Document Enhancement Using Text Specific Prior

Thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science (by Research)

in

Computer Science

by

Jyotirmoy Banerjee
200507006
jyotirmoy@research.iiit.ac.in

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA

Feb 2009

International Institute of Information Technology

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Document Restoration using Text Specific Prior" by Mr. Jyotirmoy Banerjee, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Advisor: Dr. C.V. Jawahar

To my mother

Abstract

Document images are often obtained by digitizing paper documents like books or manuscripts. They could be poor in appearance due to degradation of paper quality, spreading and flaking of ink toner, imaging artifacts etc. All the above phenomena lead to different types of noise at the word level including boundary erosion, dilation, cuts/breaks and merges of characters. Further, with the advent of modern electronic gadgets like PDAs, cellular phones, and digital cameras, the scope of document imaging has widened. Document image analysis systems are becoming increasingly visible in everyday life. For instance, one may be interested in systems that process, store, understand document images obtained by cellular phones. Processing challenges in this class of documents are considerably different from the conventional scanned document images. Many of this new class of documents are characterized by low resolution and poor quality. Super resolution provides an algorithmic solution to the resolution enhancement problem by exploiting the image-specific apriori information. In this thesis we study and propose new methods for restoration and resolution enhancement of document images.

We present a single image super-resolution algorithm for gray level document images without using any training set. Super-resolution of document images is characterized by bimodality, smoothness along the edges as well as subsampling consistency. These characteristics are enforced in a Markov Random Field (MRF) framework by defining an appropriate energy function. In our case, subsampling of super-resolution image will return the original low-resolution one, proving the correctness of the method. The restored image, is generated by iteratively reducing the energy function of the MRF, which is a nonlinear optimization problem. This approach is a single frame approach and is useful when you do not have multiple low-resolution images.

Document images have repetitive structural nature as the characters and words are found more than once in a page/book. The extraction of a single high-quality text image from a set of degraded images is benefited from the apriori information. A character segmentation is performed to extract the characters. A total variation based prior model is used in a Maximum A

Posteriori (MAP) estimate, to smoothen the edges and preserve the corners, so characteristic of text images. Dependence on character segmentation still remains a bottle-neck. Character segmentation problem is not a completely solved problem. The segmentation accuracy depends on the quantity of noise in the text image. In our next approach, we shall overcome the dependency on character segmentation. We shall look for a restoration approach that does not perform a explicit character segmentation, but still uses the repetitive component nature of document images.

In document images degradation is varied at different places in a document. Context plays an important role in textual image understanding. A MRF framework that exploits the contextual relation between image patches, is proposed. Using the topological/spacial constraints between the image patches, the impossible combinations are eliminated from the initial set of matchings, resulting in an unambiguous textual output. The local consistency is adjusted to the global consistency using the belief propagation algorithm. As we are working with patches and not characters, we avoid performing an explicit segmentation. The ability to work with larger patch sizes allows us to deal with severe degradations including cuts, blobs, merges and vandalized documents. This approach can also integrate document restoration and super-resolution into a single framework, thus directly generating high quality images from degraded documents.

To conclude, the thesis presents an approach for reconstructing document images. Unlike other conventional reconstruction methods, the unknown pixel values are not estimated based on their local surrounding neighbourhood, but on the whole image. We exploit the multiple occurrence of characters in the scanned document. A great advantage of our proposed approach over conventional approaches is that we have more information at our disposal, which leads to a better enhancement of the document image. Experimental results show significant improvement in image quality on document images collected from various sources including magazines and books, comprehensively demonstrate the robustness and adaptability of the approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Document image analysis has carved a niche out of the more general problem of computer vision because of its distinctness from regular class of images. Optical character recognition (OCR) was taken as one of the first clear applications of pattern recognition. Even today, the challenges of complex content, noisy data, and use of new imaging devices keep the field active. It is increasingly becoming important to provide people with regular and effective access to the information. Document images are information rich. Computer systems are used to develop the digital technology systems, which enables easy access to the vast reservoir of information. These system have an OCR at their core. Modern OCRs donot perform well in the case where the document image is substantially degraded. Adequate enhancement approaches are required to make the document images fit for OCRing. Further, the degraded image are not aesthetically appealing. These images are all departure from an ideal version of the document image, which is unambiguously well defined in the domain of machine-printed textual documents. The goal of this thesis is to revert back the degradation process and reach the ideal version of the document image.

The ultimate objective of the document image analysis is to recognize the text components in images of documents, and to extract the intended information as a human would. With the advent of modern publishing technologies, document analysis systems will become increasingly more evident in the form of everyday document systems.

1.2 Motivation

Images of paper documents are almost inevitably degraded in the course of printing, photocopying, Faxing, and scanning, and this loss of quality - even when it appears negligible to human eyes - can be responsible for an abrupt decline in accuracy by the current generation of text recognition (OCR) systems. This fragility of OCR systems when confronted by low image quality is well known to the OCR community [80]. The accuracy of today's document recognition algorithms falls abruptly when image quality degrades even slightly [3]. The physical causes of image degradation are myriad: spreading and flaking of ink toner; uneven paper surface; low print contrast; non-uniform illumination; defocusing; finite spatial sampling rate; variations in pixel sensor sensitivity and placement; noise in electronic components; binarization (e.g. fixed and adaptive thresholding). And, images may result from more than one stage of printing and imaging. By "degradation" (or "defects") we mean a wide variety of less-than ideal properties of real images.

Traditionally, document images are scanned from pseudo binary hardcopy paper manuscripts with a flatbed, sheet-fed, or mounted imaging device. Recently, however, the community has seen an increased interest in adapting modern imaging device like digital cameras to tasks related to document image analysis. Digital camcorders, digital cameras, PCcams, PDAs, and even cellphone cameras are becoming increasingly popular, and they have shown potential as alternative imaging devices [47]. Although they cannot replace scanners, they are small, light, easily integrated with various networks, and more suitable for many document capturing tasks in less constrained environments. These advantages lead to a natural extension of the document processing community where cameras are used to image hardcopy documents or natural scenes containing textual content. This has given rise to new potential applications, though most of the time handicapped by low-resolution. For text and document analysis, as the application areas extend to lower resolution camera enabled devices, super-resolution methods are becoming more important and necessary. Digital video compression algorithms can benefit from successful text resolution expansion techniques. Video could be indexed and retrieved based on text information, text observed in these types of images is often low-resolution. In these con-

ditions, it is virtually impossible to do character segmentation independently from recognition. Resolution enhancement is one of the approach that can assist the cause of recognition in low-resolution images. Super-resolution methods are useful where physical limitations exist preventing higher resolution images from being obtained. Whenever dynamic image enlargement is needed, such as text in camera-based imagery, super-resolution techniques can be utilized.

1.3 Text Enhancement

Image processing modifies pictures to improve them (enhancement, restoration), extract information (analysis, recognition), and change their structure (composition, image editing). Images can be processed by optical, photographic, and electronic means, but image processing using digital computers is the most common method because digital methods are fast, flexible, and precise.

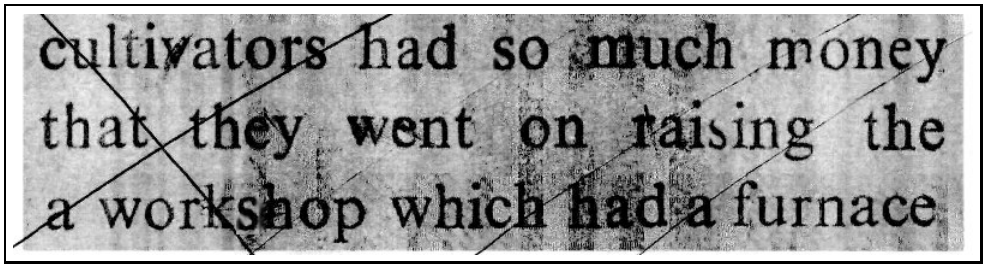
Image enhancement improves the quality (clarity) of images for human viewing. Removing blurring and noise, increasing contrast, and revealing details are examples of enhancement operations. For example, an image might be taken of an monument, which might be of low contrast and somewhat blurred. Reducing the noise and blurring and increasing the contrast range could enhance the image. The original image might have areas of very high and very low intensity, which mask details. An enhancement algorithm reveals these details. Adaptive algorithms adjust their operation based on the image information (pixels) being processed. In this case the mean intensity, contrast, and sharpness (amount of blur removal) could be adjusted based on the pixel intensity statistics in various areas of the image.

The aim of image enhancement is to improve the interpretability or perception of information in images for human viewers, or to provide 'better' input for other automated image processing techniques.

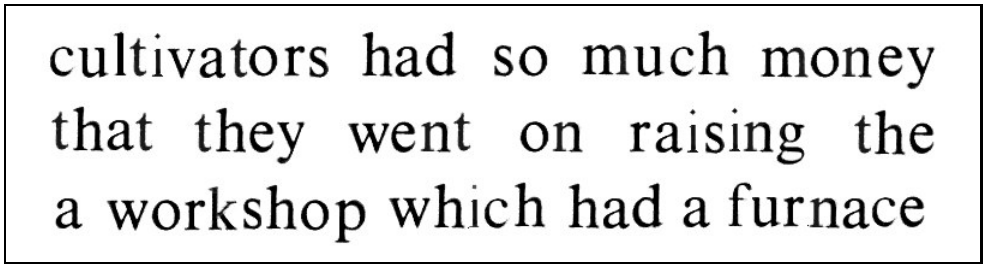
Image enhancement techniques can be divided into two broad categories:

- Spatial domain methods, which operate directly on pixels, and
- frequency domain methods, which operate on the Fourier transform of an image.

Unfortunately, there is no general theory for determining what is 'good' image enhancement when it comes to human perception. If it looks good, it is good! However, when image enhancement techniques are used as pre-processing tools for other image processing techniques, then quantitative measures can determine which techniques are most appropriate. Traditional methods for image enhancement can be classified into two categories: image restoration, and resolution expansion.



(a) Portion of a degraded text image

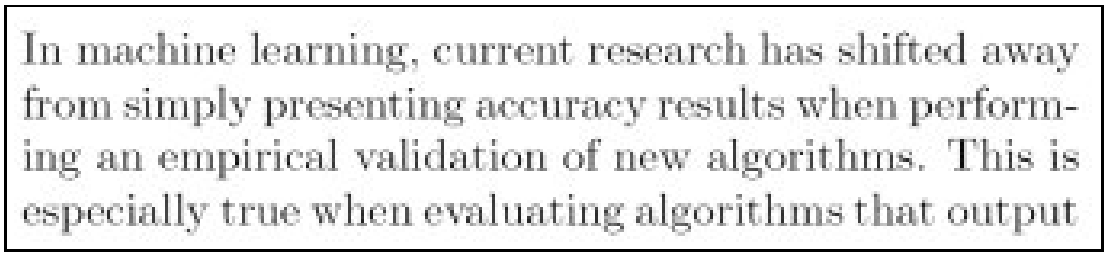


(b) Restoration output image

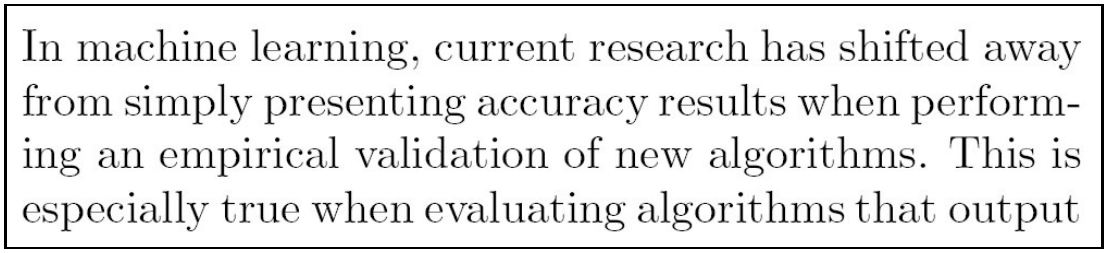
Figure 1.1 Restoration of document images.

Our Interest in this thesis - In this thesis we are interested in the problem image restoration and resolution expansion of text images. Our algorithm deals with only the textual part of a document image. In case there are graphic object in the document page, then a suitable segmentation algorithm should be used to separate the textual content from the non textual one.

- **Text Restoration -** Document images are often obtained by digitizing paper documents like books or manuscripts. They could be poor in appearance due to degradation of paper quality, spreading and flaking of ink toner, imaging artifacts etc as shown in Figure 1.1.



(a) Portion of a low-resolution text image



(b) Super-resolution output image

Figure 1.2 Restoration of document images.

Restoration of such images has many applications in enhancing the performance of character recognizers as well as in book readers used in digital libraries.

- **Text Super-resolution -** The goal of resolution expansion is to create an expanded image with improved definition from observed low-resolution imagery. Acquisition of this low-resolution imagery can be modeled by averaging a block of pixels within a high-resolution image. The image acquisition process consists of converting a continuous image into discrete values obtained from a group of sensor elements. Each sensor element produces a value which is a function of the amount of light incident on the device. For 8-bit grayscale quantization, the allowable range of values for each sensor are integers from 0 (black) to 255 (white). The sensors are typically arranged in a non-overlapping grid of square elements, smaller elements result in higher resolution imagery. A high-resolution image is shown in Figure 1.2 where the number of sensors is adequate to represent the desired text image. The majority of pixels within the image are either white or black, with a small number of gray pixels occurring at the edges. Figure 1.2 illustrates a low-resolution image where the number of sensors has been reduced by a factor of $q = 4$

in both the horizontal and vertical directions. This low-resolution acquisition results in significant blockiness and is insufficient to accurately represent this image. Each sensor element effectively averages the image within its section of the grid, resulting in an increased amount of gray pixels.

1.4 Image Enhancement as an Inverse Problem : MAP

In an image enhancement problem, we assume that an ideal image, $\mathbf{f}$, has been corrupted to create the measured image, $\mathbf{g}$. The usual model for the corruption is a distortion operation, denoted by D, followed by the addition of random noise

$$\mathbf{g} = D(\mathbf{f}) + \mathbf{n} \tag{1.1}$$

where $\mathbf{g} = [g_1, \ldots, g_N]$ and g_i denotes the i^{th} pixel in a column vector representation of the image $\mathbf{g}$. Here, $\mathbf{f}$ and $\mathbf{n}$ are also similarly represented. The restoration problem then, is the problem of finding the best estimate of $\mathbf{f}$ given the measurement, $\mathbf{g}$, some knowledge of the distortion (e.g. blur), and the statistics of noise.

Restoration is often referred to as an inverse problem. That is, we have a process (in this case blur) which takes an input and produces an output. We can only measure the output, and we wish to infer the input.

Inverse problems and ill-posedness - A problem $\mathbf{g} = D(\mathbf{f})$ is said to be well-posed if

- for each $\mathbf{f}$, a solution, $\mathbf{g}$, exists
- the solution is unique
- the solution $\mathbf{g}$ continuously depends on the data $\mathbf{f}$.

If these three conditions do not all hold, the problem is said to be "ill-posed". Ill-posedness is normally caused by the ill-conditioning of the problem. Conditioning of a mathematical problem is measured by the sensitivity of output to changes in input. For a well-conditioned problem, a small change of input does not affect the output much; while for an ill-conditioned problem, a small change of input can change the output a great deal.

A simple example of ill-conditioning is as follows: consider the linear system described by a blur $\mathbf{A}$, and unknown image f, and a measurement g, where

$$\mathbf{g} = \mathbf{A}\mathbf{f}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1.01 \end{bmatrix} \qquad \mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \qquad \mathbf{g} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

This system has solution $f_1 = 1, f_2 = 0$. Now, suppose the measurement, $\mathbf{g}$, is corrupted by noise, producing $\mathbf{g} = [1\ 1.01]^T$. Then, the solution is $f_1 = 0, f_2 = 1$. A trivially small change in the measured data causes a dramatic change in the solution. Thus in all such situations, the vector $\mathbf{f} = \mathbf{A^{-1}g}$ (or in the full ranked overdetermined case $\mathbf{A}^+\mathbf{g}$, with the pseudo inverse $\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$), if it exists at all, is usually a poor approximation of $\mathbf{f}$ (This can be seen from an analysis in terms of the singular value decomposition [71]).

Importance of well-posedness has been noted long before the dawn of the computer age by Maxwell who in 1873 wrote [6]:

> *There are certain classes of phenomena, as I have said, in which a small error in the data only introduces a small error in the result. Such are, among others, the larger phenomena of the solar system, and those in which the more elementary laws in dynamics contribute the greater part of the result. The course of events in these cases is stable.*

There are many ways to approach these ill-posed restoration problems. They all share a common structure: the *regularization theory*. Generally speaking, any regularization method tries to analyze a related well-posed problem whose solution approximates the original ill-posed problem.

For example, the first approach one might think of is to produce an image estimate which has the minimum linear least squares error. That is, find the unknown image $\mathbf{f}$ which minimizes

$$E = ||\mathbf{g} - \mathbf{A}\mathbf{f}||^2$$

However, the matrix $\mathbf{A}$ may be ill-conditioned or singular yielding a large number of solutions. Directly minimizing E does not work, as the problem is still ill-conditioned. In order to give

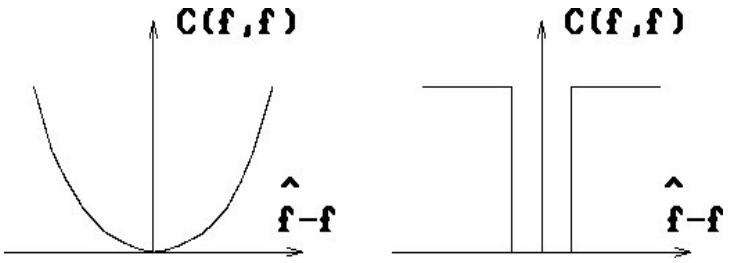


Figure 1.3 Two choices of cost functions.

preference to a particular solution with desirable properties, the regularization term is included in this minimization. A regularization term or a prior is indeed needed to derive a solution to an ill-posed problem. The Maximum A Posteriori (MAP) approach, is one such framework where the prior is used to derive the solution.

The Maximum A Posteriori (MAP) approach - Bayes image reasoning is a theory of fundamental importance in estimation and decision making. According to this theory, when both the prior distribution and the likelihood function of a pattern are known, the best that can be estimated from these sources of knowledge is the Bayes labeling. The maximum a posterior (MAP) solution, as a special case in the Bayes framework, is sought in many vision works.

Bayes Estimation - In Bayes estimation, a risk is minimized to obtain the optimal estimate. The Bayes risk of estimate f^* is defined as

$$R(f^*) = \int_{f \in F} C(f^*, f) P(f|d) df$$

where d is the observation, $C(f^*, f)$ is a cost function and $P(f|d)$ is the posterior distribution. First of all, we need to compute the posterior distribution from the prior and the likelihood. According to the Bayes rule, the posterior probability can be computed by using the following formulation

$$P(f|d) = \frac{p(d|f)P(f)}{p(d)}$$

where $P(f)$ is the prior probability of labellings f, $p(d|f)$ is the conditional p.d.f. of the observations d, also called the likelihood function of f for d fixed, and $p(d)$ is the density of d which is a constant when d is given.

The cost function $C(f^*, f)$ determines the cost of estimate f when the truth is f^*. It is defined according to our preference. Two popular choices are the quadratic cost function

$$C(f^*, f) = ||f^* - f||^2$$

where $||a - b||$ is a distance between a and b, and the δ cost function

$$C(f^*, f) = \begin{cases} 0 & \text{if } ||f^* - f|| \leq \delta \\ 1 & 1 \text{ otherwise} \end{cases}$$

where $\delta > 0$ is any small constant. A plot of the two cost functions are shown in Figure 1.3. The Bayes risk under the quadratic cost function measures the variance of the estimate

$$R(f^*) = \int_{f \in F} ||f^* - f||^2 P(f|d) df$$

Letting $\frac{\partial R(f^*)}{\partial f} = 0$, we obtain the minimal variance estimate

$$f^* = \int_{f \in F} f P(f|d) df$$

The above is the mean of the posterior probability.

For the δ cost function, the Bayes risk is

$$R(f^*) = \int_{||f^* - f|| > \delta} P(f|d) df = 1 - \int_{||f^* - f|| \leq \delta} P(f|d) df$$

When $\delta \to 0$, the above is approximated by

$$R(f^*) = 1 - \kappa P(f|d)$$

where κ is the volume of the space containing all points f for which $||f^* - f|| \leq \delta$. Minimizing the above is equivalent to maximizing the posterior probability. Therefore, the minimal risk estimate is

$$f^* = \underset{f \in F}{\operatorname{argmax}} P(f|d)$$

which is known as the MAP estimate. Because $p(d)$ is a constant for a fixed d, $P(f|d)$ is proportional to the joint distribution

$$P(f|d) \propto P(f, d) = P(d|f) P(f)$$

Then the MAP estimate is equivalently found by

$$f^* = \underset{f \in F}{\operatorname{argmax}} \{P(d|f)P(f)\}$$

Obviously, when the prior distribution, $P(f)$, is flat, the MAP is equivalent to the maximum likelihood. Hence prior plays a important role in the enhancement process.

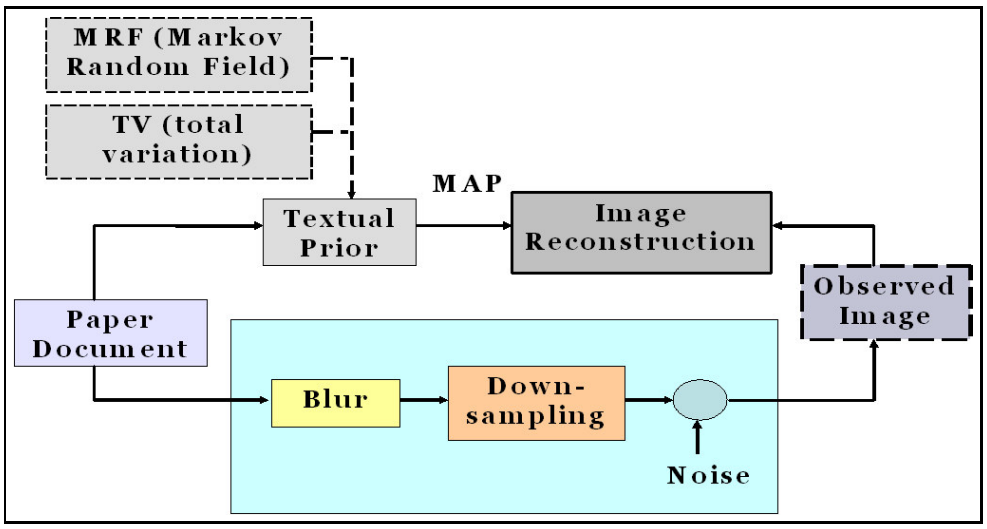


Figure 1.4 Observation model of the document image acquisition

1.5 Design of Prior

Restoration from a still image is a well recognized example of an ill-posed inverse problem. Such problems may be approached using regularization methods which constrain the feasible solution space by employing a-priori knowledge. This may be achieved in two complimentary ways; (1) obtain additional novel observation data and (2) constrain the feasible solution space with a-priori assumptions on the form of the solution. Both techniques feature in modern restoration methods which utilize (1) image sequences which provide additional spatio-temporal observation constraints (typically in the form of novel data arising from sub-pixel motion) and (2) various a-priori constraints on the degraded image (e.g. local smoothness, edge preservation, positivity, energy boundedness, etc.). The use of non-linear a-priori

constraints provides the potential for bandwidth extension beyond the diffraction limit of the optical system.

It remains however to compute the solution to the ill-posed enhancement inverse problem. Amongst the numerous solution techniques featuring in the literature, the Bayesian Maximum A Posteriori (MAP) estimation method, is promising. MAP estimation provides a rigorous theoretical framework, several desirable mathematical properties and makes explicit use of a-priori information in the form of a prior probability density on the solution image.

The relationship between the paper document, the observed inferior image and the desired enhanced/restored image is once illustrated in Figure 1.4. In this thesis, we formulate the text prior using two mathematical framework : 1) Total Variation (TV) 2) Markov Random Field (MRF). The TV formulation provides an edge preserving smoothness prior. Since their introduction in a classic paper by Rudin, Osher and Fatemi [81], total variation minimizing models have become one of the most popular and successful methodology for image restoration. More recently, there has been a resurgence of interest and exciting new developments, some extending the applicability to inpainting, blind deconvolution and vector-valued images, while others offer improvements in better preservation of contrast, geometry and textures. The spatial property can also be modeled through different aspects, among which, the contextual constraint is a general and powerful one. MRF theory provides a convenient and consistent way to model context-dependent entities such as image pixels and correlated features. In the next chapter, we shall discuss in detail about these mathematical frameworks. In this thesis we shall see how to construct document specific prior and use them for our enhancement.

1.6 Observations

Document images are a distinct class of images widely different from natural images. The problem of document restoration and super-resolution is a special case of image restoration because

- document images are pseudo binary in nature,

- the regularity of the patterns used in this "visual" language distinguishes the document images from natural scenes, and
- the relatively small size of the (character) image, makes them more susceptible to degradation.
- in a document image it is quite possible that the same character image at different physical location in a document may be degraded differently.

These factors, lead to an array of interesting observations, specific to this domain.

This thesis focuses on the issue of restoration and super-resolution of a document image using text specific prior information. The restoration of text is an ill-posed problem and one that is highly sensitive to the additional assumptions or information needed to establish its well-posedness. These assumptions are generally reflected in the priors that are imposed in the formulation. Generic smoothness constraint tend to smooth over the important details and produce improper restoration.

A successful document restoration and super-resolution algorithm needs to use the text-specific *apriori* information. Further, a mathematical framework is needed that incorporate the prior information and handles the text related uncertainties. We exploit the properties of document images to develop a specific restoration technique, specially suited for the same.

1.7 Related Work

Related articles to this thesis are referred and discussed in detail in the relevant parts of the next four chapters. However, a brief sketch of the related work is provided to give the background of the thesis.

1.7.1 Related Work in Image Restoration

- Image Denoising has remained a fundamental problem in the field of image processing [70, 62]. Spatial filtering for image denoising works only for additive noise. Median filters, mean filters, max and min filter and various spatially adaptive versions [34] are

commonly used. The simplest method for noise removal is Gaussian filtering, which is equivalent to solving anisotropic heat diffusion equation [89], a second-order linear PDE. To keep sharp edges, anisotropic diffusion can be performed [77], wavelets give a superior performance in image denoising due to properties such as sparsity and multiresolution structure. With wavelet transform gaining popularity in the last two decades various algorithms for denoising in wavelet domain were introduced. The focus was shifted from the Spatial and Fourier domain to the wavelet transform domain. Ever since Donoho's wavelet based thresholding approach was published in 1995 [23, 66], there was a surge in the denoising papers being published.

1.7.2 Related Work in Document Restoration

- **Text Restoration - Filter based Approaches -** Filter based approaches were widely used in general imagery. There are few works, where these techniques are applied on document images. In Stubberud et al. [90], by using the output from an OCR system and a distorted text image, their technique trains an adaptive restoration filter and then applies the filter to the distorted text image that the OCR system could not recognize. Ramponi et al. [79] have used quadratic filters to enhance the document. Fan et al. [26] propose to exploit the spatial correlations between wavelet coefficients by replacing the thresholding process with a diffusion process.
- **Text Restoration and Enhancement -** A border following algorithm is used in [101] to reconstruct the borders and missing links of noisy and broken handwritten digits. Shi et al. [87] performs selective and adaptive stroke filling with a neighborhood operator which emphasizes stroke connectivity. Allier et al. [2] proposed a method for accurate character reconstruction based on the active contour model. Some of the restoration efforts are based on morphological filters [60, 103] where the size of the morphological filter directly depends on the font size. Some other methods [1, 5] use model based approaches. A variety of methods have been proposed in order to improve contrast within text images. They include methods based on multi-resolution pyramid with fuzzy edge

detectors [85], and a mixed approach using topological features and contour beautification [73]. Resolution expansion is also attempted using text bitmap averaging [39]. This method depends on the segmentation and then clustering of character images which is often hard to obtain. Combination of interpolation and binarization [55] is also used to improve quality of text in images.

- **Text Restoration in Historical Documents -** The general problem in historical documents is the "ink bleed-through" problem. There are many non-blind approaches, mainly based on the comparison between the front and back page which requires a registration of the two sides of the document of the two sides of the document in order to identify the interfering strokes to be eliminated. Techniques. Sharma's approach [86] simplifies the physical model of these effects to derive a linear mathematical model and then defines an adaptive linear-filtering scheme. Approach proposed by Dubois and Pathak [25] is mainly based on processing both sides of a gray-level manuscript simultaneously using a six parameter affine transformation to register the two sides. In [92], a wavelet reconstruction process is applied to iteratively enhance the foreground strokes and smear the interfering strokes. A blind restoration approach i.e, it does not need of the both sides of the document, is generally based on steered filters. An approach proposed by Wang et al. [100, 99] uses directional wavelets to remove images of interfering strokes. Other more flexible techniques exist, among which, we can cite techniques based on Independent Component Analysis [95], adaptive binarization [32], self-organizing maps [88], color analysis [52]. Drira's [24] approach consists of combining both Principal Component Analysis (PCA) and K-means. These techniques are applied recursively to separate original text from interfering and overlapping areas of text.
- **Text Enhancement in Video -** Li et al. [55] applied Shannon interpolation method to increase image resolution and Niblack's adaptive thresholding [72] to binarize the image with complex backgrounds. In [53], Li et al. use multi-frame integration to enhance captions in video. The influence of the background is reduced on the basis of motion clues. Sato [84] enhances the text on the basis of its sub-structure: line element, by using filters

with four orientations: vertical, horizontal, left diagonal and right diagonal in the located text block. Asymmetric Gabor filters have been proposed by Chen et al. [18], which can efficiently extract the orientation and scale of the stripes present in a video image. This information is used to enhance contrast at only those edges most likely to represent text. In Kwak et al. [49], after the multiple video text frames containing the same captions are detected and the caption area in each frame is extracted, five different image enhancement techniques are serially applied to the image: multi-frame integration, resolution enhancement, contrast enhancement, advanced binarization, and morphological smoothing operations.

1.7.3 Related Work in Image Super-resolution

- Super-resolution techniques may be divided into two main classes; frequency domain and spatial domain. All frequency domain approaches are, to a greater or lesser extent, unable to accommodate general scene observation models including spatially varying degradations, non-global relative camera/scene motion, general a-priori constraints or general noise models. Spatial domain formulations can accommodate all these and provide enormous flexibility in the range of degradations and observation models which may be represented and are thus the methods of choice. Spatial domain observation models facilitate inclusion of additional data in the observation equation with the effect of reducing the feasible solution space.

  Tipping et al. developed a Bayesian treatment of the super-resolution problem in which the likelihood function for the image registration parameters is based on a marginalization over the unknown high-resolution image [94]. A texture based approach is provided in Pickup et al. [78] where a domain-specific image prior in the form of a p.d.f. based upon sampled images. Single image interpolation algorithms which use a database of training images to create plausible high-frequency details in zoomed images is proposed by Freeman et al. [31]. A comprehensive review with directions for future research can be found in [8]. Limits of super-resolution are discussed in [4, 61].

1.7.4 Related Work in Document Super-resolution

- **Multi-frame Approaches -** Super-resolution is the process of simulating a high-resolution, high-quality camera from blurred, noisy images captured using a low-resolution camera. SR algorithms are divided into two categories *viz.* multi-frame and learning based single-image super-resolution. Li and Doermann [54] used the method of projection onto convex sets (POCS), to deblur scene text in video sequences. In a parallel work, Capel and Zisserman [14] used a projective transform motion model for super-resolution of text specifically for image sequences in which the point-to-point image transformation was of enough complexity to demand such consideration. In a recent work, Teager filter (a quadratic unsharp masking filter) was adopted by Mancas-Thillou and Mirmehdi [67] for the extraction of high frequencies thus enhancing character edges. Donaldson and Myers [22] proposed a text specific prior model, which modeled the bimodality and the local smoothness with step discontinuity. They use the Gibbs prior with a Huber gradient penalty function as their smoothness function. This piecewise smoothness prior is good at reducing false speckles in the results, but it undermines the importance of enhancing edges. Dalley *et al.* [20] employed a training-based method, in a Bayesian framework. A database is built that indicates which high-resolution patch should be output given an input low-resolution patch. Park et al. [74] developed an alternative approach, an edge-based super-resolution technique. It attempts to locate the edges to subpixel accuracy in a sequence of images taken from training examples, and then fuses the conglomerated edge information into the super-resolved image using a MRF formulation.
- **Single-frame Approaches -** Thouin and Thouin et al. [93] used nonlinear optimization on a gray scale input image to minimize a Bimodal Smoothness Average (BSA) score.

1.8 Contributions

In this thesis we have proposed new methods for enhancement with focus on restoration and super-resolution of document images. In particular, we address the following:

- First, we present a method for restoration of document images, using a Maximum a Posteriori formulation. The advantage of our method is that the prior need not be learned from the training images. The extraction of a single high-quality enhanced text image from a set of degraded images can benefit from a strong prior knowledge, typical of text images. The restoration process should allow for discontinuities but at the same time discourage oscillations. These properties were represented in a total variation based prior model.
- Second, we formulate the text image restoration problem in a relaxation framework. Text images are very different from natural images. The regularity of the patterns used in this "visual" language distinguishes these pseudo binary document images from natural scenes. Context plays an important role in textual image understanding. A stochastic framework that exploits the contextual relation between image patches, is proposed in this paper. Using the topological/spacial constraints between the image patches, the impossible combinations are eliminated from the initial set of matchings, resulting in an unambiguous textual output. The local consistency is adjusted to the global consistency using the belief propagation algorithm.
- Lastly, we present an edge-directed, single image super-resolution algorithm for document images without using any training set. This technique creates an image with smooth regions in both the foreground and the background, while allowing sharp discontinuities across and smoothness along the edges. Our method preserves sharp corners in text images by using the local edge direction, which is computed first by evaluating the gradient field and then taking its tangent. Super-resolution of document images is characterized by bimodality, smoothness along the edges as well as subsampling consistency. These characteristics are enforced in a Markov Random Field (MRF) framework by defining an appropriate energy function. In our method, subsampling of super-resolution image will return the original lowresolution one, proving the correctness of the method. The super-resolution image, is generated by iteratively reducing this energy function.

1.9 Organization of the Thesis

This thesis focuses on, namely, restoration, and super-resolution framework of document images. So far in this chapter we have presented an overview of image restoration. We discussed in Section 1.3, that restoration is an ill-posed problem and how regularization helps in solving this ill-posed problem. It indirectly meant that for restoration, prior information should be incorporated. Then we discussed Maximum A Posteriori method to incorporate the prior information. In Section 1.5 we discussed the related work in the field of text restoration in the field of document restoration and super-resolution.

- In Chapter 2, we give an overview of the mathematical methods used in this thesis. In Section 2.2, the total variation based formulation for noise removal is explained. A iterative algorithm used to solve the same is discussed. In Section 2.3, we discuss the labeling problem and the markov random field. A number of concepts relating to the formulation of an energy function and its justification in a Bayesian framework is explained.
- In Chapter 3, we discuss our method on single image document super-resolution. Section 3.2 discuss the super-resolution in document images. In Section 3.3 we discuss the text specific prior information. The MRF based formulation for document Super-resolution is discussed in Section 3.4. Experimental results are shown in Section 3.5 and Conclusion in Section 3.6.
- In Chapter 4, we discuss our first method on document restoration using bayesian inference. In Section 4.2 we present some related work. Section 4.3 describes our text restoration algorithm in detail and Section 4.3.1 presents a discussion on the algorithm. We present experimental result in Section 4.4 and conclusion and future work in Section 4.5.
- In Chapter 5, we discuss our second method on document restoration using relaxation framework. In Section 5.2 we analyze how to restore by labeling. Section 5.3 describes the Markov network construction. Experimental results are shown in Section 5.4 and Conclusion in Section 5.5.
- In Chapter 6, we conclude and discuss future work.

Chapter 2

Preliminaries

2.1 Introduction

In this chapter, we discuss the mathematical preliminaries required in the chapters ahead. We had introduced the importance of MAP based formulation for the document restoration ill-possed problem. We had also discussed the use of prior information in the Maximum A Posteriori (MAP) formulation in the previous chapter. Various a-priori, constraints on the degraded image (e.g. local smoothness, edge preservation, positivity, energy boundedness, etc.). In this chapter we will now look at a natural way to incorporate the priori knowledge. Total variation and Markov Random Field (MRF) models help in designing the prior in the MAP formulation. We shall briefly discuss these mathematical frameworks in rest of the chapter.

2.2 Total Variation

Variational models have been extremely successful in a wide variety of restoration problems, and remain one of the most active areas of research in mathematical image processing and computer vision. By now, their scope encompasses not only the fundamental problem of image denoising, but also other restoration tasks such as deblurring, blind deconvolution, and inpainting. Variational models exhibit the solution of these problems as minimizers of appropriately chosen functionals. The minimization technique of choice for such models routinely

involves the solution of nonlinear partial differential equations (PDEs) derived as necessary optimality conditions.

Perhaps the most basic (fundamental) image restoration problem is denoising. It forms a significant preliminary step in many machine vision tasks, such as object detection and recognition. It is also one of the mathematically most intriguing problems in vision. A major concern in designing image denoising models is to preserve important image features, such as those most easily detected by the human visual system, while removing noise. One such important image feature are the edges; these are places in an image where there is a sharp change in image properties, which happens for instance at object boundaries. A great deal of research has gone into designing models for removing noise while preserving edges; recently there has also been a lot of effort in preserving other fine scale image features, such as texture. All successful denoising models take advantage of the fact that there is an inherent regularity found in natural images; this is how they attempt to tell apart noise and actual image information. Variational and PDE based models make it particularly easy to impose geometric regularity on the solutions obtained as denoised images, such as smoothness of boundaries. This is one of the main reasons behind their success.

2.2.1 Nonlinear total variation based noise removal

Total variation based image restoration models were first introduced by Rudin, Osher, and Fatemi (ROF) in their pioneering work [81] on edge preserving image denoising. It is one of the earliest and best known examples of PDE based edge preserving denoising. It was designed with the explicit goal of preserving sharp discontinuities (edges) in images while removing noise and other unwanted fine scale detail. Being convex, the ROF model is one of the simplest variational models having this most desirable property. The revolutionary aspect of this model is its regularization term that allows for discontinuities but at the same time disfavors oscillations. It was originally formulated in [81] for grayscale imagery in the following form:

$$\inf_{\int_\Omega (u-f)^2 dx = \sigma^2} \int_\Omega |\nabla \mathbf{x}| \tag{2.1}$$

Here, Ω denotes the image domain (for instance, the computer screen), and is usually a rectangle. The function $f(x) : \Omega \to \mathbb{R}$ represents the given observed image, which is assumed to be corrupted by Gaussian noise of variance σ^2. The constraint of the optimization forces the minimization to take place over images that are consistent with this known noise level. The objective functional itself is called the total variation (TV) of the function $u(x)$; for smooth images it is equivalent to the L^1 norm of the derivative, and hence is some measure of the amount of oscillation found in the function $u(x)$. Optimization problem in equation 2.1 is equivalent to the following unconstrained optimization, which was also first introduced in [81]:

$$\inf_{u \in L^2(\Omega)} \int_\Omega |\nabla x| + \int_\Omega \lambda (u - f)^2 dx \tag{2.2}$$

Here, $\lambda > 0$ is a Lagrange multiplier. The equivalence of problems 2.1 and 2.2 has been established in [16]. In the original ROF paper [81] there is an iterative numerical procedure given for choosing λ so that the solution $u(x)$ obtained solves 2.1.

Total variation based energies appear, and have been previously studied in, many different areas of pure and applied mathematics. For instance, the notion of total variation of a function and functions of bounded variation appear in the theory of minimal surfaces. In applied mathematics, total variation based models and analysis appear in more classical applications such as elasticity and fluid dynamics. Due to ROF, this notion has now become central also in image processing.

2.2.2 Numerical Method

There have been numerous numerical algorithms proposed for minimizing the ROF objective. Most of them fall into the three main approaches, namely, direct optimization, solving the associated Euler-Lagrange equations and using the dual variable explicitly in the solution process to overcome some computational difficulties encountered in the primal problem. We will focus on the second approach.

2.2.2.1 Artificial Time Marching and Fixed Point Iteration

In their original paper [81], Rudin et al. proposed the use of artificial time marching to solve the Euler-Lagrange equations which is equivalent to the steepest descent of the energy function. More precisely, consider the image as a function of space and time and seek the steady state of the equation

$$\frac{du}{dt} = \nabla.\left(\frac{\nabla u}{|\nabla u|_\beta}\right) - 2\lambda(u-f) \tag{2.3}$$

Here, $|\nabla u|_\beta = \sqrt{|\nabla u| + \beta^2}$ is a regularized version of $|\nabla u|$ to reduce degeneracies in flat regions where $|\nabla u| \approx 0$. In numerical implementation, an explicit time marching scheme with time step Δt and space step size Δx is used. Under this method, the objective value of the ROF model is guaranteed to be decreasing and the solution will tend to the unique minimizer as time increases. However, the convergence is usually slow due to the Courant-Friedrichs-Lewy (CFL) condition, $\Delta t \leq c\Delta x^2|\nabla u|$ for some constant $c > 0$ [68], imposed on the size of the time step, especially in flat regions where $|\nabla u| \approx 0$. CFL condition in numerical equation solving states that, given a space discretization, a time step bigger than some computable quantity should not be taken. The condition can be viewed as a sort of discrete "light cone" condition, namely that the time step must be kept small enough so that information has enough time to propagate through the space discretization.

To relax the CFL condition, Marquina and Osher use, in [68], a "preconditioning" technique to cancel singularities due to the degenerate diffusion coefficient $\frac{1}{|\nabla u|}$:

$$\frac{du}{dt} = |\nabla u|\left[\nabla.\left(\frac{\nabla u}{|\nabla u|_\beta}\right) - 2\lambda(u-f)\right] \tag{2.4}$$

which can also be viewed as mean curvature motion with a forcing term $-2\lambda(u-f)$. Explicit schemes suggested in [68] for solving the above equation improve the CFL to $\Delta t \leq c\Delta x^2|\nabla u|$ which is independent of $|\Delta u|$.

To completely get rid of CFL conditions, Vogel and Oman proposed in [98] a fixed point iteration scheme (FP) which solves the stationary Euler-Lagrange directly. The Euler-Lagrange equation is linearized by lagging the diffusion coefficient and thus the $(i+1)$-th iterate is

obtained by solving the sparse linear equation:

$$\nabla.\left(\frac{\nabla u^{i+1}}{|\nabla u^{i}|_{\beta}}\right) - \lambda(u^{i+1} - f) = 0 \tag{2.5}$$

While this method converges only linearly, empirically, only a few iterations are needed to achieve visual accuracy. In practice, one typically employs specifically designed fast solvers to solve equation 2.5 in each iteration.

2.3 Markov Random Fields

The field of computer vision is related to the task of obtaining relevant information about the real world by inferring the images of that world. Typically the task becomes difficult owing to the uncertainties in the imaging process and the ambiguities in the inference of the real world. This in turn leads to multiple solutions to a particular vision problem. An optimization approach provides an elegant technique to reduce the number of possible solutions by formulating various constraints on the problem at hand. The optimization approach consists of two major steps described as follows.

The first step is the formulation of an *objective* function. It is a function from the set of all possible solutions to real numbers. In order to formulate an objective function it is important to impose a set of constraints which should be satisfied by the final solution. The solution to an objective function which satisfies these set of constraints in the best possible manner is the desired solution. Thus, the value of the real number to which the objective function is mapped gives the measure of goodness of that solution. Conventionally, the lesser the value, the better the solution is. Two of the most commonly used constraints to formulate an objective function for a vision problem are obtained by the input data which could be an image for example and the prior knowledge about this data. The data constraint restricts the desired solution to be close to the observed data and the prior constraint confines the desired solution to have the form which is agreeable with the prior knowledge about the problem. The objective function thus formulated and containing the two constraints is referred to as an *energy function*. The data constraint is defined specific to the vision problem being solved. The prior constraint is usually

imposed by the assumption that the variables of the objective function belong to a Markov Random Field (MRF). The concept of MRFs is explained in the next section. Prior to this in Section 2.3.1, we explain the concept of Labeling in vision which is a natural representation for the study of MRFs and is imperative to understand the optimization framework for various problems in computer vision.

The second step of the optimization approach is to minimize the energy function by finding the global minima. An energy function in computer vision is typically not convex and they have multiple local minima. This makes the task of global minimization difficult. Additionally, the energy function for an image has a large number of unknowns, which makes the computational requirements for minimization high. In fact its an NP-hard problem to find the exact minima. This leads finding approximate solutions which are closer to the global minima. One of the assumption which relaxes the optimization approach to some extent is that the set of solutions is finite. This is done by discretizing the variables which are used to formulate the energy function. This makes the set of solutions countable but still too large to be explored completely. Such an optimization problem where the input solution set is countable is combinatorial in nature and is called as a discrete optimization problem. It can be shown that minimizing such an optimization function in computer vision will indeed lead to the optimal solution by using a Bayesian perspective (Maximum A Posteriori (MAP) estimation) as is explained in Section 2.3.5.

2.3.1 Labeling Problem

A number of computer vision problems can be posed as labeling problems. For example consider the problem of image segmentation: Here segmenting an image boils down to the problem of assigning a unique label out of two possible labels to each pixel. The two possible labels being either foreground or background.

A labeling problem is completely defined by two sets : site set and label set. The site set is the set of image features e.g. the pixels in an image, image regions, edges in an image etc. which can have some properties and the label set is these set of properties which can

be assigned to site set e.g. in segmentation a pixel can be in foreground or background. All the members of the label set are possible candidates which could be assigned to a particular member of the site set. This leads to a very large set of possible mappings as explained below.

Let the set of sites $\mathbb{S}$ and labels $\mathbb{L}$ be denoted as

$$\begin{aligned} \mathbb{S} &= \{1, 2, \ldots, m\}. \\ \mathbb{L} &= \{l_1, l_2, \ldots, l_k\}. \end{aligned}$$

where m is the number of sites and k is the number of labels. For segmenting an image of size $h \times w$ into foreground and background, we have $m = h \times w$, $k = 2$, (l_0 = foreground and l_1 = background). A labeling can be defined as a function g which maps sites to labels as

$$g : \mathbb{S} \rightarrow \mathbb{L}.$$

Each possible mapping where all the sites in $\mathbb{S}$ are assigned some label from the set $\mathbb{L}$ is referred to as a *configuration*. Thus, the total number of possible labeling configurations O is

$$O = \underbrace{\mathbb{L} \times \mathbb{L} \cdots \times \mathbb{L}}_{\text{m times}} = \mathbb{L}^m$$

which is exponential in size. One of these configurations will be the optimal configuration. Since the search space of all possible labeling C is large, finding optimal labeling becomes an NP - hard problem. An energy function encodes any particular labeling into an objective function and the value of that objective function becomes a quantitative measure of the goodness of the various labeling. A number of problems in Computer Vision can be addressed using this general framework of labeling:

- *Image Segmentation:* $\mathcal{S} = \{pixels\}$ and $\mathcal{L} = \{0, 1\}$ (see [10]).
- *Stereo Reconstruction:* $\mathcal{S} = \{pixels\}$ and $\mathcal{L} = \{disparities\}$ (see [12]).
- *Image Restoration:* $\mathcal{S} = \{pixels\}$ and $\mathcal{L} = \{intensities(0, \ldots, 255)\}$ (see [33]).
- *Texture Synthesis:* $\mathcal{S} = \{pixels\}$ and $\mathcal{L} = \{patches\}$ (see [50]).
- etc..

In the following section, we explain Markov Random Fields (MRF) and show its equivalence to Maximum a Posteriori(MAP) estimate of underlying labels given the input data. This equivalence leads to the formulation of an energy function which can be minimized using Belief Propagation [29].

The Markov Property - Markov Random Field (MRF) is a branch of probability theory for analyzing the spatial or contextual dependencies of a physical phenomena. The concept of MRFs has its origins from statistical physics where Ising used this model to explain certain empirically observed facts about ferromagnetic materials [46]. It is used in a labeling problem to establish probabilistic distributions of interacting labels at each site as follows.

Let $\mathbb{F} = \{F_1, \ldots, F_m\}$ be a family of random variables defined on the set $\mathbb{S}$, in which each random variable F_i takes a label l_i in $\mathbb{L}$. The family $\mathbb{F}$ is called a *random field*. We use the notation $F_i = l_i$ to denote the event that F_i takes the label l_i and the notation $(F_1 = l_1, \ldots, F_m = l_m)$ to denote the joint event. For simplicity, a joint event is abbreviated as $\mathbf{F} = \mathbf{l}$ where $l = \{l_1, \ldots, l_m\}$ is a configuration of $\mathbb{F}$, corresponding to a realization of the field. For a *discrete* label set $\mathbb{L}$, the probability that random variable F_i takes the value l_i is denoted $\Pr(F_i = l_i)$, abbreviated $\Pr(l_i)$ and the joint probability is denoted $\Pr(F = l) = \Pr(F_1 = l_1, \ldots, F_m = l_m)$ and abbreviated $\Pr(l)$. Similarly, corresponding to a *continuous* label set $\mathbb{L}$, we have probability density functions(pdf) $p(F_i = l_i)$ and $p(F = l)$.

$\mathbb{F}$ is said to be a Markov Random Field on $\mathbb{S}$ with respect to a neighborhood system N if and only if the following two conditions are satisfied:

1. $\Pr(\mathbb{F} = l) > 0 \quad \forall\, l \in \mathbb{F}$ (Positivity).

2. $\Pr(l_i | l_{\mathbb{S}-\{i\}}) = \Pr(l_i | l_{N_i})$ (Markovianity).

where $\mathbb{S} - \{i\}$ is the set difference, i is some site in $\mathbb{S}$ such that $i \leq m$, $l_{\mathbb{S}-\{i\}}$ denotes the set of labels at the remaining sites in $\mathbb{S} - \{i\}$ and

$$l_{N_i} = \{l_{i'} | i' \in N_i\}.$$

denotes the set of labels at the sites neighboring i. The first statement signifies that each configuration of the labels is probable and the second statement means that a label at a given

site i depends solely on the labeling of the neighbors of i. We describe neighboring system and the concept of *Cliques* in the next section which are useful in showing the equivalence of MRF to a Gibbs distribution.

2.3.2 Neighborhood System and Cliques

Neighborhood System - The sites in $\mathbb{S}$ are related to one another via a neighborhood system N which is defined as

$$N = \{N_i | \forall i \in \mathbb{S}\}.$$

where N_i is the set of sites neighboring the site i. The neighboring relationship has the following properties.

1. A site is not neighboring to itself : $i \notin N_i$,
2. The neighboring relationship is mutual : $i \in N_{i'} \iff i' \in N_i$.

For a regular lattice $\mathbb{S}$, the neighboring set of i : N_i is defined as the set of nearby sites within a radius of r. Thus,

$$N_i = \{i' \in \mathbb{S} | [dist(pixel_{i'}, pixel_i)]^2 \leq r, i' \neq i\}.$$

where $dist(A, B)$ denotes the Euclidean distance between A and B and r takes an integer value. Depending on the value of r, the neighborhood systems can be classified into different orders of neighborhood system e.g. first order neighborhood system where any site $x \in \mathbb{S}$ has 4 neighbors (See Fig. 2.1), second order neighborhood system has 8 neighbors around x (See Fig. 2.1). When the sites in a regular rectangular lattice $\mathbb{S} = \{(i, j) | 1 \leq i, j \leq n\}$ correspond

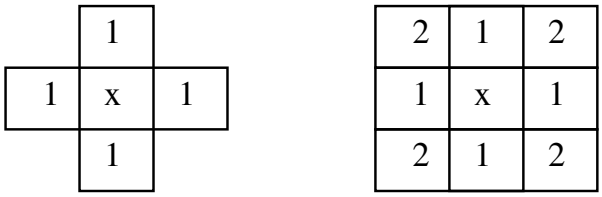


Figure 2.1 The left shows a first order neighborhood relationship between sites and the right shows a second order relationship. The numbers denote the order of neighborhood relationship.

to the pixels of an $n \times n$ image in the 2D plane, an internal site (i, j) has four nearest neighbors as $N_{i,j} = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\}$ and a site at the boundary has three and a site at the corner has two nearest neighbors.

Cliques - A 2D lattice corresponds to a regular graph where the vertices of the graph correspond to the sites and the edges in the graph correspond to the neighborhood system among the sites as described above. Thus a graph can be denoted as $G \triangleq (\mathbb{S}, N)$. A *clique* in a graph is a set of pairwise adjacent vertices, or in other words, an induced subgraph which is a complete graph. The set of cliques $\mathbb{C}$ in the graph G can consist of singe site $c = \{i\}$, pair of neighboring sites $c = \{i, i'\}$, triple of neighboring sites $c = \{i, i', i''\}$ and so on. Thus we can denote these cliques as

$$\begin{aligned} C_1 &= \{i | i \in \mathbb{S}\}. \\ C_2 &= \{\{i, i'\} | i' \in N_i, i \in \mathbb{S}\}. \\ C_3 &= \{\{i, i', i''\} | i, i', i'' \in \mathbb{S} \text{ are neighbors to one another}\}. \end{aligned}$$

The collection of all cliques of $(\mathbb{S}, N)$ is

$$\mathbb{C} = C_1 \cup C_2 \cup C_3 \cdots$$

In Fig. 2.2, we show the various sized cliques for a second order neighborhood system in a 2D lattice. As the order of the neighborhood system increases, the number of cliques grow rapidly.

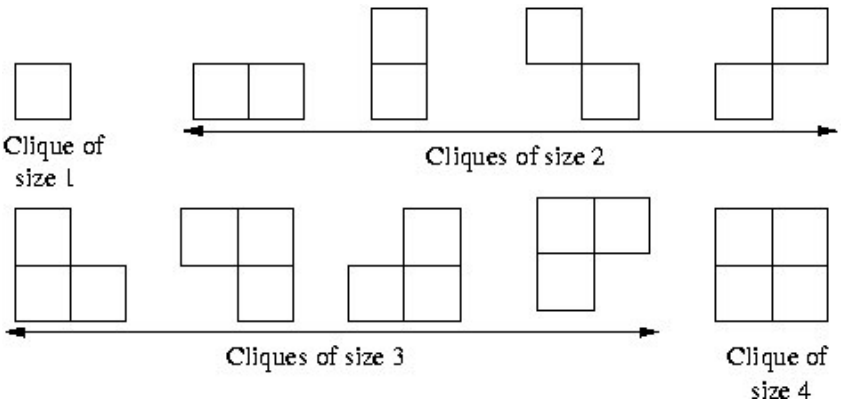


Figure 2.2 Cliques of various sizes in a second order neighborhood system.

2.3.3 Gibbs Random Fields

A set of random variables $\mathbb{F}$ is said to be a **Gibbs Random Field (GRF)** on $\mathbb{S}$ with respect to the neighborhood system N if and only if its configurations obey a Gibbs distribution. A Gibbs distribution for a given labeling l has the following form

$$\Pr(l) = Z^{-1} \times e^{-\frac{1}{T}U(l)}.$$

where

$$Z = \sum_{l \in \mathbb{F}} e^{-\frac{1}{T}U(l)}.$$

is the normalizing constant called the partition function and T is a constant called the temperature and assumed to have a value of 1. $U(l)$ is called the energy function and is given as

$$U(l) = \sum_{c \in \mathbb{C}} V_c(l).$$

which is a sum of clique potentials $V_c(l)$ over all possible cliques $\mathbb{C}$. The value of $V_c(l)$ depends on the local configuration of the clique c. Expanding the above equation in terms of cliques of various sizes we get

$$U(l) = \sum_{\{i\} \in C_1} V_1(l_i) + \sum_{\{i,i'\} \in C_2} V_2(l_i, l_{i'}) + \sum_{\{i,i',i''\} \in C_3} V_3(l_i, l_{i'}, l_{i''}) + \cdots$$

An important special case is when only cliques of size up to two are considered. In this case, the energy can also be written as

$$U(l) = \sum_{i \in \mathbb{S}} \sum_{i' \in N_i} V_2(l_i, l_{i'}).$$

Thus, the Gibbs distribution for a particular labeling l can be given as

$$\Pr(l) = Z^{-1} \times e^{-\frac{1}{T}\sum_{i \in \mathbb{S}} \sum_{i' \in N_i} V_2(l_i, l_{i'})}.$$

2.3.4 Markov-Gibbs Equivalence

An MRF is characterized by its local property (the Markovianity) whereas a GRF is characterized by its global property (the Gibbs distribution). The *Hammersley-Clifford* theorem [37]

establishes the equivalence of these two types of properties. The theorem states that $\mathbb{F}$ is an MRF on $\mathbb{S}$ with respect to N if and only if $\mathbb{F}$ is a GRF on $\mathbb{S}$ with respect to N. A proof that a GRF is an MRF is given as follows. Let $\Pr(l)$ be a Gibbs distribution on $\mathbb{S}$ with respect to the neighborhood system N. Consider the conditional probability

$$\Pr(l_i|l_{\mathbb{S}-\{i\}}) = \frac{\Pr(l_i, l_{\mathbb{S}-\{i\}})}{\Pr(l_{\mathbb{S}-\{i\}})} = \frac{\Pr(l)}{\sum_{l'_i \in \mathbb{L}} \Pr(l')}.$$

where $l' = \{l_1, \ldots, l_{i-1}, l'_i, l_{i+1}, \ldots, l_m\}$ is a configuration which agrees with l at all sites except possibly i. Using $\Pr(l) = Z^{-1} \times e^{-\sum_{c \in \mathbb{C}} V_c(l)}$ in the above equation, we get

$$\Pr(l_i|l_{\mathbb{S}-\{i\}}) = \frac{e^{-\sum_{c \in \mathbb{C}} V_c(l)}}{\sum_{l'_i} e^{-\sum_{c \in \mathbb{C}} V_c(l')}}.$$

Now, the set of cliques $\mathbb{C}$ can be divided into two sets $\mathbb{A}$ and $\mathbb{B}$ with $\mathbb{A}$ consisting of cliques containing i and $\mathbb{B}$ with cliques not containing i. Then the above can be written as

$$\Pr(l_i|l_{\mathbb{S}-\{i\}}) = \frac{\left[e^{-\sum_{c \in \mathbb{A}} V_c(l)}\right] \left[e^{-\sum_{c \in \mathbb{B}} V_c(l)}\right]}{\sum_{l'_i} \left\{\left[e^{-\sum_{c \in \mathbb{A}} V_c(l')}\right] \left[e^{-\sum_{c \in \mathbb{B}} V_c(l')}\right]\right\}}.$$

Because $V_c(l) = V_c(l')$ for any clique c that does not contain i, the term $e^{-\sum_{c \in \mathbb{B}} V_c(l)}$ cancels from both the numerator and denominator. Therefore, this probability depends only on the potentials of the cliques containing i,

$$\Pr(l_i|l_{\mathbb{S}-\{i\}}) = \frac{e^{-\sum_{c \in \mathbb{A}} V_c(l)}}{\sum_{l'_i} e^{-\sum_{c \in \mathbb{A}} V_c(l')}}.$$

that is, it depends on labels at i's neighbors. This proves that a Gibbs random field is a Markov Random Field. The reverse proof that an MRF is a GRF is given in [45]. This equivalence between MRF and GRF provides a simple way of specifying the joint probability of the labels l on the grid $\mathbb{S}$. The joint probability $\Pr(F = l)$ can be obtained by specifying the clique potential functions $V_c(l)$ and choosing the appropriate potential functions according to the problem. One of the classical potential functions of pairwise cliques C_2 is the Pott's model where we have

$$V_2(l_i, l_j) = \begin{cases} 1 & \text{if } l_i \neq l_j \\ 0 & \text{otherwise} \end{cases}$$

This simple case enforces the neighbor sites to have the same label and is applicable to many computer vision energy functions. A number of other potential functions are discussed in [97].

2.3.5 Maximum A Posteriori (MAP) - Markov Random Field (MRF) Labeling

The realization of the labeling $\mathbb{F} = l$ is not accessible directly, rather it can only be realized via the observation d. The conditional probability $\Pr(d|l)$ is the link between the realization and the observation. A classical method to estimate the configuration l is to use the Maximum A Posteriori estimation as follows. Lets denote the observed data as d and the unknown labeling configuration to be l. For the case of images, let the set of sites $\mathbb{S}$ be all the pixel positions in an image denoted as G and the size of G is m. At each pixel location (x, y) in the grid G we have an observed variable $d_{(x,y)}$ and an unknown label $l_{(x,y)}$ which is drawn from the set of labels $\mathbb{L}$. See Fig. 2.3 for an explanation of this realization setting. The posterior distribution of the

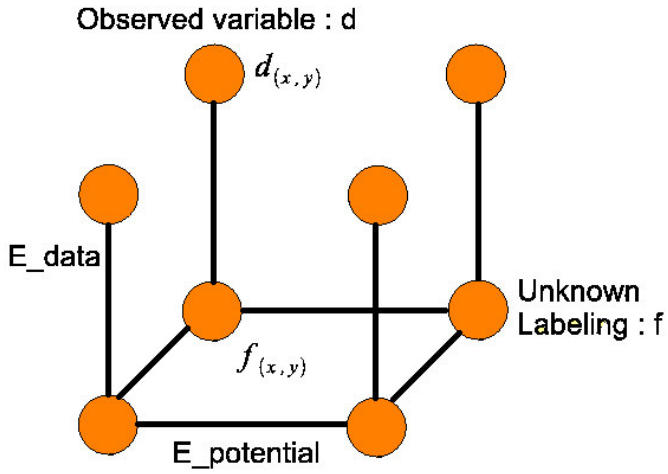


Figure 2.3 Labeling of observed variables where the unknown variables belong to a Markov Random Field

labelings l is given as $\Pr(l|d)$. From Bayes theorem

$$\underset{l}{\operatorname{argmax}} \Pr(l|d) = \underset{l}{\operatorname{argmax}} \Pr(d|l) \Pr(l).$$

where $\Pr(d|l)$ is the likelihood of generating the observation d and $\Pr(l)$ is the prior knowledge about the structure of the unknown labels l. A simple likelihood formulation can be given as

$$\Pr(d|l) = K \times \exp\left(-U(d|l)\right).$$

where K is a constant and

$$U(d|l) = \sum_{i=1}^{m} \frac{(l_i - d_i)^2}{2\sigma_i^2}.$$

The prior is given as

$$\Pr(l) = Z^{-1} \exp -U(l).$$

where from a Markov Random Field modeling of the unknown labels and a quadratic clique potential function for pairwise cliques we have

$$U(l) = \sum_{c \in \mathbb{C}} V_c(l) = \sum_{i=1}^{m} (l_i - l_{i-1})^2.$$

Here $Z = \sum_l \exp -U(l)$ and $V_c(l)$ is the clique potential defined in cliques c of size 2 in the image grid G. This potential incorporates a smoothness constraint in the final solution. Thus the posterior becomes

$$\Pr(l|d) \approx \exp\left(-U(d|l)\right) \times \exp -U(l).$$

Taking a negative log of the above equation converts the maximization of probability to minimization of an energy function. Mathematically speaking we have

$$\begin{aligned} U(l|d) &= U(d|l) + U(l) \\ &= \sum_{i=1}^{m} \frac{(l_i - d_i)^2}{2\sigma_i^2} + \sum_{i=1}^{m} (l_i - l_{i-1})^2. \end{aligned}$$

Thus the MAP estimate becomes minimizing of the posterior energy

$$l^* = \underset{l}{\operatorname{argmin}}\, U(l|d).$$

The energy function $U(l|d)$ is commonly written as $E(l)$ and consist of two terms. The first term is called the **data term** which is $\sum_{i=1}^{m} \frac{(l_i - d_i)^2}{2\sigma_i^2}$ and the second term is called **potential term** which is $\sum_{i=1}^{m} (l_i - l_{i-1})^2$ in the previous equation. As the names imply, the data term is derived from the observed data and the potential term encodes the clique potential of the underlying labeling. Thus we write

$$E(l) = E_{data}(l) + E_{potential}(l).$$

where the data term has the general form of

$$E_{data}(l) = \sum_{i \in \mathbb{S}} D_i(l_i).$$

which encodes the cost of assigning the label l_i to pixel i or in other words how much does labeling disagree. The potential term has the general form of

$$E_{potential}(l) = \sum_{\{i,j\} \in N} V_{\{i,j\}}(l_i, l_j).$$

which measure the amount of closeness in the labeling given to neighboring pixel locations i and j. Thus, the procedure of the MAP-MRF approach for solving computer vision problems is summarized in the following:

- Pose a vision problem as one of labeling and choose an appropriate MRF representation for the labeling l.
- Formulate an energy function by deriving proper likelihood and smoothness function
- Find the MAP solution by solving the energy function using optimization technique like Belief Propagation [29, 31].

The energy minimization approach has been used since long in computer vision for a number of problems e.g. image restoration and reconstruction [35, 33], shape from shading [42], stereo, motion and optical flow [40], texture [44, 19], edge detection [96], image segmentation [56], perceptual grouping [63, 69], object matching and recognition [57, 58] and pose estimation [38]. Some of the recent works which are based on optimization techniques are prominently in single view [12] and multi-view stereo [48], image restoration [33], texture synthesis [50] etc.

2.4 Summary

Total Variation -Usual choice for restoration are quadratic functionals. They give easier (lineal) mathematical problem but enforces smoothness of image and edges are not well restored. Thus the need non-quadratic functionals. Variational models exhibit the solution of

this problem as minimizers of appropriately chosen functionals. The minimization technique of choice for such models routinely involves the solution of nonlinear partial differential equations (PDEs) derived as necessary optimality conditions. Variational and PDE based models make it particularly easy to impose geometric regularity on the solutions obtained as denoised images, such as smoothness of boundaries. This is one of the main reasons behind their success.

Markov Random Field -The spatial property can be modeled through different aspects, among which, the contextual constraint is a general and powerful one. Markov random field (MRF) theory provides a convenient and consistent way to model context-dependent entities such as image pixels and correlated features. This is achieved by characterizing mutual influences among such entities using conditional MRF distributions.

Chapter 3

Super-resolution of Document Images

3.1 Introduction

With the advent of modern electronic gadgets like PDAs, cellular phones, and digital cameras, the scope of document imaging has increased. Document image analysis systems are becoming increasingly visible in everyday life. For instance, one may be interested in processing, storing, understanding a class of document images obtained by cellular phones [47]. Processing challenges in this class of documents are considerably different from the conventional scanned document images. Many of this new class of documents are characterized by low resolution and poor quality making the immediate recognition practically impossible. Super resolution provides an algorithmic solution to the resolution enhancement problem by exploiting the image-specific *apriori* information [27, 75].

Super-resolution of low resolution document images is becoming an important pre-requisite for design and development of robust document analysis systems [14, 54]. Large scale camera based book scanners employed in digital libraries could get benefited from resolution enhancement to obtain high OCR accuracies. It is also true with the text embedded in natural scenes, which could be used for indexing their images. Digital video compression algorithms can also benefit from the successful text resolution expansion techniques. Videos are often indexed and retrieved based on embedded text information. The text observed in broadcast videos is often low in resolution. Without enhancement, a simple binarization could completely remove many strokes. In these conditions, it is virtually impossible to do character recognition as most of the

OCRs are designed to work at reasonably high resolutions. Resolution enhancement algorithm increase spatial resolution, while maintaining the difference between text and background. It can further assist the recognition in low-resolution text images.

This chapter focuses on the issue of increasing the resolution of a single document image. There has been a substantial amount of previous work in super-resolution for general imagery [27, 28, 75]. However, document images are a distinct class of images widely different from natural images. The problem of document super-resolution is a special case of image super-resolution because (a) document images are pseudo binary in nature and (b) the regularity of the patterns used in this "visual" language distinguishes the document images from natural scenes. Further, due to our excessive familiarity, in the case of document images, we have fair amount of *apriori* knowledge about the high resolution image. This increases the expectations on the document super resolution algorithms. A successful document super resolution algorithm needs to use the text-specific *apriori* information. Edges are geometric regular spatial patterns, and are among the most noticeable features in document images. The visual quality near the edge areas adversely affect our perception of distortion.

In this work, we propose an algorithm for super-resolution of textual document images using an edge directed tangent field. This scheme is ideally suited for the textual content where the smoothness will have to be enforced along the edges instead of across the edges. We demonstrate the applicability of the approach on documents obtained from book scanners, cell-phone cameras and broadcast videos. We demonstrate the qualitative and quantitative improvement of this method over traditional resolution enhancement schemes.

3.2 Related Work

Simple approaches to image enhancement are popular in literature. Gaussian and Wiener filters (and a host of other linear filters) have been used for smoothing the blockiness created by the low resolution imaging [43]. Median filters (and similar nonlinear filters) tend to fare better, producing less blurry images. Interpolation methods such as cubic-spline interpolation tend to be the most common image resolution enhancement approach. There are two primary

difficulties with interpolation methods for resolution enhancement. First, smoothing in interpolation is indiscriminate. It occurs in places with gradual change, as well as across the sharp edges, producing blurring. Second, these approaches are inconsistent. Subsampling the super-resolution image will not return the original low-resolution one, which implies that the high resolution image is not the "true" high resolution image, which one is interested in estimating. Hence we need a model which not only maintains consistency but also tries to ensure that smoothing does not occur in region boundaries.

One of the earliest attempts to do super-resolution of document images was by Li and Doermann [54]. They used the method of projection onto convex sets (POCS), to deblur scene text in video sequences. This was particularly suitable for their application since overlaid text, usually have pure translation between frames. A pure translational model is a common assumption due to its simplicity and ease of implementation. In a parallel work, Capel and Zisserman [14] used a projective transform motion model for super-resolution of text specifically for image sequences in which the point-to-point image transformation was of enough complexity to demand such consideration. Both these methods successfully demonstrate the use of super-resolution to improve the document images. In a recent work, Teager filter (a quadratic unsharp masking filter) was adopted by Mancas-Thillou and Mirmehdi [67] for the extraction of high frequencies thus enhancing character edges. Most of these prior models did not reflect any text image property. This has been identified as a promising direction to derive super resolution algorithms specially suited for document images. Donaldson and Myers [22] proposed a text specific prior model, which modeled the bimodality and the local smoothness with step discontinuity. They use the Gibbs prior with a Huber gradient penalty function as their smoothness function. This piecewise smoothness prior is good at reducing false speckles in the results, but it undermines the importance of enhancing edges. Dalley *et al.* [20] employed a training-based method, in a Bayesian framework. A database is built that indicates which high-resolution patch should be output given an input low-resolution patch. Park *et al.* [74] developed an alternative approach, an edge-based super-resolution technique. It attempts to locate the edges to subpixel accuracy in a sequence of images taken from training examples, and then fuses the conglomerated edge information into the super-resolved image using a MRF formulation.

A variety of methods have been proposed to improve the contrast within a single text image. They include methods based on multi-resolution pyramid with fuzzy edge detectors [85], and a mixed approach using topological features and contour beautification [73]. There has been only limited work in the area of single frame non-training based super-resolution. Thouin and Chang [93] used nonlinear optimization on a gray scale input image to minimize a Bimodal Smoothness Average (BSA) score. Though this method works well, the processed direction of the smoothness constraint, which is a differential equation based method, is defined by the gradient magnitude of the image, where the random attribute of the image is not considered. Therefore it fails to preserve edge and texture, specially the corner edges of text image.

In general, most of the previous approaches treated document image super-resolution very similar to that of super-resolution of natural images. This resulted in adverse characteristics near the character edges and corners. Textual content in document images are primarily binary and the smoothness will have to be preserved *along* the edges and *not across* the edges. We demonstrate that such an edge preserving resolution enhancement technique is ideally suited for document images.

3.3 List of Contributions

Here are the list of contribution in this chapter

- We propose a belief propagation based discrete optimization technique for single-frame text super-resolution. We formulate a novel objective function for text images. The technique generates all the prior information on the fly and requires no training images.
- A novel way to model the bimodal nature of the document images in the MRF is shown, and is incorporated in the objective function to generate a sharp bimodal output image.
- The proposed method has edge-directed smoothing function that is tailor made for document images. It is ideally suited for the textual content where the smoothness will have to be enforced along the edges instead of across the edges.

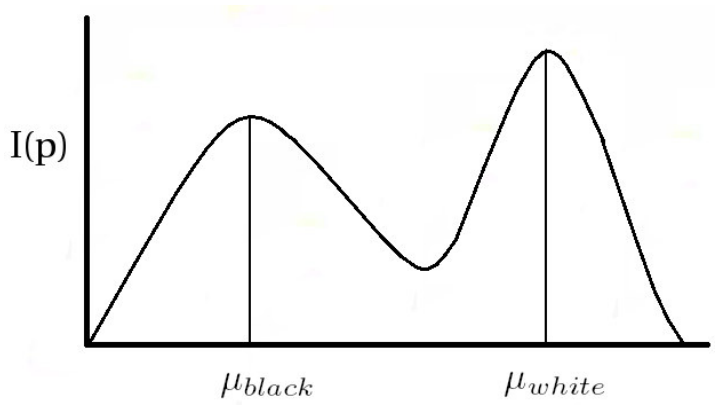


Figure 3.1 Bimodal Distribution:$I(p)$ is the gray level at pixel p. μ_{black} and μ_{white} are foreground and background peaks, respectively.

3.4 Text Specific Prior Estimation

Though some of the [73, 93] existing methods super-resolve documents, they have less emphasis on enhancing the edges. Significant amount of degradation takes place at the edges in the resolution expansion methods. Preserving character edges is most vital in document images. However, edges in low-resolution document images appear as spatially blurred edges due to degradation, sensor noise and focal blur. When edges are blurred, it is difficult to explicitly locate the edges and its digital directions. This makes the super-resolution with focus on explicit enhancement of edges in document images difficult.

To avoid the difficulties with explicit edge enhancement approach, implicit edge-directed super-resolution method is proposed in this chapter. The proposed Markov Random Field (MRF) based edge-directed super-resolution method, is an implicit edge-directed restoration. It generates the edge-directed information on the fly, making the method independent of training set.

Ideally, any algorithm to perform document image super-resolution, should have the following characteristics:

- It should be able to successfully handle the bimodal distribution so typical of a document image.
- It should preserve and enhance the edges and corners.

- Expanded images are constrained such that the subsampling the super-resolution image should return the original low-resolution one.

For practical use, we would like our method to be reasonably fast. We will also try to imbibe all the above properties in our formulation.

The general framework for the problems can be defined as follows. Let $\mathcal{P}$ be the set of pixels in an image and $\mathcal{L}$ be a set of labels. For e.g., in gray level images, there are 256 labels. The labels correspond to quantities that we want to estimate at each pixel. A labeling f assigns a label $f_p \in \mathcal{L}$ to each pixel $p \in \mathcal{P}$. The quality of a labeling is given by the energy function $E(f)$, and is defined in terms of its clique system. A neighborhood structure $\mathcal{N}_p$, which contains neighboring pixels of site p (p is not included in $\mathcal{N}_p$), is first defined. Then a clique is defined on the neighborhood structure $\mathcal{N}_p$. A set of pixel sites c in $\mathcal{N}_p$ is a clique if all pairs of sites in c are neighbors. Lastly, a function V_c, called potential function, defines the interactions of pixel sites in clique c. Spatial constraints are imposed through the formulation of potential function V_c. The potential function is related to the energy function as $E(f) = \sum_{c \in \mathcal{C}} V_c(f)$.

3.4.1 Bimodality Constraint

Images of text are usually smooth in both the foreground and background regions with sharp transitions only at the edges. Thus, text images typically have bimodal distributions, as shown in Figure 3.1, with large black and white peaks. The peak occurs at μ_{white}, the background (white) values, since the majority of pixels on a text page is background. There is a second peak at μ_{black}, representing the black letters. Additionally, there are a small number of gray values occurring between the two peaks, which represent the gray pixels that exist at transitions from white to black. The amount of these intermediate gray levels is related to the amount of blur in the document image. The textual content is almost always rendered with a high contrast, otherwise, the content provider risks, the viewer not noticing the content. In order to obtain an unblurred image, we wish to obtain a sharp bimodal distribution, pushing the intermediate gray

level towards their nearest peaks. To incorporate this property we define the energy function as

$$B_p(f_p) = (f_p - \mu_{white})^2 (f_p - \mu_{black})^2 \tag{3.1}$$

where $B_p(f_p)$ is the cost of assigning label f_p to pixel p, effecting the distribution, and is referred to as the bimodal cost. This expression measures how far the assigned label f_p is from the assigned bimodal peaks. Minimizing this expression will assign labels f_p to pixel p, values that are close to either of the peaks, making the peaks in the distribution increasingly sharper. It is interesting to notice that this energy component is capable of regulating the distribution of the document image, thus the MRF operating at a global level. As we shall see later that we try to minimize this energy component, resulting in a sharp bimodal image.

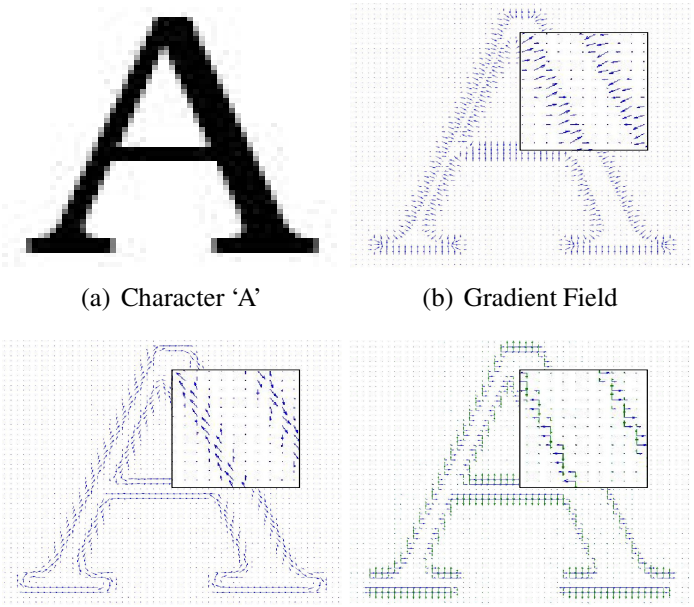
We would also like the label f_p to be as close to the gray value $I(p)$, for a pixel p. Thus, the energy term for clique with single site, is defined as

$$D_p(f_p) = (I(p) - f_p)^2 + B_p(f_p) \tag{3.2}$$

where $D_p(f_p)$ is the cost of assigning label f_p to pixel p, which is referred to as the data cost. $I(p)$ is the gray level at pixel p.

3.4.2 Smoothness Along Edges

A sharp edge in an image corresponds to relatively large intensity gradients concentrated along the edge, while a relatively smooth area is composed of a more scattered set of weaker or almost no gradients. With the exception of edges, text images tend to be very smooth in both the foreground and background regions which results in neighbors with similar values. A document image has character images with sharp curves along the boundaries as shown in Figure 3.2a. The relation between high- and low-resolution image, essentially depend on the smoothness of the edge direction. Character edges generally consist of piece wise smooth curves. The join of two curves are the corners of the characters. The enhancement approach needs to discriminate between smooth curve and the corners in the text image. Therefore, while restoring these character images, the smoothness along the character edges have to be enforced in the formulation, on the other hand maintaining sharp discontinuities across the edges. To



(a) Character 'A' (b) Gradient Field

(c) Tangent Field (d) Resolved x, y Components of Tangent Field

Figure 3.2 Tangent Field: (a) Character 'A' (b) The gradient field (c) The tangent field and (d) The resolved x and y components of the tangent field (c).

find the edge direction we first compute the gradient of image as shown in Figure 3.2b. Then we take vectors tangential to this gradient field. This tangent field consists of vectors pointing along the boundaries of character images as shown in Figure 3.2c. Let the tangent vector at pixel location p of the degraded low-resolution image I be T^p. The tangent field is further resolved into its $x-axis$ and $y-axis$ components as shown in Figure 3.2d, which are denoted as T_x^p and T_y^p, respectively. This is done because the edges are four-connected image grid graph. These potentials are used in assigning labels f_p and f_q to two neighboring pixels. We define the energy function with a quadratic cost function for the clique with two sites as

$$V(f_p, f_q) = \begin{cases} min(sT_x(f_p - f_q)^2, d) & \text{if } (p,q) \text{ are} \\ & \text{along x-axis} \\ min(sT_y(f_p - f_q)^2, d) & \text{if } (p,q) \text{ are} \\ & \text{along y-axis} \end{cases} \tag{3.3}$$

where s is the rate of increase in the cost. In order to allow for large discontinuities in the labeling the cost function stops growing after the difference becomes large. This is controlled

by the parameter d. $V(f_p, f_q)$ is the cost of assigning labels f_p and f_q to two neighboring pixels, and is normally referred to as the smoothness cost. The truncated quadratic cost changes smoothly from being almost quadratic near the origin to a constant value as the cost increases.

To understand why this approach is effective, notice that a character edge has either sharp corners or smooth curves. These geometric spatial constraints can be described by local tangent field. Our proposed MRF model-based method is an implicit edge-directed approach. In this formulation, the edge direction of an edge pixel is indicated by the continuity strength in that direction. Instead of labeling each direction as either edge or non-edge direction, we measure the continuity strength in each direction with the strength of the tangent field. These values are derived from the intensity variations, i.e., the gradient. The relative continuity strengths of the directions are used as edge direction information to formulate the geometric regular spatial constraint, which can be summarized as smoothness along edge directions and sharpness across edge directions. Areas where the gradient is zero or negligible, the smoothness cost function is very low and does not have much influence. In these places the bimodal cost is the major deciding factor, thus rendering a highly smooth surface in those regions.

3.4.3 Subsampling Consistency

The subsampling consistency should be preserved between the low-resolution and its corresponding high-resolution image, which means that when you subsample a high-resolution image generated by the method, it should recover the original input image. In this section we describe a dualscale technique to circumvent this problem. The basic idea is to impose the criteria that the expanded images are constrained such that the average of a group of high-resolution pixels is close to the original value of the low-resolution pixel from which they were derived. We use hierarchy to impose this constraint on the successive finer level.

In establishing the coarse-to-fine relation, we use the notion of dualscale image grid, as shown in Figure 3.3. The lower level corresponds to the original labeling problem we want to solve. The higher level consists of blocks of $2^m \times 2^m$ pixel locations grouped together, where m is the magnification factor, and the resulting blocks are connected in a grid structure. The

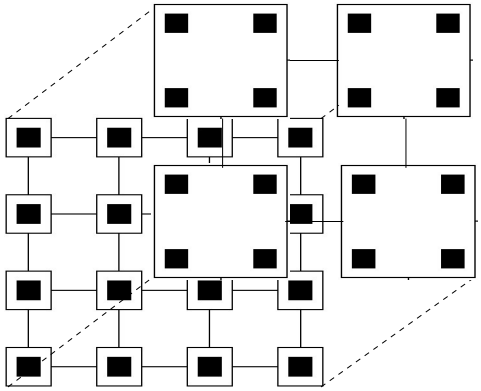


Figure 3.3 Dualscale structure: Each node in lower level(Super-resolved Image) corresponds to a block of four nodes in the higher level(Low-resolution Image). In this case the magnification factor $m = 2$.

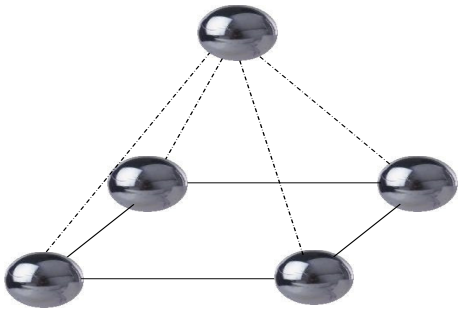
lower level and the higher level in Figure 3.3 correspond to the high- and low-resolution images respectively. A block in the higher level corresponds to a pixel in the low-resolution image. The subsampling consistency can then be conditioned as

$$S_p(f_p, b) = (I(b) - f_p)^2 \tag{3.4}$$

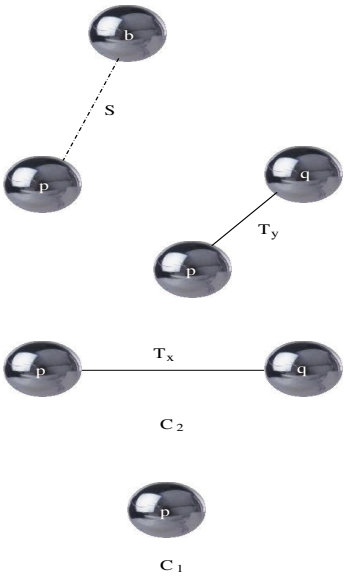
where p is a pixel at the lower level and block b is the corresponding pixel at the higher level. $I(b)$ is the gray level at pixel b. $S_p(f_p, b)$ is the cost of assigning label f_p to pixel p based on block b that measures its distance from the corresponding block at the higher level. It is referred to as the subsampling cost.

3.5 MRF Formulation to Document Super-resolution

We model the spatial relationships in images using a Markov network, which has many well-known uses in image processing [17]. This means that the probability distribution of a node on the intermediate-resolved image is conditionally independent of all but the neighborhood of the node. Figure 3.4a shows the neighborhood of a node of the MRF and Figure 3.4b shows the cliques in the neighborhood system. In Figure 3.4a, circles represent network nodes, and the lines indicate statistical dependencies between nodes. In Figure 3.4b, we define two kinds of cliques $c_1 \in C_1$ and $c_2 \in C_2$. Therefore, for each node on intermediate-resolved image,



(a) Neighborhood of a node



(b) Cliques in the neighborhood system

Figure 3.4 Clique system in the proposed MRF

there are six cliques related to it, one c_1 clique and five c_2 cliques. The clique c_1 represents the dependency between the intermediate-resolved image and the bimodality of the restored image. The clique attains higher energy value as the pixel moves away from the bimodal peak. Lowering the energy, facilitates in deriving a sharp bimodal image. The clique c_2 represents the dependency between two neighboring nodes. Clique c_2 performs two distinct tasks. First, the selective smoothing using a tangent field is performed to improve the local smoothness of each region of text region. Second, it ensures that the high resolution image does not drift far from the corresponding low resolution image. This is done by establishing a relation between the low- and high-resolution image.

The quality of a labeling in general restoration problem is given by an energy function,

$$E(f) = \sum_{(p,q)\in\mathcal{N}} (V(f_p, f_q) + S_p(f_p, b)) + \sum_{p\in\mathcal{P}} D_p(f_p) \tag{3.5}$$

where $\mathcal{N}$ are the edges in the five-connected image grid graph shown in Figure 3.4a. Here, p and q are nodes belonging to the same level and node b belongs to the immediate higher level. Finding a labeling with minimum energy corresponds to the Maximum A Posteriori (MAP) estimation problem for an appropriately defined MRF.

3.5.1 Energy Minimization using Loopy Belief Propagation

While the MRF framework yields an optimization problem that is NP hard, good approximation techniques based on graph cuts [11] and on belief propagation [29, 31] have been developed and demonstrated for problems such as stereo and image restoration. These methods are good both in the sense that the local minima they find are minima over "large neighborhoods", and in the sense that they produce highly accurate results in practice.

We start by briefly reviewing the BP approach for performing inference on Markov random fields. First we consider the max-product algorithm, which can be used to approximate the MAP solution to MRF problems. Normally this technique is defined in terms of probability distributions, but an equivalent computation can be performed with negative log probabilities, where the max-product becomes a min-sum. We use this formulation because it is less sensitive to numerical artifacts, and because it directly corresponds to the energy function definition in equation 4.6.

The max-product BP algorithm works by passing messages around the graph defined by the four-connected image grid. The method is iterative, with messages from all nodes being passed in parallel. Each message is a vector of dimension given by the number of possible labels, k. Let $m^t_{p\to q}$ be the message that node p sends to a neighboring node q at iteration t. When using negative log probabilities all entries in $m^0_{p\to q}$ are initialized to zero, and at each iteration new messages are computed in the following way,

$$m^t_{p\to q}(f_p) = \min_{f_p}\left(V(f_p, f_q) + S_p(f_p, b) + D_p(f_p) + \sum_{s\in\mathcal{N}(p)\backslash q} m^{t-1}_{s\to p}(f_p)\right) \tag{3.6}$$

where $\mathcal{N}(p)\backslash q$ denotes the neighbors of p other than q. After T iterations a belief vector is computed for each node,

$$b_q(f_q) = D_q(f_q) + \sum_{p\in\mathcal{N}(q)} m^T_{p\to q}(f_q) \tag{3.7}$$

Finally, the label f^*_q that minimizes $b_q(f_q)$ individually at each node is selected. The standard implementation of this message passing algorithm on the grid graph runs in $O(nk^2T)$ time, where n is the number of pixels in the image, k is the number of possible labels for each pixel and T is the number of iterations. It takes $O(k^2)$ time to compute each message and there are $O(n)$ messages to be computed in each iteration.

3.5.2 Algorithm Details

Proposed super-resolution process embeds the MRF super-resolution framework (Figure 3.4a) through iteration. The bicubic-interpolated image of an observed low-resolution image is given as an initial intermediate resolved image. We predict missing image details in the interpolated image to create the super-resolution output. And the intermediate-resolved image is improved by the MRF framework. The edge weights are calculated both from neighbors from same level and the immediate higher level. The quality of the final super-resolved result varies with the number of iterations. The edge weights of the same level are extracted from the tangent field and is given in Equation. 4.3. The edge that connects to the higher level in the dualscale structure (Figure 3.3) which passes the coarse-to-fine information, are estimated using Equation 4.5.

Finding the exact solution can be computationally intractable, but we find good results using the approximate solution obtained by running a fast, iterative algorithm called efficient belief propagation [29]. The algorithm runs at one level of resolution and then uses the messages at that level in order to get estimates for the messages at the next finer level, and so on, down to the original grid. Three or four iterations at each level and a maximum of five levels of grid hierarchy are sufficient. Inference algorithms based on belief propagation have been found to

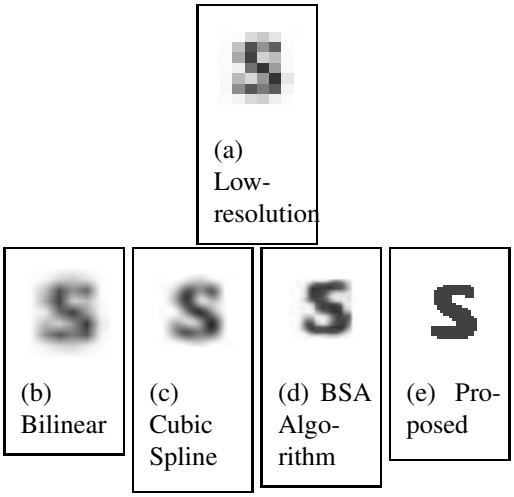


Figure 3.5 Character 's' super-resolved by a factor of 4 times

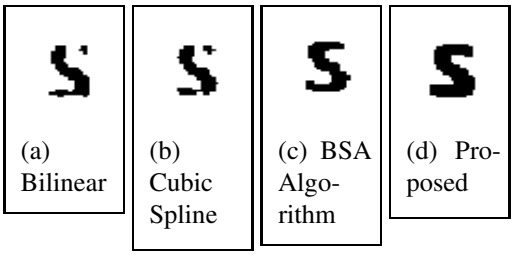
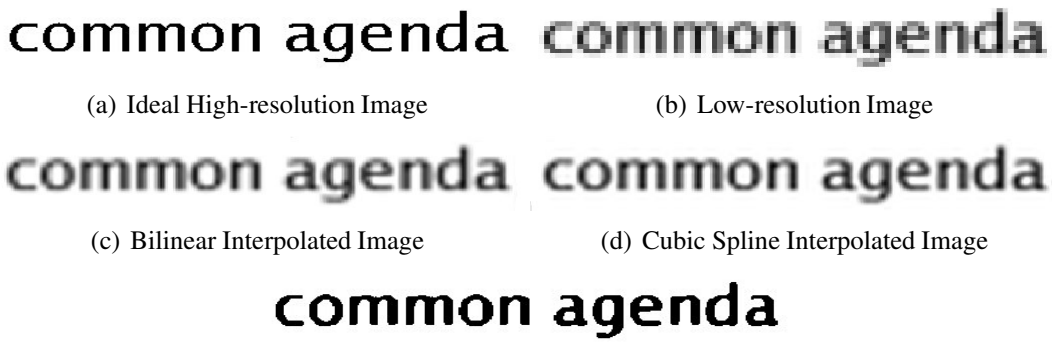


Figure 3.6 Thresholded version of the results of several methods in Figure 3.5.

yield accurate results, but despite recent advances are often too slow for practical use [29]. For a full-size page (of size 1600×2600) the processing time is way beyond any commercial use. To make it work more efficiently for a document page, we try to divide and conquer the problem. Each character in a document is visually an independent entity, not to mention about a word or a line. Geometrically it does not depend on one another. Thus they can be dealt separately without effecting the overall document image. In our method, the low resolution full size paper is segmented to word or line level as per the feasibility. These small chunks of images are then fed into the algorithm, drastically reducing the time and space complexity.



(a) Ideal High-resolution Image (b) Low-resolution Image

(c) Bilinear Interpolated Image (d) Cubic Spline Interpolated Image

(e) Using Proposed Method Restored Image

Figure 3.7 Text super-resolved by a factor of 4 times

3.6 Experimental Results

We demonstrate the performance of our algorithm on textual content in video frames as well as the document images obtained by book scanners, and cellphone cameras. We quantitatively and qualitatively demonstrate the superiority of the proposed model.

To show the effectiveness of our method, we compare the results with several common methods, including bilinear interpolation, cubic-spline interpolation and BSA algorithm [93]. Figure 3.5 shows resulting images obtained from linear interpolation, cubic spline expansion, BSA algorithm and our method. The character 's' from an image scanned at 75 dots per inch (dpi) using 8-bit gray scale quantization is shown in Figure 3.5a where significant blockiness is apparent. Bilinear interpolation results in a continuous curve, with a discontinuous derivative. These images naturally tend to be smooth, without sharp discontinuities, producing blurry results. Bilinear interpolation by a factor of four was used to create the image in Figure 3.5b, which is very blurry and lacks good contrast. Cubic-spline interpolation is an alternate popular scheme. The disadvantage of cubic splines is that they could oscillate in the neighborhood of an outlier producing a ringing effect. Figure 3.5c depicts the resulting image from cubic spline expansion which has better contrast but is still not sharp at the edges. The image obtained using BSA restoration in Figure 3.5d has superior to the images obtained using other interpolation methods for this example. This method allows for sharp edges but does not discriminate between general text edge and corners. Our method presents an edge-directed super-resolution

(a) Low-resolution Camera Based Image (b) Our method

Figure 3.8 Camera based results. A small portion of the text is magnified and displayed.

(a) Low-resolution Image from television broadcast (b) Our method

Figure 3.9 Result on text from television broadcast frames.

algorithm. Consequently the local edge direction are represented well by this method. Figure 3.5e shows, image quality is improved, strokes are reconstructed more precisely, linearity and smoothness of contours are improved, stroke width is more uniform, and shape features of fonts are reconstructed finely. Figure 3.6 shows the thresholded version of the results of several methods in Figure 3.5. Bilinear and cubic spline methods introduces cut in the thresholded image as shown in Figure 3.6a and Figure 3.6b, respectively. Figure 3.6c shows that there is still blockiness left on the smooth surface of the character 's', introduced by BSA algorithm. The reason being that the algorithm breaks the whole image into 4×4 blocks and each of these blocks are handled independently, resulting in lack of continuity across the blocks. Figure 3.6d shows that there is not much difference from the original image in Figure 3.5d even after thresholding as our method generates a sharp bimodal image. The image obtained using our method has smooth edges and is superior to the images obtained using other methods.

We demonstrate the effectiveness by creating low-resolution images from high-resolution originals, expanding the low-resolution imagery, and then measuring the distance to the originals. To achieve this an anti-aliasing process is performed by blurring (low pass filter) the image followed by block averaging (subsampling). For an image I, of r rows and c columns and a low pass filter with impulse response G, the resulting image i subsampled at each Δ

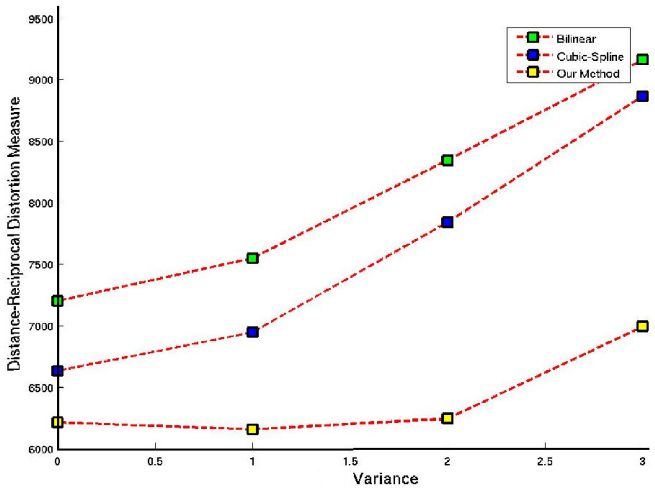
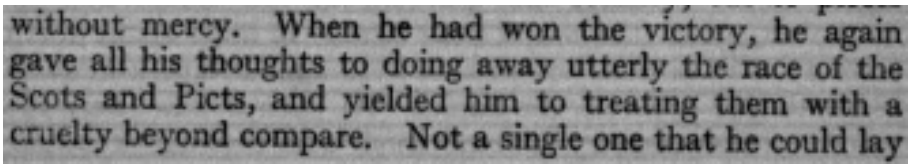


Figure 3.10 Text super-resolved by a factor of 4 times

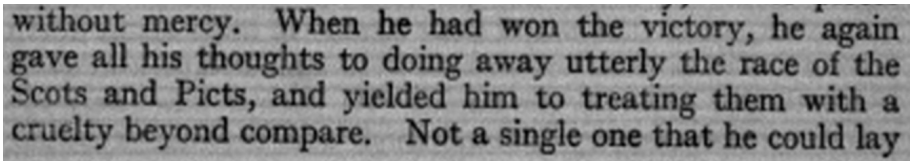
pixels would be represented by

$$i(x,y) = \sum_{j=1}^{r} \sum_{i=1}^{c} I(i,j)G(x-i, y-j),$$

where $x = 1, \ldots, c/\Delta$ and $y = 1, \ldots, r/\Delta$. Restored images are then compared with the original to determine the success of restoration numerically. For binary document images, the PSNR does not match well with subjective assessment, since it is a point-based measurement, and mutual relations between pixels are not taken into account. Hence, we use the Distance-Reciprocal Distortion Measure (DRDM) that measures the visual distortion in digital binary document images and matches well to the subjective evaluation by human visual perception [64]. The DRDM was used to compare the various methods of image resolution expansion. We initially take a 70×380 size image at 300dpi as shown in Figure 3.7a. The low resolution image is generated by the process of anti-aliasing, where a Gaussian low pass filter of standard variance $\sigma = 1$ and block averaged with $\Delta = 4$ was used shown in Figure 3.7b. The bilinear interpolation produces a severely blurred image shown in Figure 3.7c, reducing the DRDM to 7549.7. The cubic-spline gives better result in Figure 3.7d with DRDM reduced to 6945.3. Our method produced the best image shown in Figure 3.7e by reducing the DRDM to 6156.4. The sharp decline in DRDM score justifies our claim. A comparative study of the



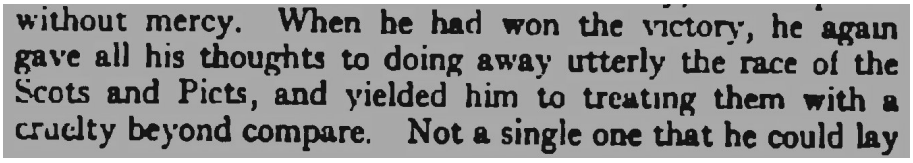
(a) Low-resolution Image



(b) Spline Interpolated Image

without mercy. When be had won the victory, he
ve all his thoughts to doing away utterly the race of the
gots and Picts, and yielded him to treating them with a
cruelty beyond compare. Not a single one that he could **hy**

(c) Spline Interpolated OCR text



(d) Our method

without mercy. When he had won the victory, he again
gave all his thoughts to doing away utterly the race **ol** the
Scots and Picts, and yielded him to treating them with a
cruelty beyond compare. Not a single one that he could lay

(e) Our method OCR text

Figure 3.11 An example text block.

reduction in DRDM for the various image expansion techniques is plotted in Figure 3.6. We observe that higher blur factor leads to greater error during restoration.

Experiments with Camera-Based images is conducted by capturing document images using a Cannon hand held camera. Result on camera-based image is displayed in Figure 3.8 Text in a video broadcast frames are rendered in very low-resolution. Result obtained by super-resolving these images is shown in Figure 3.9.

We examined effectiveness of the proposed method for improving OCR accuracy. A set of 20 page from a book were used, where a page consist of approx $350 \sim 400$ words. Each page was then scanned using 8-bit gray scale quantization at 100 dpi to create low-resolution original images using a ZEUTSCHEL OS 5000 scanner. These 100 dpi resolution pages were then expanded using various resolution expansion methods by a factor of four to create 400 dpi images which were processed by OCR. Restored images in 400 dpi were generated from input images in 100 dpi by the proposed method. The OCR accuracy, using FreeOCR Version 2.2, a freely downloadable OCR package, was compared with the results of images that were expanded using various other resolution expansion methods by a factor of four. There were a total of 28708 characters in these 20 images. Cubic spline interpolation resulted in 1558 character errors and our method had 869 character errors for an overall improvement of 44.2% for this set of images. The expansion required about $6\frac{1}{2}$ min per page for our restoration algorithm. A sample section of restored images using cubic spline expansion and our method are shown in Figure 3.11. Figure 3.11a shows the original low-resolution image. We observe that the text is bimodal where μ_{black} and μ_{white} are 20 and 170, respectively. The reason for improvement in OCR-accuracy is possibly the enhancement of the edge directed tangent field. Since many OCR algorithms use directional features along contours as primary features, the contour enhancement are effective for improving OCR accuracy as well as image quality.

3.7 Summary

An implicit edge-directed super-resolution algorithm for document images is proposed in this paper. Edge direction information is incorporated in the formulation of the energy function

in the MRF model. The edge preserving super resolution scheme provides better results on a wide class of document images. The method is quite straightforward to implement and generate good results. Our algorithm is an instance of a general non training based approach that can be useful for document image-processing, that extracts a single high-resolution frame from a single low-resolution image, where the priors are derived from same image. In this approach, the unknown pixel values are estimated based on their local surrounding neighbourhood, but not on the whole image. In particular, we donot exploit the multiple occurrence of characters in the scanned document. In the next chapter we propose to take advantage of this repetitive behaviour, we divide the image into character segments and match similar character segments to filter relevant information before the reconstruction.

Chapter 4

Text Restoration by exploiting repetitive character behaviour

4.1 Introduction

Document images are often obtained by digitizing paper documents like books or manuscripts. They could be poor in appearance due to degradation of paper quality, spreading and flaking of ink toner, imaging artifacts etc. All the above phenomena lead to different types of noise at the word level including boundary erosion, dilation, cuts/breaks and merges of characters. Restoration of such images has many applications in enhancing the performance of character recognizers as well as in book readers used in digital libraries. Often, along with the restoration, one also looks for enhancement of the resolution. Text observed from these sources is often low-resolution degraded images, and requires restoration and resolution expansion in order to improve OCR performance. Moreover, these imperfect images may be inadequate for subsequent human use. The visual and recognition ability fall due to these effects. The accuracy of today's document recognition algorithms falls abruptly when image quality degrades even slightly [3]. Significant improvement in accuracy on hard problems now depends as much, or more, on the size and quality of training sets as on algorithms and hardware [3].

Restoration and enhancement are well studied in image processing literature. The linear filters are based on the assumption of linear, space invariant degradation. The restoration technique can be carried out in the frequency domain. The linear filter is easy to design and analyze. Popular low pass noise removal filters do not make any significant assumption about the scene content. Inverse-filtering based restoration technique model the degradation (eg. motion blur)

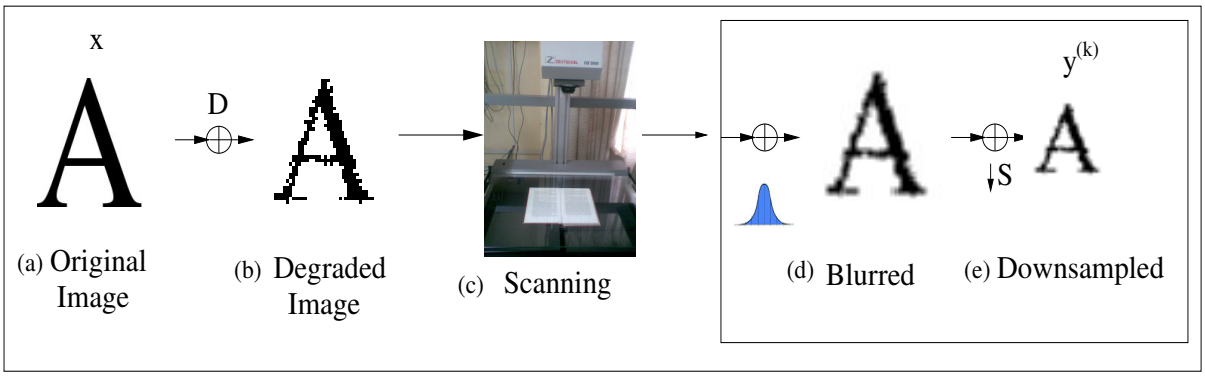


Figure 4.1 Generative Model: (a) A typical ideal image with Serif font. (b) is the Degradation version of (a) with parameters $(\alpha_0, \alpha, \beta_0, \beta) = (0.6, 1.5, 0.8, 2.0)$ [103]. (c) is the scanning process. (d) and (e) are the Blurred version and then down-sampled versions of (b), respectively. Our problem is to rectify the low resolution degraded image to a high-quality magnified document image, making it suitable for further machine and human use.

and recover the signal in a model-based framework. But document images have sharp edges. The restriction that the estimation rule be linear combination of observed values is not suitable. We exploit the properties of document images to develop a specific restoration technique, specially suited for the same. This chapter presents a document restoration technique that takes advantage of the repetitive structural nature of a document image which is further enhanced by a document specific prior information. Both prior and likelihood distributions are then formulated as a maximum a posterior (MAP) solution, which is a special case in the Bayes framework.

4.2 Related Work

There has been significant amount of research in the field of document restoration. Text enhancement efforts focus on fixing broken or touching characters [90, 102]. Traditional methods for text image enhancement can be classified into four categories: filtering, contrast enhancement, model-based image restoration, and resolution expansion. Some of the restoration efforts are based on morphological filters [103, 60] which discuss a method for binary morphological filter design to restore document images degraded by subtractive or additive noise, given a constraint on the size of filters. Bern and Goldberg [5] assume a probabilistic model of

the scanning process, and uses this model to cluster instances of the same letter and to compute super-resolved representatives of the clusters. Other methods [1] use similar model based approaches. A variety of methods have been proposed in order to improve contrast within text images. They include methods based on multi-resolution pyramid and fuzzy edge detectors [85] where document image to be enhanced is obtained from a scanner and is a blurred binary image that is corrupted by additive noise. A mixed approach using topological features and contour beautification [73] for restoring high-resolution binary images is presented to improve legibility of low-resolution document images. The initially restored image is generated by simple techniques, and is then improved by integrating a variety of features obtained through image analysis. Missing strokes of characters are complemented based on topographic features. Few of the resolution expansion approaches include text bitmap averaging [39] where the essence of the method is in finding and averaging bitmaps of the same symbol that are scattered across a text page. Outline descriptions of the symbols are then obtained that can be rendered at arbitrary resolution. Shannon interpolation is performed with text separation from the image background in [55] to improve the OCR accuracy of digital video. Restoration of images is widely considered as an example of an ill-posed inverse problem. Such problems may be approached using regularization based methods, which constrain the feasible solution space by exploiting the *a priori* knowledge [9].

A number of research efforts investigated combining text enhancement with resolution expansion in order to improve low-resolution text images. Perhaps the most salient property of text is that it is generally bimodal. By its very nature, text characters must have some contrast with the background to make them human-readable. This constraint has been successfully applied to the resolution enhancement of text in single images [93, 21]. This technique creates a strongly bimodal image with smooth regions in both the foreground and background, while allowing for sharp discontinuities at the edges. The restored image, which is constrained by the given low-resolution image, is generated by iteratively solving a nonlinear optimization problem. Dalley *et al.* [20] adopt a training-based method, where a database is build to map the output high-resolution patch for a given input low-resolution patch. Given a single image of text scanned in at low resolution from a piece of paper, return the image that is mostly likely

to be generated from a noiseless high-resolution scan of the same piece of paper. Though this method is efficient, it assumes that we have the font and script information, which is not always true.

This chapter describes a restoration technique with enhancement for document images that mimics image sequences by clustering similar character components. Spatiotemporal observation constraints are additionally added to constrain the feasible solution space with *a priori* assumptions on the form of the solution. The prior information in our formulation is independent of script and font information which is hard to predict. Our method differs from the previous work [39] in the context that we have focused on the requirement of the prior information, further combining the prior and data distribution in a Bayesian framework.

We propose a method for restoring high-quality binary images from degraded gray-scale images in low resolution. An effective approach to tackle this problem is to utilize a Bayesian inference approach. The restored image is generated from a collection of similar images by estimating the likelihood, and it is then improved by integrating with a prior information, making it a *Maximum a Posteriori* estimate. Here, we present a new image prior model based on Total Variational (TV) energy minimization. The basic idea stems from the need for preserving sharp edges, while discouraging degradations. In this chapter the performance of this method is demonstrated by showing the improvement in visual quality of the document image. Further, the results are quantitatively evaluated by running an OCR engine on the restored document images.

4.3 List of Contributions

Here are the list of contribution in this chapter

- We have developed a mathematical framework based on maximum a posteriori (MAP) to generate the prototype character from a set of similar degraded characters. The method of maximum a posteriori estimation is used to obtain a estimate of an unobserved quantity, in this case the prototype characters, on the basis of empirical data i.e, the degraded characters.

- We have proposed a prior smoothness function for document image restoration. The smoothness prior is based on variational model. The variational based method imposes geometric regularity on the solution obtained as denoised image and ensures smoothness of boundaries.
- We have proposed a document restoration algorithm that takes advantage of the repetitive structural nature of text in document images.

4.4 Document Restoration by Bayesian Inference

Given an input page as a gray-scale image, we first perform skew detection and page layout analysis upto character segmentation. We need to find images of the same character symbol that are scattered on a document page. For restoring document images, we assume that the input image is obtained by digitizing and down-sampling a degraded character. A pictorial explanation of the imaging process is given in Figure 4.1, where we see that the image gets degraded on the paper as well as while imaging. Given input pages of a document as a binary image, we segment them to obtain the word images. Connected components within this word image are then extracted from all the segmented words. The bitmaps of the segmented character images are initially clustered using a correlation based method [5]. (An alternate method is also available in [39].) We say component C_1 is equivalent to component C_2 if:

$$r(C_1/C_2) > \theta_1 \text{ and } r(C_2/C_1) > \theta_2 \tag{4.1}$$

where θ_1 and θ_2 are the tight thresholds. For our experimentation we assume θ_1 and θ_2 to be 0.85. The value $r(C_1/C_2)$ is computed as:

$$r(C_1/C_2) = \frac{\max_{x_{i,j}} corr(C_1, C_2)}{\max_{x_{i,j}} corr(C_2, C_2)}$$

where $x_{i,j}$ is an element of the correlation matrix.

Document restoration problem can now be formulated as generation of good prototypes corresponding to each cluster, where in our case the clustering is done using the Equation 4.1. We exploit the simple fact that a textual region is generated by repetition of character images

according to a language/script model. We assume that the document image being processed has enough repetitive characters to take advantage of their multiple occurrences. Since the whole page is from one book or collection, it is also in a single font.

The imaging model (Figure 4.1) specifies how the high-resolution text is transformed to generate a low-resolution degraded image. This typically involves blurring, spatial sampling and adding of noise. A high-resolution scene $\mathbf{x}$ with N pixels, is assumed to have generated a set of K low-resolution images $\mathbf{y}^{(k)}$, each with M pixels. The generative model for the kth image is

$$\mathbf{y}^{(k)} = \mathbf{W}^{(k)}\mathbf{x} + \boldsymbol{\epsilon_G}^{(k)} \tag{4.2}$$

where $\boldsymbol{\epsilon_G}$ represents noise on the low-resolution image, and consists of *i.i.d.* samples from a zero-mean Gaussian with precision β_G (equivalent to *standard deviation* $\sigma_N = {\beta_G}^{(-1/2)}$). For each image, the blurring and sub-sampling of the scene is modeled by an $M \times N$ sparse matrix $\mathbf{W}^{(k)}$ which is assumed to be parameterized by some vector $\boldsymbol{\theta}^{(k)}$. In other words, $\mathbf{W}^{(k)}$ is a function of $\boldsymbol{\theta}^{(k)}$. Given the sequence $\{\mathbf{y}^{(k)}\}$, the goal is to recover $\mathbf{x}$, without any explicit knowledge of the registration parameters $\{\boldsymbol{\theta}^{(k)}, \boldsymbol{\epsilon_G}^{(k)}\}$.

We argue that the image registration parameters may be determined *a priori*. For an individual low-resolution image, given registrations and $\mathbf{x}$, the likelihood is

$$p\left(\mathbf{y}^{(k)}|\mathbf{x}, \boldsymbol{\theta}^{(k)}, \boldsymbol{\epsilon_G}^{(k)}\right) = \left(\frac{\beta_G}{2\pi}\right)^{M/2} exp\left[-\frac{\beta_G}{2}\|\mathbf{y}^{(k)} - \mathbf{W}^{(k)}\mathbf{x}\|^2\right] \tag{4.3}$$

The vector $\mathbf{x}$ yielding the maximal value of Equation 4.3, would be the Maximum Likelihood (ML) estimation to the problem. But super-resolution images recovered in this way often tend to be dominated by a great deal of high-frequency noise [78]. Moreover, the super-resolution problem is almost always poorly conditioned, so a prior over $\mathbf{x}$ is usually required to avoid solutions that are subjectively implausible to the human viewer.

In real world applications, it is critical that we use an accurate prior model. The problem becomes more challenging when we deal with document images, because of its pseudo binary

nature and the regularity of the patterns used in this "visual" language. Images of text are also usually smooth in both the foreground and background regions with sharp transitions only at the edges. In addition, expanded images are constrained so the average of a group of high-resolution pixels is close to the original value of the low-resolution pixel from which they were derived. The challenges of complex content, various types of structures (e.g., corners, edges or surfaces) has to be incorporated in the model accurately.

We present the prior over the high resolution image by employing a total variational energy minimization function. A major concern in designing image denoising models is to preserve important image features, such as those most easily detected by the human visual system, while removing noise. One such important image feature are the edges typical of a document image; these are places in an image where there is a sharp change in image properties, which happens for instance at object boundaries. Total variation (TV) based image restoration models were first introduced by Rudin, Osher, and Fatemi in their pioneering work [81] on edge preserving image denoising. It is one of the earliest and best known examples of PDE based edge preserving denoising. It is designed with the explicit goal of preserving sharp discontinuities (edges) in images while removing noise and other unwanted fine scale detail. The revolutionary aspect of this model is its regularization term that allows for discontinuities but at the same time discourages oscillations. This algorithm seeks an equilibrium state (minimal energy) of an energy functional comprised of the TV norm of the image $\mathbf{x}$ and the fidelity of this image to the noisy input image $\mathbf{x_0}$. The minimizing energy function is:

$$\mathbf{E_{TV}} = \int_{\Omega} (|\nabla \mathbf{x}|) + \frac{1}{2}\lambda(\mathbf{x} - \mathbf{x_0})^2 \mathrm{d}u\, \mathrm{d}v \tag{4.4}$$

Here, Ω denotes the image domain, and is usually a rectangle and λ is a Lagrange multiplier.

If we assume a uniform prior over the input images, the *Maximum a Posteriori* (MAP) solution is found using the Bayes' rule. The posterior distribution over $\mathbf{x}$ is of the form

$$p\Big(\mathbf{x}|\mathbf{y}^{(k)}, \boldsymbol{\theta}^{(k)}, {\boldsymbol{\epsilon_G}}^{(k)}\Big) = p(\mathbf{x}) \prod_{k=1}^{K} p\Big(\mathbf{y}^{(k)}|\mathbf{x}, \boldsymbol{\theta}^{(k)}, {\boldsymbol{\epsilon_G}}^{(k)}\Big) \tag{4.5}$$

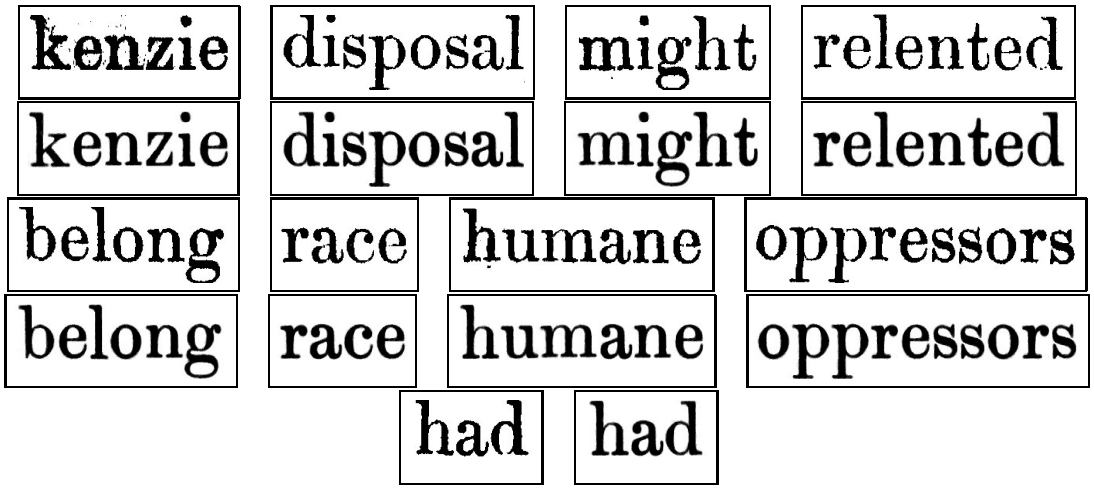


Figure 4.2 Restoration of words.

As the prior probability distribution on the super-resolution image is available, this information is used to "regularize" the estimation. Inserting this prior into Equation 4.5, the posterior over $\mathbf{x}$, and taking the negative log, the MAP *(maximum a posterior)* estimator has the form:

$$\mathbf{x}_{MAP} = \underset{x}{\operatorname{argmax}}(-\mathcal{L}) \tag{4.6}$$

where

$$\mathcal{L} = \beta\, \mathbf{E_{TV}} + \sum_{k=1}^{K} \|\mathbf{y}^{(k)} - \mathbf{W}^{(k)}\mathbf{x}\|^2$$

where the right-hand side has been scaled to leave a single unknown ratio β between the data error term and the prior term. We optimize the objective function of Equation 4.6 using conjugate gradient method to obtain an approximation to our resultant image. Here, we assume that the matrix $\mathbf{W}^{(k)}$ is available. To estimate $\mathbf{W}^{(k)}$ we have used a method suggested by Tipping and Bishop [94]. These enhanced images form the high-quality representatives of their respective clusters.

Our restoration and enhancement algorithm is based on the basic Bayesian framework. The Algorithm 1 shows the flow of our procedure. It is an iterative procedure, where at every stage we infer a better estimate of restored image $\mathbf{x}$. Assuming a set of K low-resolution degraded observation images, $\{\mathbf{y}^{(k)}\}$, the algorithm finds the corresponding high-quality image $\mathbf{x}$ such

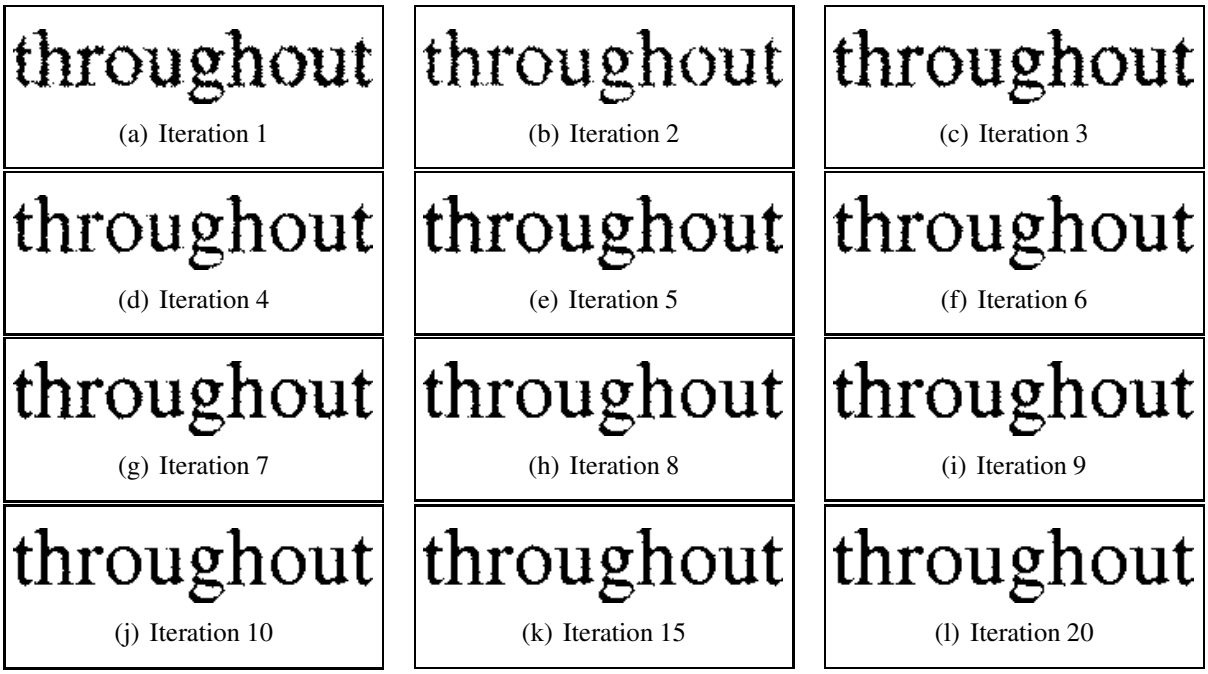


Figure 4.3 Evolution of a word image. (a) Degraded Input (b)-(k) Intermediate restored images and (l) Final restored image.

that the conditional probability of $\mathbf{x}$, given the observed images $\{\mathbf{y}^{(k)}\}$, $p(\mathbf{x}|\mathbf{y}^{(k)})$, is maximized. In our case this is difficult to calculate directly. Thus using Bayes' law, we obtain $p(\mathbf{x}|\mathbf{y}^{(k)}) \propto p(\mathbf{x})p(\mathbf{y}^{(k)}|\mathbf{x})$, which is the MAP estimator. Once $p(\mathbf{x})p(\mathbf{y}^{(k)}|\mathbf{x})$ are defined, the output image, $\mathbf{x}$, that maximizes $p(\mathbf{x}|\mathbf{y}^{(k)})$, is iteratively calculated by stepping down the gradient of the negative log likelihood of $p(\mathbf{x})p(\mathbf{y}^{(k)}|\mathbf{x})$ until a minimum is reached or a maximum number of iterations are executed. Finally, reassembling the output page by replacing each member of the cluster by its representative we restore the document.

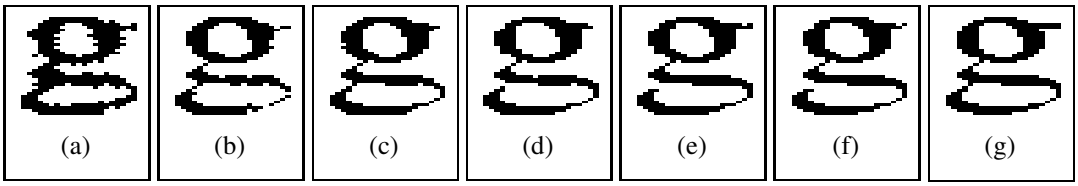


Figure 4.4 Evolution of a character image. (a) Degraded Input (b)-(f) Intermediate restored images and (g) Final restored image.

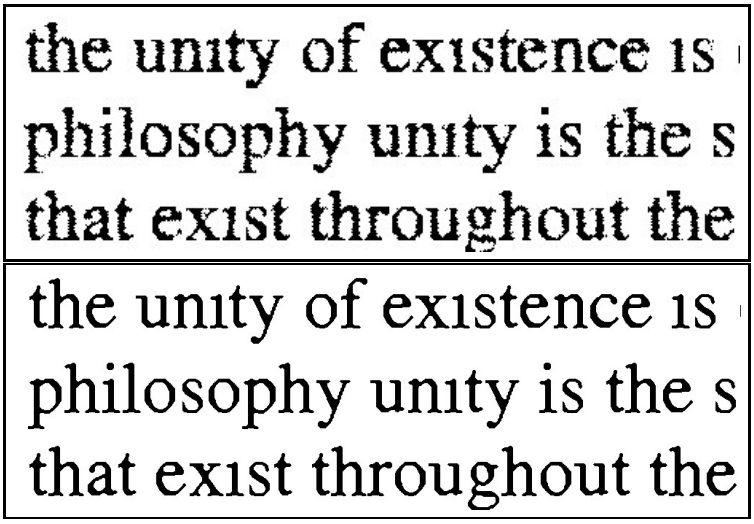


Figure 4.5 (a) Portion of text from original image (b) Portion of text from restored image.

Input: Given the document image and parameter $\mathbf{W}^{(k)}$ [**?**].
Perform a character level segmentation.
Here $\mathbf{y}$ is the input image.
Output: Here $\mathbf{x}$ is the output image.
initialization - Perform the initial clustering [Equation 1].;
foreach *cluster bin* **do**
  foreach *element of the bin* **do**
    repeat
      1) Parameterize the posterior distribution as a function of $\mathbf{x}$
      by substituting the values of $\mathbf{y}$ in Equation 4.3;
      2) The equation is then minimized using conjugate gradient algorithm
      to get a estimate of $\mathbf{x}$;
      3) Total energy minimization of $\mathbf{x}$ is then performed
      to get the next estimation on $\mathbf{x}$ [Equation 4.4];
    until *the energy is minimized*;
  end
end

Algorithm 1: MAP formulation

4.4.1 Discussion

We make the following comments about our method and its implementation. Document image processing algorithms to detect text regions, and then segmenting them to obtain word and component images are not described here. There exists significant amount of material in this respect [15]. In real-life situations, character images could be split into multiple components or merged to form single component. They may affect the clustering process. In [39] a procedure is discussed to find images of the same character symbol that are scattered on a document page. They employ a sequence of different clustering techniques, each applied to a different set of shape features derived from the character images. The motivation is to progressively divide all characters on a page into groups of decreasing sizes, and delay the uses of more expensive techniques until later stages when the groups are sufficiently small. This method is experimentally verified to be quite effective. However, in our case by defining appropriate similarity measure in clustering, they are taken care of. It is important to classify the character images into as few clusters as possible, since this is how the algorithm achieves its benefits. Yet it is even more important to avoid clustering incompatible character images since this leads to "mistakes" in the output. The clustering results are important side products of the procedure and they have other potential uses that remain to be explored. The computational requirement of this algorithm is directly proportional to the number of similar components in the cluster and the conjugate gradient method used in the optimization process. Further, it is worth while mentioning that our method differs from the previous super-resolution methods in following three aspects: (i) we do not learn a low-resolution to high-resolution match to build up our output image; (ii) since we are using energy function (i.e., total variation minimizing process) to determine our prior, we need not have any font or script information; (iii) our approach of image restoration cum resolution expansion adopts a *Maximum a Posteriori* estimation approach as it provides a rigorous theoretical framework with several desirable mathematical properties.

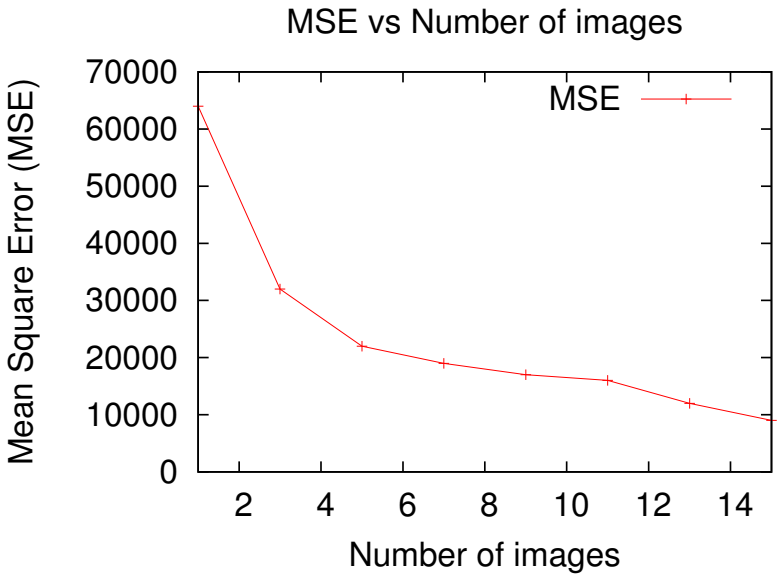


Figure 4.6 MSE with ground truth for the character image "g".

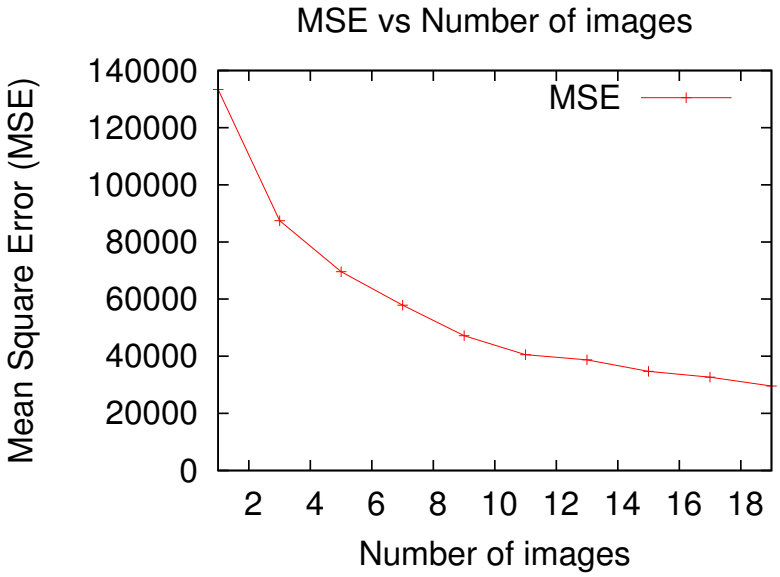


Figure 4.7 MSE with ground truth for the word image "throughout".

Specification	**Noisy Page**	**Restored Page**
Number of words	325	325
Recognized words	268	325
% Accuracy	82%	100%

Table 4.1 OCR Evaluation of image restoration results.

4.5 Experimental results

The complete algorithm is implemented on degraded document images scanned at a specific resolution. We expect restored document images as output, at the end of our experimentation. We show the effectiveness of our algorithm by demonstrating the results using samples collected mainly from degraded books. We scan these books in 200dpi using a *ZEUTSCHEL OS 5000* scanner shown in Figure 4.1(c). The scanning device used here has a mounted camera on top of the flat bed where the book is kept. The focus of the camera has to be adjusted to get a sharp image. We have scanned 20 pages from four different variety of books containing different fonts and styles. The document books already contain degradations. After the scanning process the resultant image gets blurred and down-sampled. We proceed with binarizing and skew correcting the scanned images. After a character level segmentation, we cluster the components. For a character we get an approximate of 10-15 or more similar components.

Effectiveness is demonstrated for improving image quality. Fig. 4.5b shows the generated binary image with resolution enhanced by a factor of two, along with the original image in 200dpi shown in Fig.4.5. Image quality is improved as resolution increases; strokes are reconstructed more precisely, linearity and smoothness of contours are improved, stroke width is more uniform, and shape features of fonts are reconstructed more finely. The proposed method is effective for Latin scripts as well as oriental scripts. The plot in Fig. 4.6 depicts the evolution of the degraded character "g". The *x-axis* shows the number of connected components used and the *y-axis* determines the Mean Square Error (MSE). The performance of our algorithm was evaluated with respect to the mean square error (MSE). The figure shows how the mean square error function decreases steadily as the number of collection of the similar components increases. We see that the number of similar components is directly related to the accuracy of the result. The step-by-step changes in the output of the image is shown in Fig. 4.4 where the image in the left is the degraded image and image in the extreme right is the restored image. If there are sufficient number of similar components then we get a high-quality restored image.

We examined effectiveness of the proposed method for improving OCR accuracy. Binary images in 400dpi were generated from input images in 200dpi by the proposed method, and

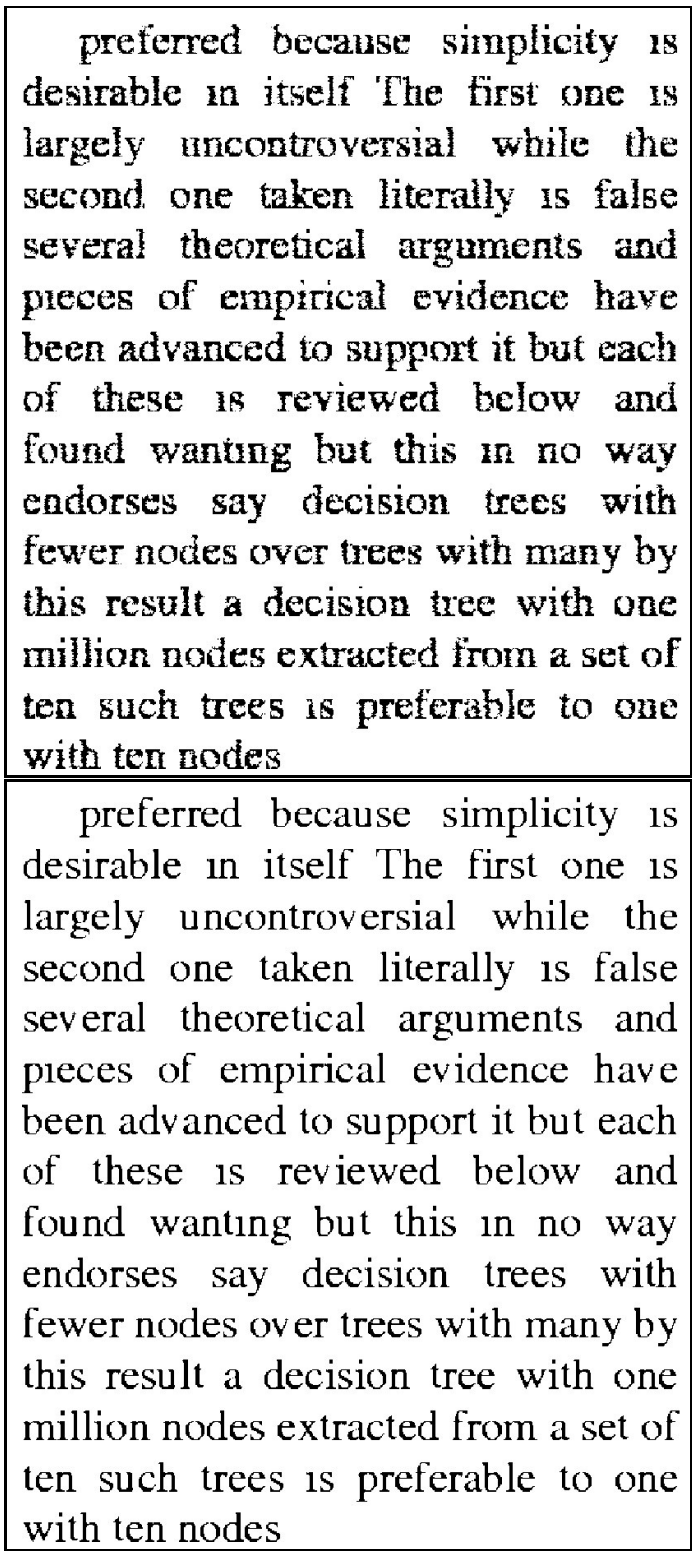


Figure 4.8 The document page on the left suffers from degradation and low-resolution. The second image on the right shows the content restored using the algorithm presented in Algorithm 1

Word Images	OCR recognized as
	**urity**
	**piHlar**
	**pur¿ty**
	**appeanng**
	**egottsrn**

Table 4.2 OCR recognition output for few of the degraded words using a commercial OCR(CuneiForm OCR).

the OCR accuracy using these images as input was compared with the results using bilinear interpolation. Gray-scale images in 400dpi generated from input images in 200dpi by bilinear interpolation which gives around 82% accuracy as shown in Table 4.1. Few of the words incorrectly recognized during the whole process are listed in Table 4.2. Our method gives around 100% accuracy. The page level output to our algorithm is shown in Figure 4.8.

4.6 Limitations of this approach

In this work, we exploit the repetitive behaviour and propose a reconstruction framework for degraded low-resolution document images. This assumes that we focus on locating characters and segmenting them. The document image acquisition process consists of making a (discrete) digital image out of a paper document. However in practice, the acquired image is corrupted by noise and blur. This makes the whole segmentation process inaccurate. The higher the degradation or noise the greater is the unpredictability of the segmentation process. Hence, the limitation of this work is that the work is built on top of character segmentation, which can be a bottle-neck in the whole process. Moreover character segmentation is not a completely solved problem [51].

4.7 Summary

To improve quality and OCR accuracy for degraded low-resolution text images, a new method has been presented for restoring high-quality binary text images from a set of low-resolution degraded image. The initially restored image is improved by MAP based approach where a suitable a priori information is used to guide the restoration, resolution enhancement being the byproduct. The proposed method can deal with various scripts, and entails relatively simple computation. Through experiments, it has been validated that the proposed method improves both OCR accuracy and image quality. But excessive dependence on character segmentation still remains a problem. In the next chapter we shall see how to overcome the dependency on character segmentation. We shall look for a restoration approach that does not perform a explicit character segmentation, but still uses the repetitive component nature of document images.

Chapter 5

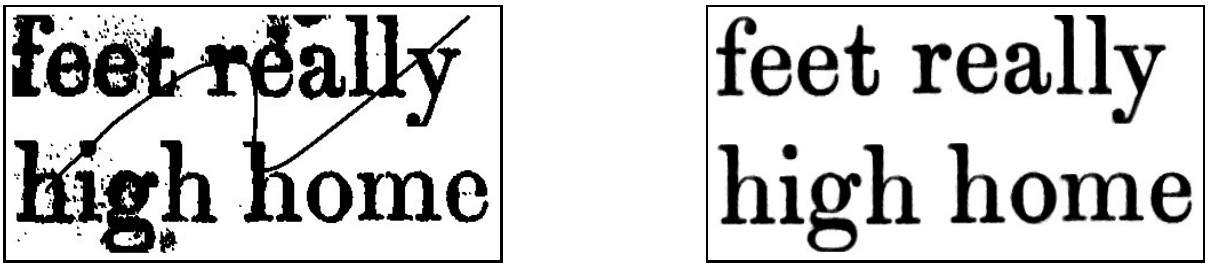
Contextual Restoration of Text Images

5.1 Introduction

Degradations in document images result from poor quality of paper, the printing process, ink blot and fading, document aging, extraneous marks, noise from scanning, etc. The goal of document restoration is to remove some of these artifacts and recover an image that is close to what one would obtain under ideal printing and imaging conditions. The ability to restore a degraded document image to its ideal condition would be highly useful in a variety of fields such as document recognition, search and retrieval, historic document analysis, law enforcement, etc.

Images with certain known noise models can be restored using traditional image restoration techniques such as Median filtering, Weiner filtering, etc. [34]. However, in practice, degradations arising from phenomena such as document aging or ink bleeding cannot be described using popular image noise models. Document processing algorithms improve upon the generic methods by incorporating document specific degradation models [83] and text specific content models [99, 21].

In image restoration the goal is to recover an image that has been corrupted or degraded. There are several techniques in image restoration, some use frequency domain concepts, others attempt to model the degradation and apply the inverse process. e.g. the blurred image that is the result of convolving a Gaussian filter with the original image, is the effect which is similar to the one observed when a photograph is taken with a camera in motion. In document images



(a) Degraded (b) Restored

Figure 5.1 Portion a vandalized degraded document and the result of our restoration process.

it is quite possible that the same character image at different physical location in a document may be degraded differently. Inverse restoration process, in this case may not possibly generate the desired result. Further, due to our excessive familiarity of document images, even a small variation in the text from the expected image will quickly draw our attention. The increased expectation and unavailability of an inverse restoration process in the case of document images, motivates us to use a patch based approach where a degraded patch is replaced by a noise free rendered patch.

5.2 Related Work

Approaches that deal with highly degraded documents (see figure 5.1) take a more focused approach by modeling specific types of degradations. For instance, ink-bleeding or backside reflection is one of the main reasons for degradation of historic handwritten documents. Huang *et al.* [41]. The success of their approach is in combining the degradation model and the document model into a powerful MRF-based optimization framework [33, 59]. To achieve generic restoration of carbon copy documents, Cao and Govindaraju [13] used a document content model. The model consisted of a set of 5×5 binary patches, trained using high quality data, which is used for restoring noise and removal of rulings on the paper. Gupta *et al.* [36] used a patch based alphabet model to remove blurring artifacts for license plate images using a camera. The authors use an MRF based optimization to find the most likely noise free patch that

generated the observations. All the above approaches consider specific instances of restoration of a single document image, and are solved by combining prior knowledge of documents with noisy observations.

In this chapter, we approach document restoration in a different, and useful setting. We consider the problem of restoration of a degraded 'collection of documents' such as those from a single book. Such a collection of documents, arising from the same source, is often highly homogeneous in the script, font and other typesetting parameters. The availability of such a uniform collection of documents for learning allows us to:

- Do robust learning of a tight model of the document content even in presence of severe degradations, as one can discard data that is potential noise.
- Do accurate parameter estimation from multiple evidences, as the amount of data available after discarding highly noisy parts is still considerable.
- Adapt to a large varieties of documents in various fonts, styles and scripts, as our model is exclusively learned from the input collection itself.

Given that we can learn an accurate and exact model of the documents content, we leverage it to compute the most probable estimate of the underlying content during document restoration. We frame the restoration process as a maximum a posteriori estimate computed from the learned document model prior and the noisy observed data in a Markov Random Field framework. Our formulation enables us to incorporate a larger context into the inferencing process, thus providing us with the ability to restore highly degraded documents.

The proposed approach is far more powerful than traditional approaches in restoring highly degraded documents as it relies on learning of a document model specific to the input. It can handle severe degradations including cuts, merges, ink blobs, or even vandalized documents. To achieve this, we address the problems of learning high quality priors and that of robust restoration in a flexible MRF-based optimization framework [30].

5.3 List of Contributions

Here are the list of contribution in this chapter

- We would like to replace the patches in the degraded image with exact noise free original patch, based on the neighboring patches. To make our algorithm efficient, we choose larger patch size and restrictive number of labels. This gives rise to serious correspondence related issues. We propose a novel overlapping markov random framework which allows us to establish the correspondences.
- An algorithmic approach, is proposed to exploits the contextual relation between image patches. This allows the system to update the constraints by reasoning about their validity in the context of an image description. Using the topological/spacial constraints between the image patches, local constraints are formulated.
- We formulate the document restoration as a labeling problem in a relaxation framework. An likelihood function encodes any particular labeling into an objective function and the value of that objective function becomes a quantitative measure of the goodness of the various labeling. An solution to the objective function is obtained using Belief Propagation [30].

5.4 Restoration by Learning

The process of restoration proceeds in two stages. In the first stage, a set of ideal patches, x_i, that can occur in the restored document are estimated, along with the spatial relationship between them. This constitutes a probabilistic document model that is specific to the input. The most likely set of patches that generated the observed patches, y_i, is estimated in the second stage, using an MRF framework. Figure 5.4 shows the construction of the patch-based MRF for a degraded word image.

The ideal (restored) patch at x_5 depends not only on the observed patch at y_5, but also on the context of x_5, given by its neighbors, $x_2, x_4, x_8,$ and x_6. For example, in figure 5.4, the restoration of y_6 depends on whether the underlying character (b, h, n, p), which is indicated by

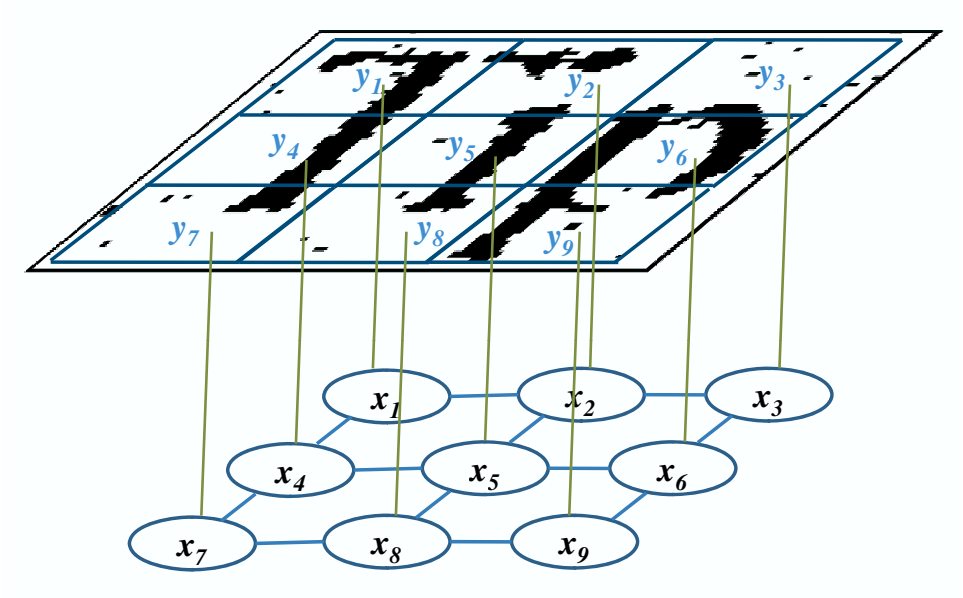


Figure 5.2 Patch-based MRF for a degraded word image (*Tip*).

its neighboring patches. The goal of the restoration stage can be thought of that of estimating the most likely set of patches, x_i, that could generate the observations, y_i, while being in their respective neighborhoods. Based on the restoration goal, the problem in the first stage is to find a set of ideal prototypes $\mathcal{X}$, that are possible in a specific document, along with the probabilities of their neighborhood values, $p(x_i, x_j)$ for each of the four neighbors.

The first stage involves the estimation of ideal prototypes from degraded ones. The primary goal as mentioned before is to identify the consistent primitives (patches in our case) in the document collection. As we have a collection of documents at our disposal, we try to estimate the model from multiple observations. The process proceeds as follows: A given document image is approximately segmented into words and characters. One could make errors in this stage. The resulting segments are clustered to identify consistent shapes in the document. Errors in segmentation or highly degraded characters are eliminated from the learning phase. The consistent, probably noisy, segments are used to compute their most probable restoration. The restored segments are then covered with patches to learn their shaped and neighborhood relations.

The challenge here is to deal with the large number of possible patches at the patch size we chose, as well as to deal with the severe degradations of characters present in the document. We refer to this step as *prototype generation*. One should also be able to generate the neighborhood relationships from the prototypes. Note that using a generic MRF model as shown in figure 5.4 will lead to a dramatic increase in the number of possible patches. Hence the second challenge is to come up with a formulation for the restoration phase, that makes the prototype generation phase easier, and the restoration, efficient. We will first look into the restoration formulation and then return to the prototype generation phase.

5.4.1 Markov Model for Restoration

The input image is segmented into words, and each word is restored independently. However, we do not assume that the word segmentation is always correct. Each word is assumed to be divided into a set of possibly overlapping patches, y_i, as shown in figure 5.4. Given a set of observed patches, y_i, form an input document image, I, we aim to compute the MAP (*maximum a posteriori*) estimate of the corresponding underlying labels, $x_i \in \mathcal{X}$.

Let $P(\bar{x}, \bar{y}) = P(x_1, \ldots, x_N, y_1, \ldots, y_N)$ be the joint probability of observing $y_1, \ldots, y_N$ when the corresponding underlying labels are $x_1, \ldots, x_N$. Let $\boldsymbol{\psi}(x_i, x_j)$ denote the pairwise compatibility of two neighboring labels, x_i and x_j, and $\boldsymbol{\phi}(y_k|x_k)$, the likelihood that the label x_k generates the observed patch, y_k. The joint probability can now be written as:

$$\begin{aligned} P(\bar{x}, \bar{y}) &= P(x_1, \ldots, x_N, y_1, \ldots, y_N) \\ &= \prod_{(i,j)} \boldsymbol{\psi}(x_i, x_j) \prod_k \boldsymbol{\phi}(y_k|x_k), \end{aligned} \tag{5.1}$$

The first product is over all neighboring pairs of nodes, i and j. To compute the MAP estimate, we solve the MRF using the belief propagation framework [76, 29]. The belief-propagation algorithm updates *messages*, m_{ij}, from node i to node j, which are used to infer the state at node j. The state of a node is updated based on the messages it receives, and the process is repeated until convergence. Let m_{kj}^t be the message being sent from the node k to j at time

instant t. The MAP estimate at node j over all label candidates x_j is:

$$\hat{x}_{jMAP} = \underset{x_j}{\operatorname{argmax}} \ \boldsymbol{\phi}(x_j, y_j) \prod_k m_{kj}^t, \tag{5.2}$$

where k runs over all neighbors of node j. We calculate m_{kj}^t as

$$\begin{aligned}
m_{j\uparrow}^t &= \max_{[x_k]} \ \vec{\boldsymbol{\psi}}(x_j, x_k)\boldsymbol{\phi}(x_k, y_k) m_{k\rightarrow}^{t-1} m_{k\uparrow}^{t-1} m_{k\leftarrow}^{t-1}, \\
m_{j\leftarrow}^t &= \max_{[x_k]} \ \vec{\boldsymbol{\psi}}(x_j, x_k)\boldsymbol{\phi}(x_k, y_k) m_{k\uparrow}^{t-1} m_{k\leftarrow}^{t-1} m_{k\downarrow}^{t-1}, \\
m_{j\downarrow}^t &= \max_{[x_k]} \ \vec{\boldsymbol{\psi}}(x_j, x_k)\boldsymbol{\phi}(x_k, y_k) m_{k\leftarrow}^{t-1} m_{k\downarrow}^{t-1} m_{k\rightarrow}^{t-1}, \\
m_{j\rightarrow}^t &= \max_{[x_k]} \ \vec{\boldsymbol{\psi}}(x_j, x_k)\boldsymbol{\phi}(x_k, y_k) m_{k\downarrow}^{t-1} m_{k\rightarrow}^{t-1} m_{k\uparrow}^{t-1}.
\end{aligned} \tag{5.3}$$

m_{lk}^{t-1} is m_{lk}^t from the previous iteration. The initial m_{kj}^0s are set to column vectors of 1's, of the dimensionality of the variable x_j. Spatial constraints are imposed through the formulation of $\vec{\boldsymbol{\psi}}(x_j, x_k)$ function. Here, $\boldsymbol{\psi}$ is not a symmetric function and depends on the orientation of x_j and x_k, enabling the prior being stronger than smoothness prior [91].

5.4.2 Localizing the Patches

As noted before, one of the main constraints in the patch based formulation is that the location of structures within a patch can vary, changing the observation probabilities, $\boldsymbol{\phi}(x_k, y_k)$. To deal with this, we allow the patches to slide around and settle at a location that best matches the underlying label. The spirit of our approach is similar to [7]. However, we note that the label itself is unknown and further depends on its neighboring nodes. Hence we need to carry out the optimization procedure described above for all possible patch locations for each of the patches. Let each patch, y_k be offset in the horizontal and vertical directions by Δp_k and Δq_k respectively, from their initial uniformly spaced locations. The function to optimize becomes:

$$P(\bar{x}, \bar{y}) = \max_{\Delta p_k, \Delta q_k} \prod_{(i,j)} \boldsymbol{\psi}(x_i, x_j) \prod_k \boldsymbol{\phi}(y_{k,\delta p_k, \delta q_k} | x_k) \tag{5.4}$$

The direct optimization of equation 5.4 over all patch locations turns out to be prohibitively expensive. To overcome the difficulty, we enforce left-to-right and top-to-bottom orderings on the centroids of the patches and formulate a dynamic programming solution to carry out the

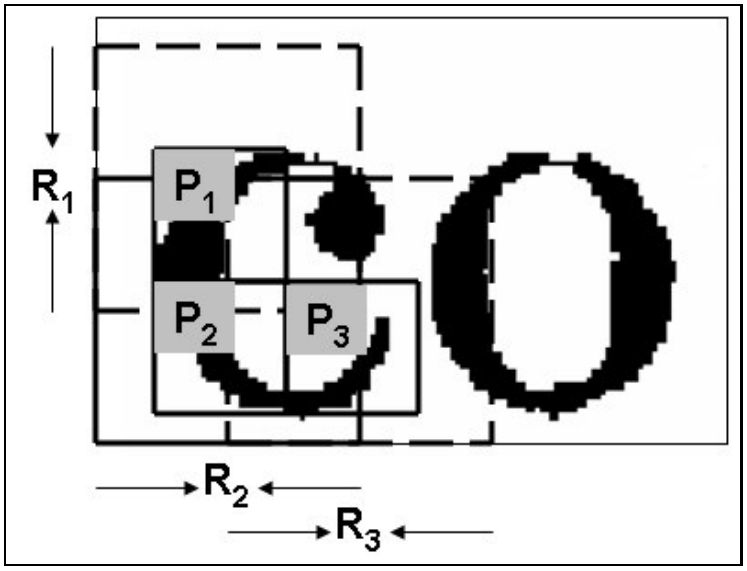


Figure 5.3 A patch can be located anywhere within a window of $m \times n$ within the word image.

computation. We further observe that the vertical sliding of a patch within a word is limited and hence for each horizontal position for a patch, we compute the best matching vertical position for every interpretation of the underlying label.

The problem is now reduced to finding the best horizontal shift for each patch. To achieve this, we apply a Viterbi decoder for every row of nodes, while keeping the patches in other rows rooted to their locations. The process is repeated, sequentially, until convergence. Note that we need to carry out an MRF optimization at every step in the Viterbi algorithm. The initialization of the patches is carried out using independent maximum likelihood estimates for the patches over all possible labels and locations within the limits. We can further improve the computation speed by restricting the range of sliding for each patch to a specific limit, restricting the most likely path (horizontal locations) within a diagonal band, referred to as the Sakoe-Chiba band [82].

A lighter version of the optimization can be obtained if we assume that the position of a patch is within one window width around its initial location. This makes the computation of path locations independent of its neighbors, and the resulting optimization function would be:

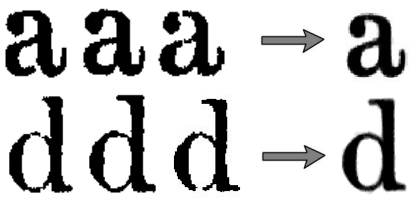


Figure 5.4 Collection of Characters and their Prototypes. A collection of 10 characters are used to generate the prototype.

$$P(\bar{x}, \bar{y}) = \prod_{(i,j)} \boldsymbol{\psi}(x_i, x_j) \prod_k \max_{\delta p, \delta q} \boldsymbol{\phi}(y_{k,\Delta p,\Delta q} | x_k) \tag{5.5}$$

$$= \prod_{(i,j)} \boldsymbol{\psi}(x_i, x_j) \prod_k \boldsymbol{\phi'}(y_k | x_k) \tag{5.6}$$

Note that the above equation leads to a regular MRF formulation. In most practical cases, we found that the direct MRF formulation using equation 5.6 leads to the same solution as the more complex Viterbi optimization using equation 5.4.

5.5 Learning the Labels and Context

To generate the label set, we generate a collection of similar characters by segmentation and clustering. Outliers in each cluster are usually errors in segmentation process or highly corrupted samples and are removed [39]. These similar characters are used to generate high quality prototypes, by bitmap averaging and restoration. Similarly, prototypes are generated from all the clusters. Figure 5.4 shows examples of two prototypes, corresponding to characters 'a' and 'd' being generated from the noisy samples.

To learn the context relationship, each character prototype is divided into collection of patches. Different characters have different dimensions. They are divided into equal $N \times N$ sized labels. Similarly labels are extracted from other character prototypes. The collection of patches from all the characters form the possible set of labels. These patches are typically of

size 25×25 for a 600 dpi image. Large patch size means that the prior is defined on large neighborhood, making it more powerful.

We sample the patches so that they overlap with each other by few pixels. In the overlap region, the pixel values of compatible neighboring patches should agree. We measure $d(x_i, x_j)$, the sum of squared differences between patch candidates x_i and x_j in their overlap regions at nodes i and j. The compatibility matrix between nodes i and j is then

$$\vec{\boldsymbol{\psi}}(x_i, x_j) = \exp\left(-\frac{\vec{d}(x_i, x_j)}{2\sigma^2}\right), \tag{5.7}$$

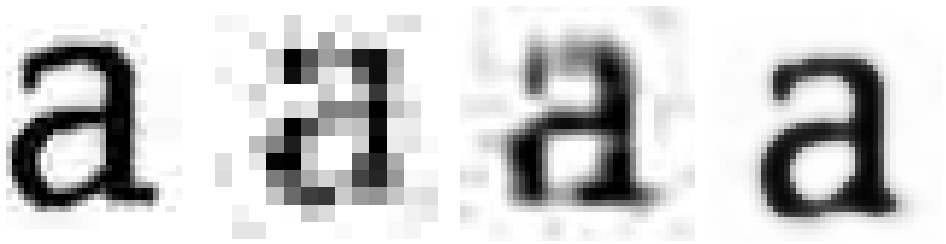
where σ is a noise parameter [31]. We use a correlation based penalty on differences between the observed degraded image patch, y_i, and the candidate label patch, x_i, found from the prototype to specify $\boldsymbol{\phi}(y_k|x_k)$.

5.5.1 Document Image Super-resolution

One of the advantages of our formulation of learning ideal patches from multiple degraded or low-resolution patches is that we can directly estimate the ideal patches at high resolution, thus combining document restoration and super-resolution into one process. We propose the use of a MRF-based MAP estimation to generate the super-resolved prototypes. The overall process proceeds as follows:

- Upsample the low-resolution, degraded prototypes using cubic spline interpolation.
- Register the prototypes at high resolution using correlation and compute mean prototypes.
- Obtain the super-resolved patches by computing the MAP estimate of the underlying high resolution prototype.

To achieve the third step, we use a text specific prior and formulate the estimation in an MRF framework. Images of text are usually smooth in both the foreground and background, with sharp transitions only at the edges. Thus, text images typically have bimodal distributions, with large black and white peaks [21]. The peak occurs at μ_{white} for the background (white),



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

Figure 5.5 Super-resolution by a factor of 3: (a) original high-resolution image, (b) low-resolution input (c) cubic spline interpolation of (b), and (d) super-resolved prototype.

since the majority of pixels on a text page is background. There is a second peak at μ_{black}, representing the text. Additionally, there are a small number of gray values occurring between the two peaks, which represent the gray pixels that exist at transitions from white to black. The amount of these intermediate gray levels is related to the amount of blur in the document image. In order to obtain an unblurred image, we wish to obtain a sharp bimodal distribution, pushing the intermediate gray level towards their nearest peaks. To incorporate this property we define the conditional probability as

$$\boldsymbol{\zeta}(y_k|F) = (y_k - \mu_{white})^2(y_k - \mu_{black})^2, \tag{5.8}$$

where $\boldsymbol{\zeta}(y_k|F)$ is the cost of assigning label x_k to pixel y_k, effecting the (bimodal) distribution F, and is referred to as the bimodal cost prior.

We would also like the label x_k to be as close to the gray value y_k, for a pixel. Thus, the conditional probability is defined as

$$\boldsymbol{\phi}(y_k|x_k) = (y_k - x_k)^2\boldsymbol{\zeta}(y_k|F) \tag{5.9}$$

where $\boldsymbol{\phi}(y_k|x_k)$ is the cost of assigning label x_k to pixel y_k, which is referred to as the data cost. We use the edge stopping function to ensure sharp edges. Thus we use Lorentzian edge penalty function [65] which determines the penalty between the two nodes of a MRF:

$$\boldsymbol{\psi}(x_i, x_j) = \log\left\{1 + \frac{1}{2}\frac{(x_i - x_j)^2}{\sigma_L}\right\}, \tag{5.10}$$

where σ_L is called the contrast parameter, which controls the shape of the edge stopping function [65]. The quality of a labeling in general restoration problem is given by an probability estimate in equation 5.1. Thus, unlike Luong et al. [65], we formulate the problem as MRF that provides us with a better optimum. The belief propagation [29] based optimization is both fast and robust for the purpose.

5.6 Experimental Results and Discussions

We conducted extensive experiments that analyze the performance of the algorithms as well as give insights into its working and potential applications. We now discuss some of the quantitative and qualitative results on various input documents.

5.6.1 Restoration of Degraded Documents

We have carried out a variety of restoration experiments with different document images and differing levels of degradation. For the first experiment, we selected a degraded English book containing 40 pages with close to $50,000$ words and $237,000$ characters. The pages of the book were scanned using an HP flatbed scanner at 600dpi. A document model was learned for the complete book after segmentation, and restoration was performed on all the pages. Figure 5.6 shows a selection of 10 words from the book containing cuts, merges, blobs and erosion artifacts, along with the restoration output of our algorithm.

The first class of degradations that we notice is ink blobs and smears, as present in the word *surely*, *convening*, *permitting*, etc. We note that our algorithm is able to handle most of the cases very well. Especially, the word *little*, which had three of its characters connected by an ink smear was restored correctly. Sever and minor cuts and erosion were also present in the dataset. For example, the word *several* has a severe cut in the character v and the character m in *imprisonment* is separated into three parts due to erosion. As the overall shape of all the characters are discernible in spite of these degradations, the restoration algorithm is able to replace the noisy regions with the correct ideal patches.

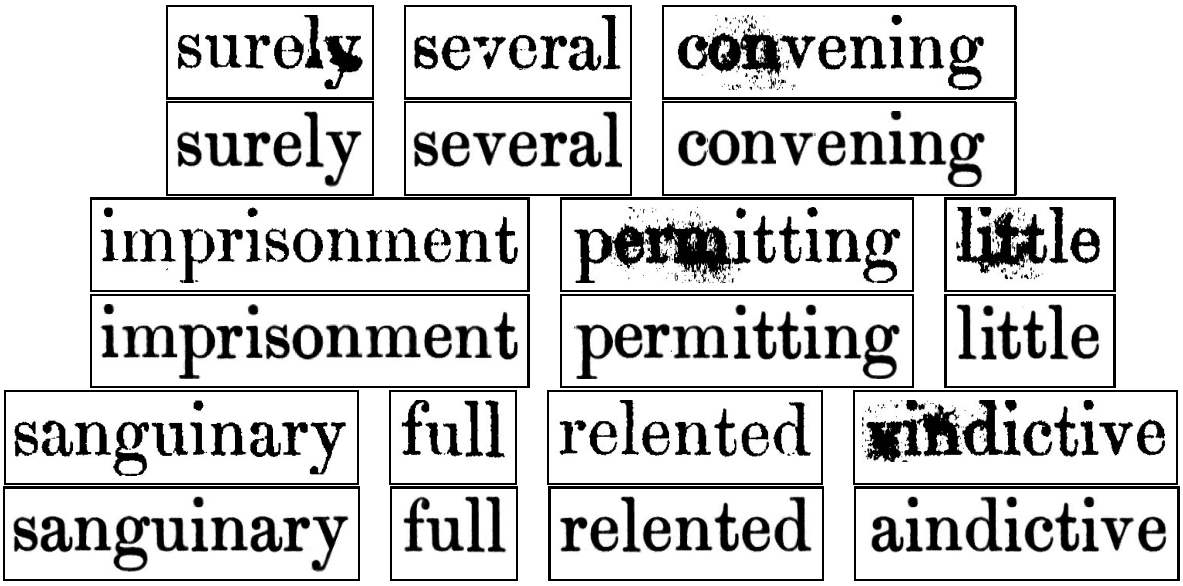


Figure 5.6 Restoration of words containing cuts, merges, blobs and erosion.

However, we note that for the word *vindictive*, the ink smear on the character v is so severe that the resultant patches were not correctly matched. As the restoration always tries to find a set of patches whose neighborhood relations are correct as per the document model, we notice the patch replacements have resulted in replacements by patches of character a.

The restoration should also improve the recognition results of any off-the-shelf OCR system. To verify this, we ran the Tesseract-2.01 OCR from Google on all the pages of the above book, which resulted in an error rate of 3.7%. We note that the modern day OCRs are trained to perform well even in presence of common types of noise, and the accuracy on the original document is already very good. However, after restoration by our proposed algorithm, the error rate further reduced to 1.9%. The accuracy was measured at character level and the book contained $236,861$ characters.

Figure 5.8 shows a portion of an input page and the restored version, along with the OCR results. We note that the recognition of the restored document is in fact highly accurate, and most of the errors are introduced during the rectification process. Two types of errors are of interest here. The first one is due to the erroneous restoration of the word *vindictive*, where the ink-blotted v was restored as an a. The second set of errors is due to missing punctuation marks.

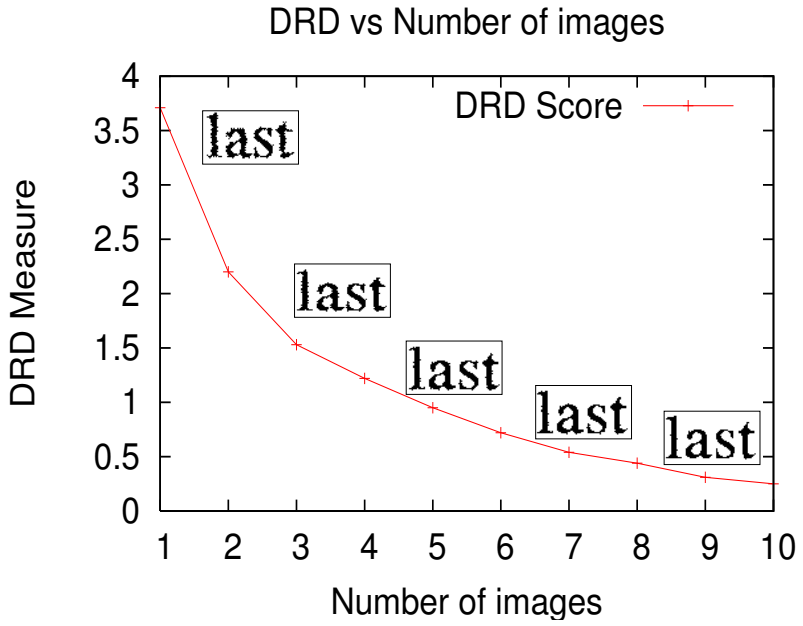


Figure 5.7 Distance-Reciprocal Distortion Measure for a word "last".

This is primarily because of the assumption of heavy noise in documents during prototype learning, which discards small marks. One can tune the restoration to the noise levels present in the document to avoid this.

To study the effect of the size of the document on the restoration results, we analyze the restoration quality with increasing number of prototypes available in each cluster. To measure the restoration quality, we use *distance-reciprocal distance measure* [64], defined as: $DRD = (\sum_{k=1}^{S} DRD_k)/NUBN$, where $NUBN$ is the estimate the nonempty area in the image and DRD_k is the weighted sum of the pixels in the block of the original image that differ from the flipped pixel in the degraded image. We select one word from the above book and plot the DRD score as the number of prototypes used in the learning stage increases. We note that with around 7 prototypes, the DRD score is already very low, which keeps improving further over iterations. We also show a sample restored word if performed at different stages of the learning process for illustration purposes. One can clearly notice the increase in visual quality of the sample word as the number of prototypes in a cluster increases.

5.6.2 Script Independence

As mentioned before, the restoration process does not make the assumption that the document contains a specific font or script. We can perform restorations of multiple scripts using the same approach. As the learning happens at a patch level, and the document priors are generic to text, the algorithm is directly applicable to any script or font. The result of restoration of the proposed algorithm on a document containing Greek text is shown in figure 5.9. Several touching and broken characters are effectively corrected by our restoration algorithm.

5.6.3 Restoration of Vandalized Documents

One of the main strength of the approach is that it models the contents of the document image. This allows us to discard any additions to the document that do not follow the learned document model. We are able to restore even severely degraded and vandalized documents, as long as the actual content is discernible. As the learning is done at a patch level, one can learn the document model from the degraded/vandalized document itself, assuming parts of the document has segmentable patches that can be used for learning. Figure 5.10 shows an example document that has severe scratches/overwriting on the original document. We notice that our approach is able to completely recover the original document. Restoration results of document images from a magazine with degradations and vandalization is given in figure 5.11.

It is interesting to note that even in presence of severe degradations, our algorithm is able to perform extremely well, as long as the overall shape of a patches in a character or its neighbors are visible.

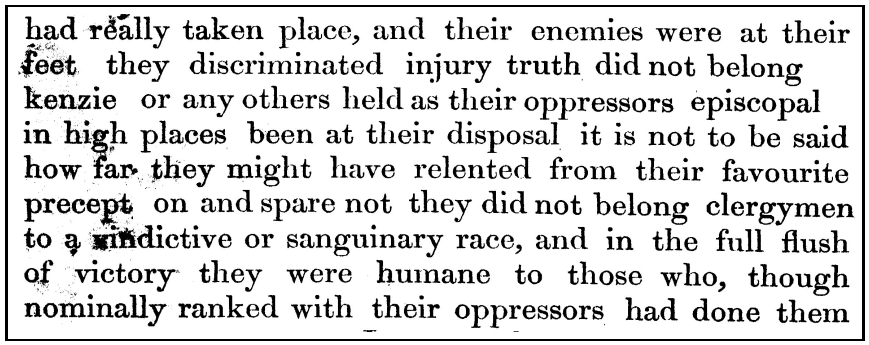
5.6.4 Restoration with Super-resolution

Another aspect of the approach of using multiple degraded prototypes to learn the ideal one is that one can infer super-resolved prototypes for patch models and restoration. Figure 5.12 shows a sample document at 100dpi that is super-resolved to 300dpi. Comparison with the original document scanned at 300 dpi reveals that while the super-resolved text is close to the original, the process has achieved its intended goal of restoration (noise removal) also.

5.7 Summary

We presented a novel approach to document restoration, that builds a tight model of the document content from the input document itself, and uses it to restore severe degradations, including cuts, merges, blobs and erosion. Modeling the document as an MRF on larger patches allows us to use a larger context for restoration. As the approach works on a generic model of the document content, we are able to handle vandalized documents as well as multiple scripts and fonts. The estimation of the content model can also incorporate generation of high quality prototypes, leading to super-resolution of the restored document.

The current approach primarily uses a content model that is learned from the input document. Integration of the approach with a complementary mechanism that models the nature of degradations could further improve the restoration performance. Another potential direction is to combine recognition with restoration in an iterative fashion.


had really taken place, and their enemies were at their feet they discriminated injury truth did not belong kenzie or any others held as their oppressors episcopal in high places been at their disposal it is not to be said how far they might have relented from their favourite precept on and spare not they did not belong clergymen to a indictive or sanguinary race, and in the full flush of victory they were humane to those who, though nominally ranked with their oppressors had done them

(a) Input Paragraph

hadrreally taken place, and their enemies were at their **$$$$** they discriminated injury truth did not belong kenzie or any others held as their oppressors episcopal **inhi** places been at their disposal it is not to be said how **$$$** they might have relented from their favourite precept on and spare not they did not belong clergymen to a **éjgindictive** or sanguinary race, and in the full **Hush $$ pgvictoryr** they were humane to those who, though nominally ranked with their oppressors had done them

(b) OCR Result of (a)

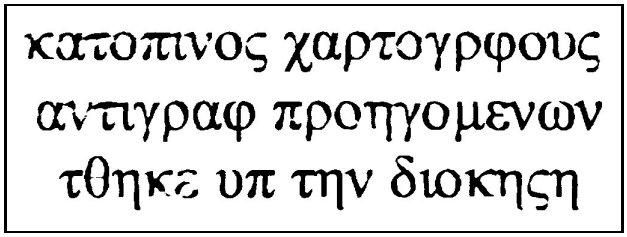
had really taken place and their enemies were at their feet they discriminated injury truth did not belong kenzie or any others held as their oppressors episcopal in high places been at their disposal it is not to be said how far they might have relented from their favourite precept on and spare not they did not belong clergymen to a aindictive or sanguinary race and in the full flush of victory they were humane to those who though nominally ranked with their oppressors had done them

(c) Output Paragraph

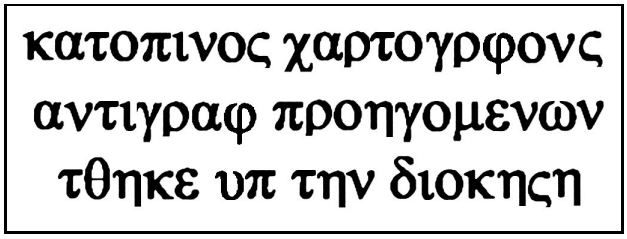
had really taken place and their enemies were at their feet they discriminated injury truth did not belong kenzie or any others held as their oppressors episcopal in high places been at their disposal it is not to be said how far they might have relented from their favourite precept on and spare not. they did not belong clergymen to a **aindictive** or sanguinary race and in the full flush of victory they were humane to those who though nominally ranked with their oppressors had done them

(d) OCR Result of (c)

Figure 5.8 Result on a portion of image from the book.



(a) Input Document



(b) Restored Document

Figure 5.9 Restoration of text in Greek using the proposed approach.

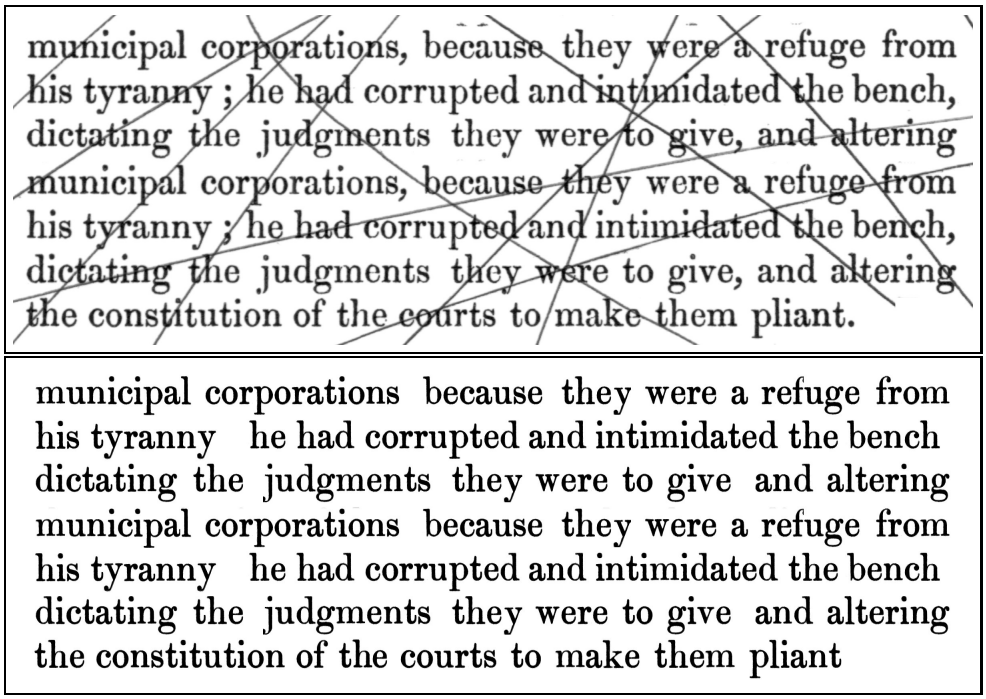
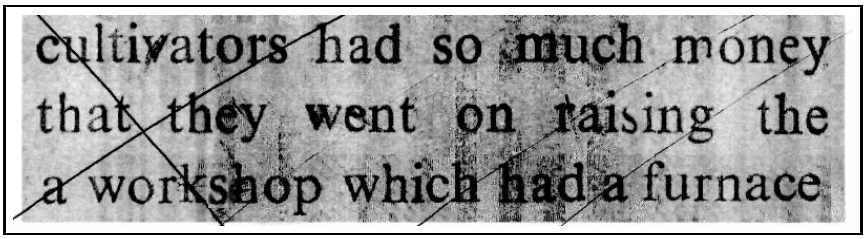
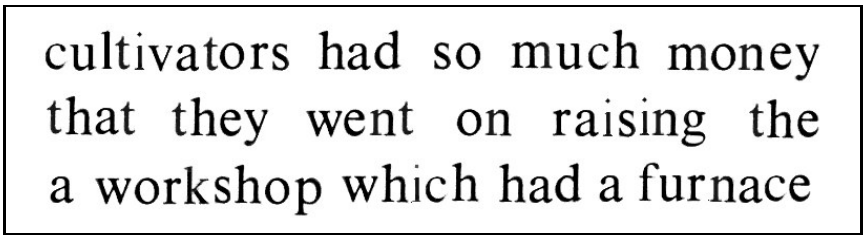


Figure 5.10 Restoration results of a page with overwritten scratches and ink spray marks.

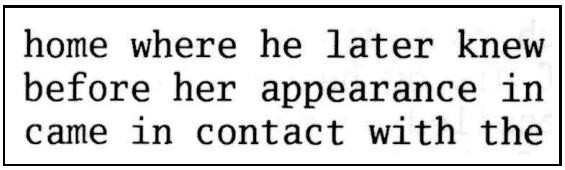


(a) A Magazine Page

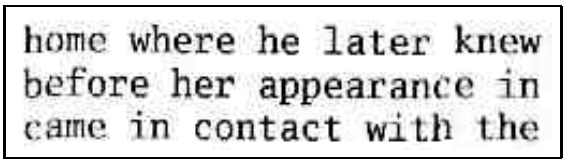


(b) Restoration Result

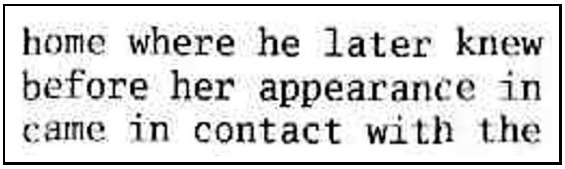
Figure 5.11 Restoration results of document images from a magazine.



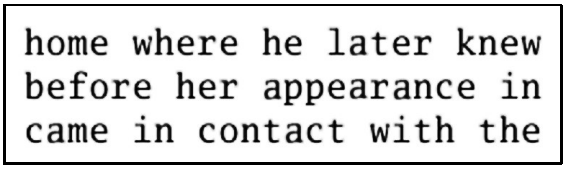
(a) Original High-res. Image (300dpi)



(b) Low-res. Image (100dpi)



(c) Cubic Spline Image (300dpi)



(d) Proposed Method (300dpi)

Figure 5.12 Text super-resolved by a factor of 3 times.

Chapter 6

Conclusions

In this thesis, we have described a robust reconstruction technique to enhance the quality of document images. Image restoration using resolution expansion is important in many areas of image processing. This work introduces a restoration method for low-resolution text images which produces expanded images with improved definition. An implicit edge-directed super-resolution algorithm for document images is proposed. Edge direction information is incorporated in the formulation of the energy function in the MRF model. This technique creates a strongly bimodal image with smooth regions in both the foreground and background, while allowing for sharp discontinuities at the edges. The restored image, which is constrained by the given low-resolution image, is generated by iteratively solving a nonlinear optimization problem. Low-resolution text images restored using this technique are shown to be both quantitatively and qualitatively superior to images expanded using the standard methods. The algorithm is an instance of a general non training based approach that can be useful for document image-processing, that extracts a single high-resolution frame from a single low-resolution image, where the priors are derived from same image.

Exploiting the multiple occurrence of characters brings more information at our disposal, which leads to much better estimates of the unknown pixel values. In order to take advantage of this repetitive behaviour in a practical way, we divide the image into character segments. The character segmentation reduces the computation time drastically in two ways: the algorithm only has to focus on these regions of interests and the search space for possible matching candidates is enormously reduced. Matching between the character segments filters relevant

information before the reconstruction. Information originating from other similar characters are combined and the characters are reconstructed in a Bayesian framework. Results of different experiments show the effectiveness of our proposed technique: characters and symbols are reconstructed very well and OCR results show a significant improvement of our method compared to other reconstruction methods. A trivial extension to our method is to take multiple pages of the same document, journals or book into account or to combine our method with multi-frame restoration techniques (for video applications). This would produce even better results because there is more repetitive information available. The initially restored image is improved by MAP based approach where a suitable a priori information is used to guide the restoration, resolution enhancement being the byproduct. The proposed method can deal with various scripts, and entails relatively simple computation. Through experiments, it has been validated that the proposed method improves both OCR accuracy and image quality. The limitation of this work is that the work is built on top of character segmentation, which can be a bottle-neck and is not a completely solved problem [51].

To overcome this problem we propose an approach to restore severely degraded document images using a probabilistic context model. Unlike traditional approaches that use previously learned prior models to restore the image, we are able to learn the text model from the degraded document itself, making the approach independent of script, font, style, etc. We model the contextual relationship using an MRF. The ability to work with larger patch sizes allows us to deal with severe degradations including cuts, blobs, merges and vandalized documents. This approach can also integrate document restoration and super-resolution into a single framework, thus directly generating high quality images from degraded documents. Experimental results show significant improvement in image quality on document images collected from various sources including magazines and books, comprehensively demonstrate the robustness and adaptability of the approach.

In short, we presented a novel approach to document restoration, that builds a tight model of the document content from the input document itself, and uses it to restore severe degradations, including cuts, merges, blobs and erosion. Modeling the document as an MRF on larger patches allows us to use a larger context for restoration. As the approach works on a generic model of

the document content, we are able to handle vandalized documents as well as multiple scripts and fonts. The estimation of the content model can also incorporate generation of high quality prototypes, leading to super-resolution of the restored document.

The proposed method can also deal with documents irrespective to their exotic font type, it even preserves the font type and is not restricted to characters of a particular alphabet. The strategy of using the repetitive symbol property is not restricted to the reconstruction of document images which suffer from noise, compression artefacts, low resolution scanning, wear processes (e.g. in old manuscripts), etc., but can also be applied in an example-based search engine and combined with an efficient document compression scheme. The latter is useful for the storage of large digital libraries or for transmitting documents. Repetitive characters contain redundant information, this redundancy can be removed for compression by constructing a prototype for each class/cluster of characters and encode the remaining reconstruction errors.

Future Work - The current approach primarily uses a content model that is learned from the input document. Future work could focus on the integration of the approach with a complementary mechanism that models the nature of degradations could further improve the restoration performance. Another potential direction is to combine recognition with restoration in an iterative fashion.

Our work is is an attempt of applying stochastic method to the preprocessing of badly degraded document data. The restriction of our model might be that it is essentially based on document image, but does not handle intense illumination variation, complicated background, and blurring that are common in low resolution video or pictures. However it is possible to generalize the model for more applications. Besides, there are some other issues concerning speeding-up the MRF, training multiple models to deal with different degradations.

The degradations in document images are quite complex in nature. We treat the restoration and recognition as two separate fields. But a overlap might be more effective to extract better results, for example, based the outcome of the recognition stage we can better the restoration process. This may significantly improves the restoration. To simultaneously address restoration and recognition problems for object class specific images could be a good idea. This problem cannot be consistently solved using normal MRFs due to the lack of strong priors

and the computational challenges of learning large datasets. It is also highly unlikely that pure recognition methods would work in the cases of severely blurred images. This work has potentially very interesting extensions. One of them is to overcome the need for manually segmented images by performing the segmentation jointly with recognition and restoration. This would be a potentially significant contribution to the active area of joint recognition and segmentation.

Chapter 7

Related Publications

Conferences

- Jyotirmoy Banerjee, Anoop M. Namboodiri and C.V. Jawahar. **"Contextual Restoration of Severely Degraded Document Images"**. *In the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Jun 22-25, 2009, Miami, US.*
- Jyotirmoy Banerjee, and C.V. Jawahar. **"Super-resolution of Text Images Using Edge-Directed Tangent Field"**. *In DAS, Sep 17-19, 2008, Nara, Japan.*
- Jyotirmoy Banerjee, and C.V. Jawahar. **"Restoration of Document Images Using Bayesian Inference"**. *In NCVPRIPG, Jan 11-13, 2008, Gandhinagar, India.*

Journals

- Jyotirmoy Banerjee, and C.V. Jawahar. **"Restoration of Document Images using Bayesian Inference by Exploiting Repetitive Character Behaviour"**. *Submitted to NASI (under review).*

Awards

- Jyotirmoy Banerjee, and C.V. Jawahar Received Honorable Mentions - Nakano Award in DAS, Sep 17-19, 2008, Nara, Japan, for their work on **"Super-resolution of Text Images Using Edge-Directed Tangent Field"**.

Bibliography

[1] T. Akiyama, N. Miyamoto, M. Oguro, and K. Ogura. Faxed document image restoration method based on local pixel patterns. In *SPIE98*, volume 3305, pages 253–262, Apr. 1998.

[2] B. Allier, N. Bali, and H. Emptoz. Automatic accurate broken character restoration for patrimonial documents. *IJDAR*, 8(4):246–261, September 2006.

[3] H. S. Baird. Document image defect models and their uses. In *ICDAR '93*, pages 62–67, Tsukuba, Japan, Oct 1993.

[4] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1167–1183, 2002.

[5] M. Bern and D. Goldberg. Scanner-model-based document image improvement. In *ICIP00*, pages Vol II: 582–585, 2000.

[6] P. Bochev. A discourse on variational and geometric aspects of stability of discretizations. In H. Deconinck, editor, *33rd Computational Fluid Dynamics Lecture Series*, VKI LS 2003-05, Chaussee de Waterloo, 72, B-1640 Rhode Saint Genese, Belgium, 2003. Von Karman Institute for Fluid Dynamics. 90 pages.

[7] E. Borenstein and S. Ullman. Combined top-down/bottom-up segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(12):2109–2125, 2008.

[8] S. Borman and R. Stevenson. Spatial resolution enhancement of low-resolution image sequences - a comprehensive review with directions for future research. *Technical Report, University of Notre Dame*, 1998.

[9] S. Borman and R. Stevenson. Spatial resolution enhancement of low-resolution image sequences: A comprehensive review with directions for future research. Technical report, Dept. of Electrical Engineering, University of Notre Dame, July 1998.

[10] Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *International Conference on Computer Vision*, volume 1, pages 105–112, 2001.

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

[12] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, November 2001.

[13] H. Cao and V. Govindaraju. Handwritten carbon form preprocessing based on markov random field. In *CVPR07*, pages 1–7, 2007.

[14] D. Capel and A. Zisserman. Super-resolution enhancement of text image sequences. In *ICPR00*, pages Vol I: 600–605, 2000.

[15] R. G. Casey and E. Lecolinet. A survey of methods and strategies in character segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(7):690–706, 1996.

[16] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.

[17] R. Chellappa and A. Jain. *Markov Random Fields: Theory and Applications*. Academic Press, 1993.

[18] D. Chen, K. Shearer, and H. Bourlard. Text enhancement with asymmetric filter for video ocr. In *11th International Conference on Image Analysis and Processing*, pages 192–197, 2001.

[19] G. C. Cross and A. K. Jain. Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(1):25–39, 1983.

[20] G. Dalley, B. Freeman, and J. Marks. Single-frame text super-resolution: a bayesian approach. In *ICIP04*, pages V: 3295–3298, 2004.

[21] K. Donaldson and G. K. Myers. Bayesian super-resolution of text in video with a text-specific bimodal prior. In *CVPR '05 - Volume 1*, pages 1188–1195. IEEE Computer Society, 2005.

[22] K. Donaldson and G. K. Myers. Bayesian super-resolution of text in videowith a text-specific bimodal prior. *IJDAR*, 7(2-3):159–167, 2005.

[23] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.

[24] F. Drira. Towards restoring historic documents degraded over time. In *DIAL '06*, pages 350–357. IEEE Computer Society, 2006.

[25] E. Dubois and A. Pathak. Reduction of bleed-through in scanned manuscript documents. In *PICS*, pages 177–180, 2001.

[26] L. Fan, L. Fan, and C. Tan. Wavelet diffusion for document image denoising. In *ICDAR03*, pages 1188–1192, 2003.

[27] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Advances and challenges in super-resolution. *IJIST*, 14(2):47–57, 2004.

[28] R. Fattal. Image upsampling via imposed edge statistics. *ACM Trans. Graph.*, 26(3):95, 2007.

[29] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54, 2006.

[30] W. Freeman, E. Pasztor, and O. Carmichael. Learning low-level vision. *IJCV*, 40(1):25–47, October 2000.

[31] W. T. Freeman, T. R. Jones, and E. C. Pasztor. Example-based super-resolution. *IEEE Comput. Graph. Appl.*, 22(2):56–65, 2002.

[32] B. Gatos, I. Pratikakis, and S. Perantonis. An adaptive binarization technique for low quality historical documents. In *DAS04*, pages 102–113, 2004.

[33] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[34] R. Gonzalez and R. Woods. *Digital Image Processing*. Prentice Hall, 2002.

[35] W. E. L. Grimson. *From Images to Surfaces: A Computational Study of the Human Early Visual System.* MIT Press, Cambridge, MA, 1981.

[36] M. Gupta, S. Rajaram, N. Petrovic, and T. Huang. Restoration and recognition in a loop. In *CVPR05*, pages I: 638–644, 2005.

[37] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished. Clifford published a simplified veresion of the theorem in 1990, 1971.

[38] R. M. Haralick, H. Joo, C. Lee, X. Zhuang, V. Vaidya, and M. Kim. Pose estimation from corresponding point data. *IEEE Transactions on Systems, Man and Cybernetics*, 19:1426–1446, 1989.

[39] J. D. Hobby and T. K. Ho. Enhancing degraded document images via bitmap clustering and averaging. In *ICDAR '97*, pages 394–400, 1997.

[40] B. K. P. Horn and B. G. Schunk. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[41] Y. Huang, M. S. Brown, and D. Xu. A framework for reducing ink-bleed in old documents. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[42] K. Ikeuchi and B. K. P. Horn. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 17:141–184, 1981.

[43] A. K. Jain. *Fundamentals of Digital Image Processing*. Prentice Hall, 1989.

[44] R. L. Kashyap, R. Chellappa, and A. Khotanzad. Texture classification using features derived from random process models. *Pattern Recognition Letters*, 1:43–50, 1982.

[45] J. G. Kemeny, J. L. Snell, and A. W. Knapp. *Denumerable Markov Chains*. Springer-Verlag, 1976.

[46] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.

[47] K. Kise and D. Doermann. *Second International Workshop on Camera-Based Document Analysis and Recognition*. 2007.

[48] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *European Conference on Computer Vision*, pages 82–96, London, UK, 2002. Springer-Verlag.

[49] S. Kwak, Y. Choi, and K. Chung. Video caption image enhancement for an efficient character recognition. In *ICPR00*, pages Vol II: 606–609, 2000.

[50] V. Kwatra, A. Schodl, I. Essa, G. Turk, and A. Bobick. Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics, SIGGRAPH 2003*, 22(3):277–286, July 2003.

[51] S.-W. Lee, D.-J. Lee, and H.-S. Park. A new methodology for gray-scale character segmentation and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(10):1045–1050, 1996.

[52] Y. Leydier, F. Le Bourgeois, and H. Emptoz. Serialized k-means for adaptative color image segmentation. In *DAS04*, pages 252–263, 2004.

[53] H. Li and D. Doermann. Text enhancement in digital video using multiple frame integration. In *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 19–22, New York, NY, USA, 1999. ACM.

[54] H. Li and D. Doermann. Superresolution-based enhancement of text in digital video. In *ICPR00*, pages Vol I: 847–850, 2000.

[55] O. Li, H. Kia and D. Doermann. Text enhancement in digital video. In *SPIE99*, volume 3651, pages 2–9, 1999.

[56] S. Z. Li. Invariant surface segmentation through energy minimization with discontinuities. *International Journal of Computer Vision*, 5(2):161–194, 1990.

[57] S. Z. Li. Towards 3D vision from range images: An optimization framework and parallel networks. *Computer Vision, Graphics and Image Processing*, 55(3):231–260, 1992.

[58] S. Z. Li. A Markov random field model for object matching under contextual constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 866–869, Seattle, Washington, 1994.

[59] S. Z. Li. *Markov Random Field Modeling in Computer Vision*. Springer-Verlag, 1995.

[60] J. Liang and R. M. Haralick. Document image restoration using binary morphological filters. In *SPIE96*, volume 2660, pages 274–285, 1996.

[61] Z. Lin and H.-Y. Shum. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(1):83–97, 2004.

[62] C. Liu, R. Szeliski, S. B. Kang, C. L. Zitnick, and W. T. Freeman. Automatic estimation and removal of noise from a single image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):299–314, 2008.

[63] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, 1985.

[64] H. Lu, A. Kot, and Y. Shi. Distance-reciprocal distortion measure for binary document images. *SPLetters*, 11(2):228–231, February 2004.

[65] H. Q. Luong and W. Philips. Robust reconstruction of low-resolution document images by exploiting repetitive character behaviour. *International Journal of Document Analysis and Recognition*, 11(1):39–51, 2008.

[66] S. Mallat and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2):617–643, 1992.

[67] C. Mancas-Thillou and M. Mirmehdi. Super-resolution text using the teager filter. In *First International Workshop on Camera-Based Document Analysis and Recognition*, pages 10–16, 2005.

[68] A. Marquina and S. Osher. Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal. *SIAM J. Sci. Comput.*, 22(2):387–405, 2000.

[69] R. Mohan and R. Nevatia. Using perceptual organization to extract 3-d structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:1121–1139, 1989.

[70] M. Motwani, M. Gadiya, R. Motwani, and J. Frederick C. Harris. A survey of image denoising techniques. In *Proceedings of Global Signal Processing Expo and Conference*, Santa Clara Convention Center, Santa Clara, CA., September 2004.

[71] A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Rev.*, 40(3):636–666, 1998.

[72] W. Niblack. *An Introduction to Digital Image Processing*. Prentice Hall, 1986.

[73] H. Nishida. Restoring high-resolution binary images for text enhancement. In *ICIP (2)*, pages 506–509, 2005.

[74] J. Park, Y. Kwon, and J. H. Kim. An example-based prior model for text image super-resolution. In *ICDAR05*, pages I: 374–378, 2005.

[75] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *Signal Processing Magazine*, 20:21–36, 2003.

[76] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann, 1988.

[77] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(7):629–639, 1990.

[78] L. C. Pickup, S. J. Roberts, and A. Zisserman. A sampled texture prior for image super-resolution. In *NIPS*, 2003.

[79] G. Ramponi and P. Fontanot. Enhancing document images with a quadratic filter. *Signal Process.*, 33(1):23–34, 1993.

[80] S. Rice, G. Nagy, and T. Nartker. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer, May 1999.

[81] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[82] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Speech and Audio Processing*, 26:43–49, 1978.

[83] P. Sarkar, H. Baird, and X. Zhang. Training on severely degraded text-line images. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 38–43, 2003.

[84] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith. Video ocr for digital news archive. In *CAIVD '98: Proceedings of the 1998 International Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, page 52, Washington, DC, USA, 1998. IEEE Computer Society.

[85] F. Sattar and D. Tay. Enhancement of document images using multiresolution and fuzzy logic techniques. *Signal Processing Letters*, 6(10):249–252, October 1999.

[86] G. Sharma. Cancellation of show-through in duplex scanning. In *ICIP00*, pages Vol II: 609–612, 2000.

[87] Z. Shi and V. Govindaraju. Character image enhancement by selective region-growing. *Pattern Recogn. Lett.*, 17(5):523–527, 1996.

[88] E. Smigiel, A. Belaid, and H. Hamza. Self-organizing maps and ancient documents. In *DAS04*, pages 125–134, 2004.

[89] G. Strang. *Introduction to Applied Math*. Wellesley-Cambridge Press, 1986.

[90] P. Stubberud, J. Kanai, and V. Kalluri. Adaptive image restoration of text images that contain touching or broken characters. In *ICDAR '95*, pages 778–781, 1995.

[91] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. F. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, 2008.

[92] C. Tan, R. Cao, and P. Shen. Restoration of archival documents using a wavelet technique. *PAMI*, 24(10):1399–1404, October 2002.

[93] P. D. Thouin and C.-I. Chang. A method for restoration of low-resolution document images. *IJDAR*, 2(4):200–210, 2000.

[94] M. E. Tipping and C. M. Bishop. Bayesian image super-resolution. In *NIPS*, pages 1279–1286, 2002.

[95] A. Tonazzini, E. Salerno, M. Mochi, and L. Bedini. Bleed-through removal from degraded documents using a color decorrelation method. In *DAS04*, pages 229–240, 2004.

[96] V. Torre and T. Poggio. On edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(2):147–163, 1986.

[97] O. Veksler. *Efficient graph based energy minimization methods in computer vision*. PhD thesis, Cornell University, 1999.

[98] C. R. Vogel and M. E. Oman. Iterative methods for total variation denoising. *SIAM J. Sci. Comput.*, 17(1):227–238, 1996.

[99] Q. Wang, T. Xia, L. Li, and C. Tan. Document image enhancement using directional wavelet. In *CVPR03*, pages II: 534–539, 2003.

[100] Q. Wang, T. Xia, C. Tan, and L. Li. Directional wavelet approach to remove document image interference. In *ICDAR03*, pages 736–740, 2003.

[101] A. Whichello and H. Yan. Linking broken character borders with variable sized masks to improve recognition. *PR*, 29(8):1429–1435, August 1996.

[102] A. Whichello and H. Yan. Linking broken character borders with variable sized masks to improve recognition. *PR*, 29(8):1429–1435, August 1996.

[103] Q. Zheng and T. Kanungo. Morphological degradation models and their use in document image restoration. In *ICIP01*, pages I: 193–196, 2001.