$P(\cdot)$	Probability mass.
$p(\cdot)$	Probability density.
$P(a, b)$	The joint probability-that is, the probability of having both a and b.
$Pr[\cdot]$	The probability of a condition being met.
$p(x\|\theta)$	The conditional probability density of x given θ.
w	Weight vector.
$\lambda(.,.)$	Loss function.
θ^*_{ML}	Maximum-likelihood estimate of θ.
$\hat{\theta}_{MAP}$	Maximum a posterior estimate of θ.

BV space

$\bar{\mathbb{R}}$	$\mathbb{R} \cup \{-\infty, +\infty\}$
$\mathbb{R}^+$	the set of non-negative real numbers, i.e.$[0, \infty)$
$\mathbb{Z}^+$	the set of positive integers, i.e. 1, 2, 3...
$\mathbb{N}$	the set of natural numbers, i.e. 0, 1, 2, 3,...

Let $n \in \mathbb{Z}^+$; $\alpha = (\alpha_1, ..., \alpha_n) \in \mathbb{N}^n$, $x = (x_1, ..., x_n) \in \mathbb{R}^n$:

$\|\alpha\|$	the modulus of $\alpha, \|\alpha\| = \alpha_1 + ... + \alpha_n$
∂_i	the partial derivative with respect to x_i,$\partial_i = \partial/\partial x_i$
D^α	$D^\alpha f = \partial_1^{\alpha 1}, ..., \partial_n^{\alpha_n} f$
Ω	a bounded domain in $\mathbb{R}^n$; with Lipschitz boundary
$\bar{\Omega}$	the closure of Ω in $\mathbb{R}^n$
Γ	the boundary of Ω, $\Gamma = \partial\Omega$
ν	the outward unit normal to Γ
$C^\infty(\mathbb{R}^n), C^\infty(\Omega)$	the space of infinitely differentiable functions
$C^k(\mathbb{R}^n), C^k(\Omega)$	the space of differentiable functions whose partial derivatives up to order kth are continuous
$C_c^k(\mathbb{R}^n), C_c^k(\Omega)$	the space of functions in $C^k(\mathbb{R}^n), C^k(\Omega)$, with compact support
$\mathcal{D}(\mathbb{R}^n), \mathcal{D}(\Omega)$	the subspace of functions in $C^\infty(\mathbb{R}^n), C^\infty(\Omega)$, with compact support in $\mathbb{R}^n, \Omega$; also denoted $C^\infty(\mathbb{R}^n), C^\infty(\Omega)$ $\alpha \in \mathbb{N}$(also called the Schwartz space)
$\mathcal{D}(\mathbb{R}^n)^n$	the space of n-tuples $(\varphi_1, ..., \varphi_n)$ where $\varphi_i \in \mathcal{D}(\mathbb{R}^n)$ for $i = 1, ..., n$
$\mathcal{D}(\bar{\Omega})$	the space consisting of $\varphi\|_\Omega$ for all $\varphi \in \mathcal{D}(\mathbb{R}^n)$
$\mathcal{D}(\bar{\Omega})^n$	the space consisting of $\varphi\|_\Omega$ for all $\varphi \in \mathcal{D}(\mathbb{R}^n)^n$
$BV(\Omega), BV(\mathbb{R}^n)$	the space of functions of bounded variations on $\Omega, \mathbb{R}^n$;
$SBV(\Omega), SBV(\mathbb{R}^n)$	the space of special functions of bounded variations on $\Omega, \mathbb{R}^n$
$X', or X^*$	the dual of the space X
$S'(\mathbb{R}^n)$	the space of tempered distributions on $\mathbb{R}^n$ i.e. the set of continuous linear functionals on $S(\mathbb{R}^n)$;
$\mathcal{D}'(\Omega)$	the space of distributions on Ω, i.e. the set of continuous linear functionals on $\mathcal{D}(\Omega)$
$\mathcal{M}(\Omega)$	the space of Radon measures on Ω
Df	Distributed derivative $Df = (D_1 f, ..., d_n f)$ in bounded total variational space.
C_f	Cantor part of Df with respect to Lebesgue measure in the BV space

Let $m \in \mathbb{N}, p \geq 1$ and $s \in \mathbb{R}$:

$L^p_{loc} = L^p_{loc}(\mathbb{R})$, or $L^p_{loc}(\Omega)$	the space of classes of measurable functions on $\mathbb{R}^n$, or Ω, such that $\|f(x)\|^p$ is locally integrable
$L^p = L^p(\mathbb{R}^n)$, or $L^p(\Omega)$	the space of classes of measurable functions on $\mathbb{R}^n$, or Ω, such that $\|f(x)\|^p$ is integrable
$L^\infty = L^\infty(\mathbb{R}^n)$, or $L^\infty(\Omega)$	the space of classes of measurable functions on $\mathbb{R}^n$, or Ω, such that $\|f(x)\|$ is essentially bounded
$W^{m,p} = W^{m,p}(\mathbb{R}^n)$, or $W^{m,p}(\Omega)$	Sobolev space consisting of all $f \in L^p$ (resp. $f \in L^p(\Omega)$) such that $D^\alpha f \in L^p$ (resp. $L^p(\Omega)$), $\forall \alpha \in \mathbb{N}^n, \|\alpha\| \leq m$
$W_0^{m,p}(\Omega)$	the closure of $\mathcal{D}(\Omega)$ in $W^{m,p}(\Omega)$;
$H^m = H^m(\mathbb{R}^n)$ or $H^m(\Omega)$	$W^{m,2}, (resp. W^{m,2}(\Omega))$
$H_0^m(\Omega)$	the closure of $\mathcal{D}(\Omega)$ in $H^m(\Omega)$
$H^{-m}(\Omega)$	the dual space of $H_0^m(\Omega)$
$H^s = H^s(\mathbb{R}^n)$	the Sobolev space of functions or distributions on $\mathbb{R}^n$

$H^s_{loc}(\mathbb{R}^n)$	the space $\{f \in S'(\mathbb{R}^n)$ s.t. $f\varphi \in H^s, \forall \varphi \in S(\mathbb{R}^n)\}$
$\lvert \cdot \rvert_X$	the semi-norm in the space X
$\lVert \cdot \rVert_X$	the norm in the space X
$\lVert \cdot \rVert_p$	the norm on L^p
$\lVert \cdot \rVert_\infty$	the norm on L^∞, i.e. essential sup norm
$\lVert \cdot \rVert_{-s}$	the norm in the space H^{-s}
$\mathcal{L}^n$	the n-dimensional Lebesgue measure
$\lvert\Omega\rvert$	the volume of Ω in $\mathbb{R}^n$, $\lvert\Omega\rvert = \mathcal{L}^n(\Omega)$
$\mathcal{H}^{N-1}$	the (N-1)-dimensional Hausdorff measure