SUPER-RESOLUTION IMAGING

edited by

Subhasis Chaudhuri

KLUWER ACADEMIC PUBLISHERS
Boston / Dordrecht / London

SUPER-RESOLUTION IMAGING

THE KLUWER INTERNATIONAL SERIES
IN ENGINEERING AND COMPUTER SCIENCE

SUPER-RESOLUTION IMAGING

Edited by
SUBHASIS CHAUDHURI
Department of Electrical Engineering
Indian Institute of Technology - Bombay
Mumbai, India 400 076.

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47004-7
Print ISBN: 0-792-37471-1

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

Contents

This page intentionally left blank

Preface

The field of image processing is now quite a mature one. Many important developments have taken place over the last three or four decades. It finds applications in many diverse areas and a plenty of new applications are being suggested on a regular basis. The bottom line of all image processing applications is that the quality of the input images should be good. Further, the area of interest in the digital picture should be represented at a sufficiently high spatial resolution. One way to increase this resolution is to go for a very high resolution CCD camera which is often not a viable option. Thus, a need for generating a super-resolution image from a set of low resolution observations was felt by the researchers and this book is an outcome of such an effort.

I, as the editor of the book, requested a team of authors to cover a wide range of problems and methods coming under the topic of super-resolution imaging so that the readers get a survey of the present state of the field. I believe that the field super-resolution imaging has reached a first stage of maturity, justifying the timeliness of the book. We primarily concentrate on three different issues in this book – summarization of the existing results, exploration of new ideas, and preparation of ground for further investigations. I hope that the book will become a widely used general reference and that it will motivate further research in this topic and stimulate communication between mathematicians, scientists and practicing engineers. If the book serves to demonstrate that there is a basic need for further exploration in this topic, I would consider my goal of editing this book to have been achieved.

The book is addressed to a broad audience. The senior undergraduate and the graduate students in electrical engineering, computer science and mathematics disciplines may find it a good reference for their courses and seminars on image processing and computer vision. The academicians and the researchers at the universities and various laboratories may find it suitable to supplement their research interest. The practicing engineers will also find the book to be quite useful as the con-

tributing authors have discussed the methodologies in sufficient details so that they can be very easily implemented.

The book being very focused on a particular topic, it is mostly self contained. One does not require a very strong background in image processing to appreciate the contents of the book. However, some knowledge on image restoration would definitely help.

The contributors of this book would be happy if the readers find the book to be useful. I would very much appreciate if readers send me some constructive suggestions.

Acknowledgments

I would like to express my sincere thanks to all authors for readily agreeing to my request to participate in this exercise and for their excellent contributions. It was wonderful that I did not have to send too many reminders to the contributors for their submissions. Thanks are also due to Jennifer Evans and Anne Murray of Kluwer Academic Publishers for their constant help and encouragement. They allowed me liberal extensions of deadlines whenever I needed. I should also thank my student Deepu Rajan for helping me in typesetting the book. Some financial support from the Curriculum Development Programme at IIT Bombay is gratefully acknowledged. My acknowledgments would be incomplete if I do not thank three most important people in my life – Sucharita, Ushasi and Syomantak for their understanding and support during the period of preparation for this book.

It is not very common to write a dedication for an edited book. Yet I could not help but to do so. For the last few years I had to helplessly watch my father surrendering to the wrath of Parkinson's disease. He lost his all forms of communication – the very precious thing he inculcated in me since my childhood days. He breathed his last just three weeks ago. I dedicate this book to all such sufferers of Parkinson's disease.

SUBHASIS CHAUDHURI

Contributing Authors

Kiyoharu Aizawa, Department of Electrical Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
aizawa@hal.t.u-tokyo.ac.jp

Simon Baker, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
simonb@cs.cmu.edu

Terrance E. Boult, Weisman Chair Professor and Chairman, Computer Science and Engineering Department, Lehigh University, Pennsylvania, USA.
tboult@eecs.lehigh.edu

Subhasis Chaudhuri Department of Electrical Engineering, Indian Institute of Technology-Bombay, Powai, Mumbai-400 076. India.
sc@ee.iitb.ac.in

Ming-Chao Chiang, Computer Science and Engineering Department, Lehigh University, Pennsylvania, USA.
mingchao@earthlink.net

Uday B. Desai, Department of Electrical Engineering, Indian Institute of Technology, Bombay, Mumbai 400 076, India.
ubdesai@ee.iitb.ac.in

Nikolas P. Galatsanos, Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA.
npg@ece.iit.edu

Takeo Kanade, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
tk@cs.cmu.edu

Aggelos Katsaggelos, Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208. USA.
aggk@ece.nwu.edu

Narasimha Kaulgud, Department of Electronics and Communications, SJ College of Engineering, Mysore 570 006, India.
kaulgud@sjce.ac.in

Takashi Komatsu, Department of Electrical Engineering, Kanagawa University, 3-27-1 Rokkakubashi,Kanagwa-ku, Yokohama,221-8686,Japan.
komatt01@kanagawa-u.ac.jp

Javier Mateos, Departamento de Ciencias de la Computación, Universidad de Granada, 18071 Granada, Spain.
jmd@decsai.ugr.es

Ross J. Micheals, Computer Science and Engineering Department, Lehigh University, Pennsylvania, USA.
rjm2@eecs.lehigh.edu

Rafael Molina, Departamento de Ciencias de la Computación, Universidad de Granada, 18071 Granada, Spain.
rms@decsai.ugr.es

Shmuel Peleg, School of Computer Science and Engineering, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel.
peleg@cs.huji.ac.il

Deepu Rajan, School of Biomedial Engineering, Indian Institute of Technology-Bombay, Powai, Mumbai-400 076. India.
dr@doe.cusat.edu

Takahiro Saito, Department of Electrical Engineering, Kanagawa University, 3-27-1 Rokkakubashi, Kanagwa-ku, Yokohama,221-8686,Japan.
saitot01@kanagawa-u.ac.jp

C. Andrew Segall, Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA.
asegall@ece.nwu.edu

Brian C. Tom, Center for MR Research, Evanston Northwestern Healthcare, Evanston, IL 60201, USA.
briant@cmr.nunet.net

Assaf Zomet, School of Computer Science and Engineering, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel.
zomet@cs.huji.ac.il

This page intentionally left blank

Chapter 1

INTRODUCTION

Subhasis Chaudhuri
Department of Electrical Engineering
Indian Institute of Technology-Bombay, Powai
Mumbai-400 076, India.
sc@ee.iitb.ac.in

It has been well over three decades now since the first attempts at processing and displaying images by computers. Motivated by the fact that the majority of information received by a human being is visual, it was felt that a successful integration of the ability to process visual information into a system would contribute to enhancing its overall information processing power. Today, image processing techniques are applied to a wide variety of areas such as robotics, industrial inspection, remote sensing, image transmission, medical imaging and surveillance, to name a few. Vision-based guidance is employed to control the motion of a manipulator device so as to move, grasp and then place an object at a desired location. Here the visual component is embedded in the feedback loop in the form of a camera which looks at the scene, a frame grabber which digitizes the analog signal from the camera into image data and a computer which processes these images and sends out appropriate signals to the manipulator actuators to effect the motion. A similar set-up is required for an industrial inspection system such as a fault detection unit for printed circuit boards or for detecting surface faults in machined parts. In remote sensing, multi-spectral sensor systems aboard spacecraft and aircraft are used to measure and record data.

In almost every application, it is desirable to generate an image that has a very high resolution. Thus, a high resolution image could contribute to a better classification of regions in a multi-spectral image or to a more accurate localization of a tumor in a medical image or could facilitate a more pleasing view in high definition televisions (HDTV) or web-based images. The resolution of an image is dependent on the resolution of the image acquisition device. However, as the resolution of the image generated by a sensor increases, so does the cost of the sensor and hence it may not be an affordable solution. The question we ask in this book is that given the resolution of an image sensor, is there any algorithmic way of enhancing the resolution of the camera? The answer is definitely affirmative and we discuss various such ways of enhancing the image resolution in subsequent chapters. Before we proceed, we first define and explain the concept of resolution in an image in the remainder of the chapter.

1. The Word Resolution

Resolution is perhaps a confusing term in describing the characteristics of a visual image since it has a large number of competing terms and definitions. In its simplest form, *image resolution* is defined as the smallest discernible or measurable detail in a visual presentation. Researchers in optics define resolution in terms of the modulation transfer function (MTF) computed as the modulus or magnitude of the optical transfer function (OTF). MTF is used not only to give a resolution limit at a single point, but also to characterize the response of the optical system to an arbitrary input [1]. On the other hand, researchers in digital image processing and computer vision use the term resolution in three different ways.

- *Spatial resolution* refers to the spacing of pixels in an image and is measured in pixels per inch (ppi). The higher the spatial resolution, the greater the number of pixels in the image and correspondingly, the smaller the size of individual pixels will be. This allows for more detail and subtle color transitions in an image. The spatial resolution of a display device is often expressed in terms of dots per inch (dpi) and it refers to the size of the individual spots created by the device.

- *Brightness resolution* refers to the number of brightness levels that can be recorded at any given pixel. This relates to the quantization of the light energy collected at a charge-coupled device (CCD) element. A more appropriate term for this is quantization level.

The brightness resolution for monochrome images is usually 256 implying that one level is represented by 8 bits. For full color images, at least 24 bits are used to represent one brightness level, i.e., 8 bits per color plane (red, green, blue).

- *Temporal resolution* refers to the number of frames captured per second and is also commonly known as the frame rate. It is related to the amount of perceptible motion between the frames. Higher frame rates result in less smearing due to movements in the scene. The lower limit on the temporal resolution is directly proportional to the expected motion during two subsequent frames. The typical frame rate suitable for a pleasing view is about 25 frames per second or above.

In this book, the term resolution unequivocally refers to the spatial resolution, and the process of obtaining a high resolution image from a set of low resolution observations is called super-resolution imaging.

2. Illustration of Resolution

Modern imaging sensors are based on the principle of charge-coupled devices [2] that respond to light sources. A sensor with a high density of photo-detectors captures images at a high spatial resolution. But a sensor with few photo-detectors produces a low resolution image leading to pixelization where individual pixels are seen with the naked eye. This follows from the sampling theorem according to which the spatial resolution is limited by the spatial sampling rate, i.e., the number of photo-detectors per unit length along a particular direction. Another factor that limits the resolution is the photo-detector's size. One could think of reducing the area of each photo-detector in order to increase the number of pixels. But as the pixel size decreases, the image quality is degraded due to the enhancement of shot noise. It has been estimated that the minimum size of a photodetector should be approximately $50\mu m^2$ [3]. This limit has already been attained by current charge-coupled device (CCD) technology. These limitations cause the point spread function (PSF) of a point source to be blurred. On the other hand, if the sampling rate is too low, the image gets distorted due to aliasing.

Consider a pin hole model of a camera which focuses an object of length a at a distance u onto the image plane which is at a distance f from the pin-hole (see Figure 1.1). Assume a square detector array of side x *mm* containing N^2 pixels. If the field of view is described by the angle α in Figure 1.1, then

$$\tan\alpha = \frac{a}{2u} = \frac{x}{2f}.$$

For $x = 10$ *mm* and $N = 512$, we have a resolution of about 51 pixels/*mm* which can focus objects at a distance $u = \frac{fa}{10}$. However, as the object is moved closer to the camera to the new position indicated by the dotted line, for the same field of view, the same number of pixels on the imaging plane are now used to represent only a fraction of the earlier object. Hence, one has a higher resolution representation of the same (or part of the) scene.

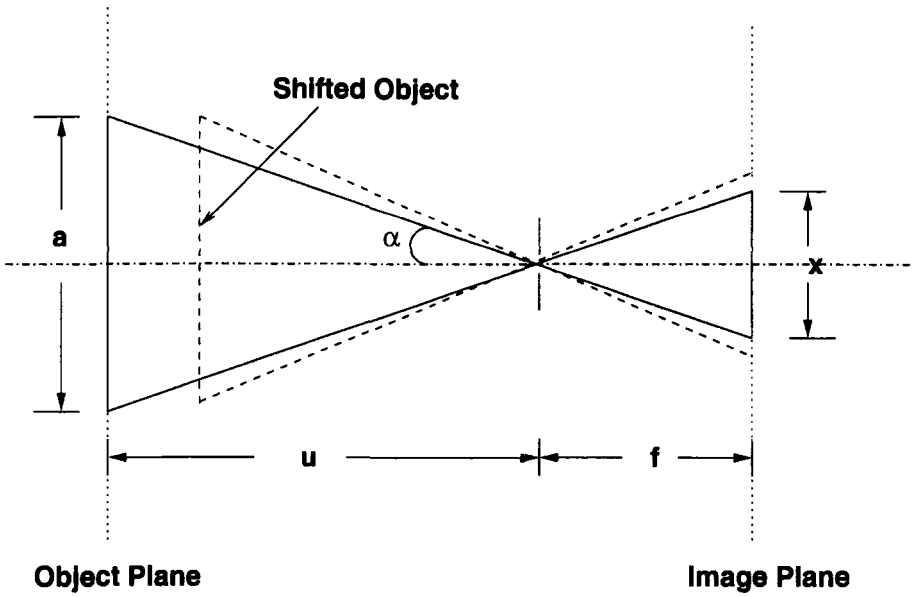


Figure 1.1. The concept of *spatial resolution* illustrated for a pin-hole camera.

We can also explain the limit to the resolution of an image from the principle of optics. The total amount of light energy which enters the optical system is limited by a physically real pupil or aperture that exists somewhere in the optical system. If this limiting pupil is described as an aperture function $a(x,y)$, then the OTF $H(u,v)$ is the auto-correlation of the aperture function [4], i.e.,

$$H(u,v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a(x,y)a(x+u,y+v)dx\ dy.$$

While within the aperture, transmission is perfect and $a(x,y) = 1$, outside the aperture the transmission $a(x,y) = 0$ and no wave can propagate. Thus the OTF goes to zero outside of a boundary that is defined from the auto-correlation of the aperture function and all spatial frequency information outside the region of support is lost. The limit to

the maximum spatial frequency that can pass through the aperture and form the image is given by $f_c = \frac{\lambda F_l}{D}$, where λ is the wavelength of the light, F_l is the focal length of the optics, D is the diameter of the circular limiting aperture and f_c is the spatial cut off frequency.

3. Image Zooming

Mere re-sizing of the image does not translate into an increase in resolution. In fact, re-sizing should be accompanied by approximations for frequencies higher than those representable at the original size and at a higher signal to noise ratio. We may call the process of re-sizing for the purpose of increasing the resolution as *upsampling* or image zooming. The traditional method of upsampling has been to use interpolating functions wherein the original data is fit with a continuous function (strictly speaking, this is called interpolation) and then resampled at a finer sampling grid. In implementing resampling, interpolation and sampling are often combined so that the signal is interpolated at only those points which will be sampled [5]. Sampling the interpolated image is equivalent to interpolating with a sampled interpolating function.

The simplest interpolation algorithm is the so-called nearest neighbor algorithm or a zero-order hold where each unknown pixel is given the value of the sample closest to it. But this method tends to produce images with a blocky appearance. More satisfactory results can be obtained with bilinear interpolation or by using small kernel cubic convolution techniques [6]. Smoother reconstructions are possible using bicubic spline interpolation [7] and higher order splines in general. See [8, 9] and [10] and references therein for more recent literature on image interpolation.

The quality of the interpolated image generated by any of the single input image interpolation algorithms is inherently limited by the amount of data available in the image. Image zooming cannot produce the high frequency components lost during the low resolution sampling process unless a suitable model for zooming can be established. Because of this reason image zooming methods are not considered as super-resolution imaging techniques. To achieve further improvements in this area, the next step requires the investigation of multi-input data sets in which additional data constraints from several observations of the same scene can be used. Fusion of information from various observations of the same scene allows us a super-resolved reconstruction of the scene.

4. Super-Resolution Restoration

The phenomenon of aliasing which occurs when the sampling rate is too low results in distortion in the details of an image, especially at the edges. In addition, there is loss of high-frequency detail due to the low resolution point spread function (PSF) and the optical blurring due to motion or out-of-focus. Super-resolution involves simultaneous up-conversion of the input sampling lattice and reduction or elimination of aliasing and blurring. One way to increase the sampling rate is to increase the number of photo-detectors and to decrease their size thereby increasing their density in the sensor. But there is a limit to which this can be done beyond which the shot noise degrades image quality. Also, most of the currently available high resolution sensors are very expensive. Hence, sensor modification is not always a feasible option. Therefore, we resort to image processing techniques to enhance the resolution. Super-resolution from a single observed image is a highly ill-posed problem since there may exist infinitely many expanded images which are consistent with the original data. Although single input super-resolution yields images that are sharper than what can be obtained by linear shift invariant interpolation filters, it does not attempt to remove either the aliasing or blurring present in the observation due to low resolution sampling. In order to increase the sampling rate, more samples of the image are needed. The most obvious method seems to be to capture multiple images of the scene through sub-pixel motion of the camera. In some cases, such images are readily available, e.g., a Landsat satellite takes pictures over the same area on the ground every 18 days as it orbits around the earth.

With the availability of frame grabbers capable of acquiring multiple frames of a scene (video), super-resolution is largely known as a technique whereby multi-frame motion is used to overcome the inherent resolution limitations of a low resolution camera system. Such a technique is a better posed problem since each low resolution observation from neighboring frames potentially contains novel information about the desired high-resolution image. Most of the super-resolution image reconstruction methods consist of three basic components : (i) motion compensation (ii) interpolation and (iii) blur and noise removal. Motion compensation is used to map the motion from all available low resolution frames to a common reference frame. The motion field can be modeled in terms of motion vectors or as affine transformations. The second component refers to mapping the motion-compensated pixels onto a super-resolution grid. The third component is needed to remove the sensor and optical blurring.

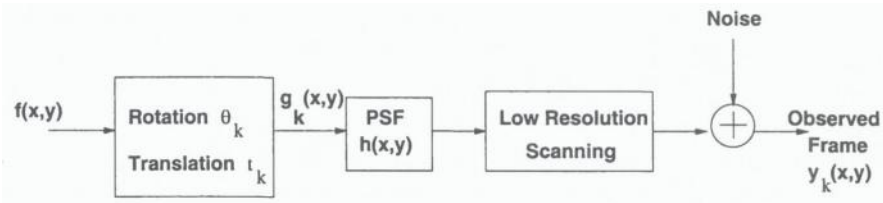


Figure 1.2. Observation model relating a high resolution image to the low resolution observed frames.

The observation model relating a high resolution image to the low resolution observed frames is shown in Figure 1.2. The input signal $f(x,y)$ denotes the continuous (high resolution) image in the focal plane co-ordinate system (x,y). Motion is modeled as a pure rotation $\boldsymbol{\theta_k}$ and a translation $\boldsymbol{t_k}$. The shifts are defined in terms of low-resolution pixel spacings. This step requires interpolation since the sampling grid changes in the geometric transformation. Next the effects of the physical dimensions of the low resolution sensor (i.e., blur due to integration over the surface area) and the optical blur (i.e., out-of-focus blur) are modeled as the convolution of $\boldsymbol{g_k(x,y)}$ with the blurring kernel $h(x,y)$. Finally, the transformed image undergoes low-resolution scanning followed by addition of noise yielding the low resolution $\boldsymbol{k^{th}}$ frame/observation $\boldsymbol{y_k(x,y)}$.

Most of the multi-frame methods for super-resolution proposed in the literature are in the form of a three-stage registration, interpolation, and restoration algorithm. They are based on the assumption that all pixels from available frames can be mapped back onto the reference frame, based on the motion vector information, to obtain an upsampled frame. Next, in order to obtain a uniformly spaced upsampled image, interpolation onto a uniform sampling grid is done. Finally, image restoration is applied to the upsampled image to remove the effect of sensor PSF blur and noise. The block diagram of constructing a high resolution frame from multiple low resolution frames is shown in Figure 1.3. Here, the

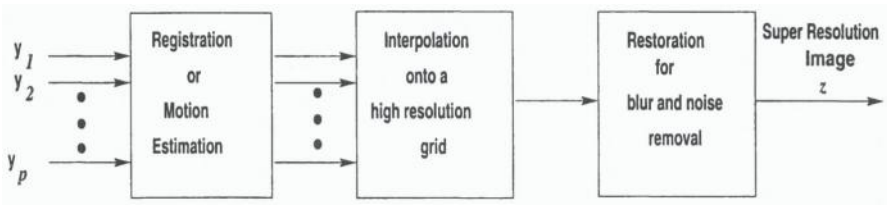


Figure 1.3. Scheme for super-resolution from multi-frame, shifted observations.

low resolution frames $y_1, y_2, \ldots, y_p$ are input to the motion estimation or registration module, following which the registered image is interpolated onto a high resolution grid. Post-processing of the interpolated image through blur and noise removal algorithms results in the generation of a super-resolution image. As discussed in subsequent chapters in this book, other cues such as the relative blurring between observations can also be used in generating the super-resolution images.

5. Earlier Work

The literature on super-resolution can be broadly divided into methods employed for still images and those for video. Most of the research in still images involves an image sequence containing sub-pixel shifts among the images. Although some of the techniques for super-resolution video are extensions of their still image counterpart, a few different approaches have also been proposed. In this section, we briefly review the available literature for generation of super-resolution images or frames from still images or a video sequence. Further references can be found in other chapters in the book as per their contextual relevance to the topic discussed therein.

5.1. Super-resolution from still images

Tsai and Huang [11] were the first to address the problem of reconstructing a high resolution image from a sequence of low resolution undersampled images. They assume a purely translational motion and solve the dual problem of registration and restoration - the former implies estimating the relative shifts between the observations and the latter implies estimating samples on a uniform gird with a higher sampling rate. Note that their observations are free from degradation and noise. Thus, the restoration part is actually an interpolation problem dealing with non-uniform sampling. Their frequency domain method exploits the relationship between the continuous and the discrete Fourier transforms of the undersampled frames. Kim *et al*. extend this approach to include noise and blur in the low resolution observations and develop an algorithm based on a weighted recursive least squares theory [12]. This method is further refined by Kim and Su who consider the case of different blurs in each of the low resolution observations and use Tikhonov regularization to determine the solution of an inconsistent set of linear equations [13].

Ur and Gross utilize the generalized multichannel sampling theorem of Papoulis [14] and Brown [15] to perform a non-uniform interpolation of an ensemble of spatially shifted low resolution pictures. This is fol-

lowed by a deblurring process. The relative shifts of the input pictures are assumed to be known precisely. Another registration-interpolation method for super-resolution from sub-pixel translated observations is described in [16]. Irani and Peleg describe a method based on the principle of reconstruction of a 2D object from its 1D projections in computer aided tomography [17]. Whereas in tomography, images are reconstructed from their projections in many directions, in the super-resolution case, each low resolution pixel is a "projection" of a region in the scene whose size is determined by the imaging blur. Here image registration is carried out using the method described in [18] followed by an iterative super-resolution algorithm in which the error between the set of observed low resolution images and those obtained by simulating the low resolution images from the reconstructed high resolution image is minimized. Since registration is done independently of the high resolution image reconstruction, the accuracy of the method depends largely on the accuracy of estimated shifts. The sub-pixel registration method in [18] looks at two frames as two functions related through the horizontal and vertical shifts and the rotation angle. A Taylor series expansion of the original frame is carried out in terms of the motion parameters and an error function is minimized by computing its derivatives with respect to the motion parameters.

The interdependence of registration, interpolation and restoration has been taken into account by Tom and Katsaggelos in [19] where the problem is posed as a maximum likelihood (ML) estimation problem which is solved by the expectation-maximization (EM) algorithm. The problem is cast in a multi-channel framework in which the equation describing the formation of the low resolution image contains shifts, blur and noise variables. The structure of the matrices involved in the objective function enables efficient computation in the frequency domain. The ML estimation problem then solves the sub-pixel shifts, the noise variances of each image, and the high resolution image. In [3], Komatsu *et al.* use a non-uniform sampling theorem proposed by Clark *et al.* [20] to transform non-uniformly spaced samples acquired by multiple cameras onto a single uniform sampling grid. However if the cameras have the same aperture, it imposes severe limitations both in their arrangement and in the configuration of the scene. This difficulty is overcome by using multiple cameras with different apertures. Super-resolution via image warping is described in [21]. The warping characteristics of real lenses is approximated by coupling the degradation model of the imaging system into the integrating resampler [22].

Wirawan *et al.* propose a blind multichannel high resolution image restoration algorithm by using multiple finite impulse response (FIR)

filters [23]. Their two stage process consists of blind multi-input-multi-output (MIMO) deconvolution using FIR filters and blind separation of mixed polyphase components. Due to the downsampling process, each low resolution frame is a linear combination of the polyphase components of the high resolution input image, weighted by the polyphase components of the individual channel impulse response. Accordingly, they pose the problem as the blind 2D deconvolution of a MIMO system driven by polyphase components of a bandlimited signal. Since blind MIMO deconvolution based on second order statistics contains some coherent interdependence, the polyphase components need to be separated after the deconvolution.

Set theoretic estimation of high resolution images was first suggested by Stark and Oskoui in [24] where they used a projection onto convex sets (POCS) formulation. Their method was extended by Tekalp *et al.* to include observation noise [25]. In addition, they observe that the POCS formulation can also be used as a new method for the restoration of spatially variant blurred images. They also show that both the high resolution image reconstruction and the space variant restoration problems can be reduced to the problem of solving a set of simultaneous linear equations, where the system is sparse but not Toeplitz. Calle and Montanvert state the problem of increasing the resolution as an inverse problem of image reduction [26]. The high resolution image must belong to the set of images which best approximates the reduced estimate. The projection of an image onto this set provides one of the possible enlarged images and is called *induction.* Hence the super-resolution problem is addressed by elaborating a regularization model which restores data losses during the enlargement process.

Improved definition image interpolation or super-resolution from a single observed image is described in Schultz and Stevenson [27]. They propose a discontinuity preserving nonlinear image expansion method where the MAP estimation technique optimizes a convex functional. Although they consider both noise-free and noisy images, they exclude any kind of blur in their model. A MAP framework for jointly estimating image registration parameters and the high resolution image is presented by Hardie *et al.* in [28]. The registration parameters, horizontal and vertical shifts in this case, are iteratively updated along with the high resolution image in a cyclic optimization procedure. A two stage process of estimating the registration parameters followed by high resolution image reconstruction with the knowledge of the optical system and the sensor detector array is presented in [29]. The high resolution image estimate is formed by minimizing a regularized cost function based on the observation model. It is also shown that with the

proper choice of tuning parameter, the algorithm exhibits robustness in presence of noise. Both gradient descent and conjugate-gradient descent optimization procedures are used to minimize the cost function. In [30], Baker and Kanade propose an algorithm that learns recognition-based priors for specific classes of scenes and illustrate it on faces and text images. They also show that for large enough magnification factors, the super-resolution reconstruction constraints do not provide much useful information as the magnification increases.

Elad and Feuer propose a unified methodology that combines the three main estimation tools in image restoration, viz., ML estimator, MAP estimator and the set theoretic approach using POCS. The proposed restoration approach is general but assumes explicit knowledge of the blur and the smooth motion constraints. They also propose a hybrid algorithm that combines the benefits of the simple ML estimator and the ability of the POCS to incorporate non ellipsoidal constraints. The hybrid algorithm solves a constrained convex minimization problem, combining all the a *priori* knowledge on the required result into the restoration process. Cheeseman *et al*. applied Bayesian estimation with a Gaussian prior model to the problem of integrating multiple satellite images observed by the Viking orbiter [31]. Some extensions of this method including 3D reconstruction are also presented.

Most super-resolution algorithms proposed in the literature are confined to 2D applications. A 3D version was proposed in [32] where the high resolution albedo of a Lambertian surface was estimated with the knowledge of high resolution height and vice versa. The problem of surface reconstruction has been formulated as that of expectation maximization and has been tackled in a probabilistic framework using a Markov random field (MRF) model. The idea has been extended to the inverse problem of simultaneous reconstruction of albedo and height in [33] using the extension of Papoulis' generalized sampling theorem to *N*-dimensional cases.

As indicated earlier in Section 2, within the optics community, resolution is described in terms of the OTF. This has led to a slightly different definition of super-resolution. In [4], super-resolution is defined as the processing of an image so as to recover object information from beyond the spatial frequency bandwidth of the optical system that formed the image. The physical size of the image remains the same. This can be seen as equivalent to extrapolating the frequency spectrum [34]. The Gerchberg algorithm is one of the earliest algorithms for super-resolution [35]. Here the constraints that exist on the object are imposed on the image in the spatial domain. The modified image is then Fourier transformed after which constraints in the Fourier domain are imposed on

the Fourier data. This constraint would typically come from the knowledge of the Fourier transform below the diffraction limit. The modified Fourier transform is then inverse transformed to the spatial domain. Walsh and Delaney describe a modification to the Gerchberg algorithm, which computes the spatial frequency components above the diffraction limit directly [36]. Shepp and Vardi use an iterative technique based on an ML estimation of the Poisson statistics in the emission of positrons in positron emission tomography. A similar algorithm is proposed by Hunt and Sementilli in which Poisson statistics is assumed and a MAP estimate is iteratively reconstructed. Performances of such super-resolution algorithms have been studied in [37].

In [38], the author has proposed an interesting application of the super-resolution imaging technique. The depth related defocus blur in a real aperture image is used as a natural cue for super-resolution restoration. The concept of depth from defocus [39] has been incorporated in this scheme to recover the unknown space-varying defocus blur. Since the depth is related to the relative blurring in two (or more) observations of the same scene, a dense depth map can also be recovered. The author proposes a method for simultaneous super-resolution MAP estimation of both the image and the depth fields. Both the high resolution intensity and the depth fields have been modeled as separate MRFs and very promising results have been obtained.

5.2. Super-resolution from video

As mentioned earlier, most of the super-resolution algorithms applicable to video are extensions of their single frame counterpart. Irani and Peleg minimize the mean squared error between the observed and simulated images using the back projection method similar to that used in computer aided tomography [40]. This method is the same as the one used by them for super-resolution of a single image from shifted observations. However, here the core issue that is addressed is the accurate computation of image motion. After the motion for different image regions is computed, these regions are enhanced by fusing several successive frames covering the same region. Possible enhancements include improvement of resolution, filling-in occluded regions and reconstruction of transparent object.

Earlier Keren *et al.* minimized the same difference but their minimization method was relatively simple : each pixel was examined in turn, and its value incremented by unity, kept constant, or decreased by unity, so that a global cost function was decreased [18]. Bascle *et al.* optimize a cost function which in addition to the squared difference between the

observed and the simulated low resolution images, contains second order smoothness constraints on the reconstructed image [41]. Also, the simulated low resolution image takes into account motion blur, optical blur as well as signal averaging by each cell of the CCD array due to spatial sampling.

Schultz and Stevenson use the modified hierarchical block matching algorithm to estimate the sub-pixel displacement vectors and then solve the problem of estimating the high resolution frame given a low resolution sequence by formulating it using the MAP estimation, resulting in a constrained optimization problem with unique minimum [42]. This method is similar to their method of image expansion described in [27]. Patti *et al*. propose a complete model of video acquisition with an arbitrary input sampling lattice and a non-zero aperture time [43]. They propose an algorithm based on this model using the theory of POCS to reconstruct a super-resolution video from a low resolution time sequence of images. However, the performance of the proposed POCS-based super-resolution algorithm will ultimately be limited by the effectiveness of the motion estimation and modeling. Of course, this fact is pertinent to any motion based super-resolution algorithm. Eren *et al*. extended the technique in [43] to scenes with multiple moving objects by introducing the concept of validity maps and segmentation maps [44]. Validity maps were introduced to allow robust reconstruction in the presence of errors in motion estimation. Segmentation maps enable object-based processing, which leads to more accurate motion models within each object. In addition, the proposed method is able to address occlusion issues. The super-resolution video enhancement algorithm proposed by Shah and Zakhor also considers the fact that motion estimation used in the reconstruction process will be inaccurate [45]. To this end, their algorithm finds a set of candidate motion estimates instead of a single motion vector for each pixel and then both the luminance and chrominance values are used to compute the dense motion field at a sub-pixel accuracy. The high resolution frame estimate is subsequently generated by a method based on the Landweber algorithm. Hong *et al*. define a multiple input smoothing convex functional and use it to obtain a globally optimal high resolution video sequence [46]. Baker and Kanade propose an algorithm for simultaneous estimation of super-resolution video and optical flow taking as input a conventional video stream. This is shown to be particularly useful for super-resolution of video sequences of faces [47].

6. Organization of the Book

The need for high resolution images in a variety of applications is now established. The development of a particular technique for super-resolution is driven by the ultimate use to which the super-resolved image is put. This volume is a testimony to the success of super-resolution as a method to overcome the inherent limitations of currently available inexpensive image capturing devices. However, as with every growing area of research, much more need to be done for super-resolution to attain full maturity and eventually become part of a commercial product.

In Chapter 2, Kaulgud and Desai discuss the use of wavelets for zooming images. Although zooming of a single image does not strictly fall in the realm of super-resolution, it is nevertheless interesting to study zooming from a wavelet perspective in order to seek pointers towards use of wavelets for super-resolution. The authors use a multi-resolution analysis based on zero trees to estimate the wavelet coefficients at a finer scale after which the inverse wavelet transform is taken to obtain the zoomed image. They extend this method to color images where the K-L transform is used to generate the (monochrome) principal component of the color image which is then zoomed using the multi-resolution technique.

In Chapter 3, Rajan and Chaudhuri develop a method called generalized interpolation and use it to generate super-resolution images. In generalized interpolation, the space containing the original function is decomposed into appropriate subspaces such that the rescaling operation on individual subspaces preserves the properties of the original function. The combined rescaled sub-functions lead us back to the original space containing the interpolated function, possibly with less information loss compared to direct interpolation in the original space. This method is shown to be effective in structure-preserving super-resolution and in super-resolution rendering. In addition, the generalized interpolation is applied to perceptually organized image interpolation and to transparency images.

In Chapter 4, Tom, Galatsanos and Katsaggelos initially reviews the sub-pixel shift based methods for the generation of super-resoled images. The problem is described in both spatial and frequency domains. Finally, they propose expectation-maximization based algorithms that perform the tasks of registration, restoration and regularized interpolation simultaneously. Various experimental results are presented to validate the proposed techniques.

The use of sub-pixel displacements or motion among the low resolution observations has been widely used in algorithms for super-resolution.

However, in such cases the issue of image registration has to be addressed. In Chapter 5, Rajan and Chaudhuri describe a method of generating super-resolution images from a sequence of blurred observations with no relative motion among them. The motivation for using blur as a cue is twofold : (a) the phenomenon of blurring is inherent during the formation of an image and hence it can be seen as a *natural* cue to be exploited, and (b) the pre-requisite of registration among the images is done away with. The super-resolved image is modeled as a Markov random field and a maximum *a posteriori* estimate of the super-resolved image is obtained.

Warping is commonly used in computer graphics. In Chapter 6, Boult, Chiang and Micheals use warping to generate super-resolution images. Their method is based on a concept called integrating resampler whose purpose is to warp the image subject to some constraints. They also suggest that there is a need for evolving a mechanism to quantify the performance of a super-resolution algorithm. To this end, they make an extensive comparison among several variants of their super-resolution algorithm as applied to optical character recognition and face recognition.

In Chapter 7, Komatsu, Aizawa and Saito address the problem of increasing the spatial resolution using multiple cameras with different apertures. The motivation for using multiple apertures stems from the fact that the spatial uniformity in the generated high resolution image in the case of same apertures is guaranteed if and only if multiple cameras are coplanar and the object of imaging is a two-dimensional plate perpendicular to their optical axes. Their super-resolution algorithm consists of an iterative two stage process of registration and reconstruction from non-uniformly spaced samples and is based on the Landweber algorithm.

In Chapter 8, Zomet and Peleg present the super-resolution problem as a system of a large set of sparse linear equations which are solved using the conjugate gradient method. The algorithm is accelerated through the use of basic image operations, instead of multiplication of sparse matrices, to compute the gradient.

As mentioned earlier, the bulk of the effort on generating super-resolution images lies in estimating the sub-pixel shifts. Compare this to the problem of video compression where motion compensation between frames defines a crucial stage. In MPEG video, such a motion information is already available in the bit stream. In Chapter 9, Segall, Katsaggelos, Molina and Mateos explore the possibility of utilizing this information in generating the super-resolved image from the compressed video.

Finally, in Chapter 10, Baker and Kanade ask a fundamental question on an aspect which is the essence of super-resolution : how much extra information is actually added by having more than one image for super-resolution? It is shown analytically that various constraints imposed on the reconstruction stage provide far less useful information as the decimation ratio increases. They also propose a new super-resolution algorithm in which features are extracted from the low resolution image and it is the resolution of these features that are enhanced, leading to a super-resolved image. The performance of the method is evaluated on analyzing face images.

References

[1] E. L. Hall, *Image Processing and Recognition*, Academic Press, New York, 1979.

[2] R. J. Schalkoff, *Digital Image Processing and Computer Vision*, John Wiley, New York, 1989.

[3] T. Komatsu, T. Igarashi, K. Aizawa, and T. Saito, "Very high resolution imaging scheme with multiple different-aperture cameras," *Signal Processing : Image Communication*, vol. 5, pp. 511–526, Dec. 1993.

[4] B. R. Hunt, "Super-resolution of images : algorithms, principles and performance," *International Journal of Imaging Systems and Technology*, vol. 6, pp. 297–304, 1995.

[5] J. A. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolating methods for image resampling," *IEEE Trans. on Medical Imaging*, vol. 2, pp. 31–39, 1983.

[6] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. on Accoust., Speech and Signal Processing*, vol. 29, pp. 1153–1160, 1981.

[7] H. S. Hou and H. C. Andrews, "Cubic splines for image interpolation and digital filtering,," *IEEE Trans. on Accoust., Speech and Signal Processing*, vol. 26, no. 6, pp. 508–517, 1978.

[8] Chulhee Lee, Murray Eden, and Michael Unser, "High quality image resizing using oblique projection operator," *IEEE Trans. on Image Processing*, vol. 7, no. 5, pp. 679–691, 1998.

[9] M. Unser, A. Aldroubi, and M. Eden, "Fast B-spline transforms for continuous image representation and interpolation," *IEEE Trans. on Image Processing*, vol. 13, no. 6, pp. 508–517, 1978.

[10] G. Ramponi, "Warped distance for space-variant linear image interpolation," *IEEE Trans. on Image Processing*, vol. 8, no. 5, pp. 629–639, May 1999.

[11] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," in *Advances in Computer Vision and Image Processsing*, pp. 317–339. JAI Press Inc., 1984.

[12] S. P. Kim, N. K. Bose, and H. M. Valenzuela, "Recursive reconstruction of high resolution image from noisy undersampled multiframes," *IEEE Trans. on Accoustics, Speech and Signal Processing*, vol. 18, no. 6, pp. 1013–1027, June 1990.

[13] S. P. Kim and W.-Y. Su, "Recursive high-resolution reconstruction of blurred multiframe images," *IEEE Trans. on Image Processing*, vol. 2, pp. 534–539, Oct. 1993.

[14] A. Papoulis, "Generalized sampling theorem," *IEEE Trans. on Circuits and Systems*, vol. 24, pp. 652–654, November 1977.

[15] J. L. Brown, "Multi-channel sampling of low pass signals," *IEEE Trans. on Circuits and Systems*, vol. CAS-28, no. 2, pp. 101–106, February 1981.

[16] Lucas J. van Vliet and Cris L. Luengo Hendriks, "Improving spatial resolution in exchange of temporal resolution in aliased image sequences," in *Proc. of 11th Scandinavian Conf. on Image Analysis*, Kaugerlussuaq, Greenland, 1999, pp. 493–499.

[17] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP:Graphical Models and Image Processing*, vol. 53, pp. 231–239, March 1991.

[18] D. Keren, S. Peleg, and R. Brada, "Image sequence enhancement using sub-pixel displacements," in *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, Ann Arbor, USA, 1988, pp. 742–746.

[19] Brian C. Tom and Aggelos K. Katsaggelos, "Reconstruction of a high-resolution image by simultaneous registration, restoration and interpolation of low-resolution images," in *Proc. of Int Conf. Image Processing*, Washington D.C., 1995, pp. 539–542.

[20] J. J. Clark, M. R. Palmer, and P. D. Lawrence, "A transformation method for the reconstruction of functions from non uniformly spaced samples," *IEEE Trans. on Accoutics, Speech and Signal Processing*, vol. 33, pp. 1151–1165, Oct. 1985.

[21] M. C. Chiang and T. E. Boult, "Efficient super-resolution via image warping," *Image and Vision Computing*, vol. 18, pp. 761–771, 2000.

[22] K. M. Fant, "A non aliasing, real time spatial transform technique," *IEEE Computer Graphics and Applications*, vol. 6, no. 1, pp. 71–80, 1986.

[23] Wirawan, Pierre Duhamel, and Henri Maitre, "Multi-channel high resolution blind image restoration," in *Proc. of IEEE ICASSP*, Arizona, USA, 1999, pp. 3229–3232.

[24] H. Stark and P. Oskui, "High-resolution image recovery from image-plane arrays using convex projections," *J. Optical Society of America*, vol. 6, no. 11, pp. 1715–1726, Nov. 1989.

[25] A. M. Tekalp, M. K. Ozkan, and M. I. Sezan, "High resolution image reconstruction from lower-resolution image sequences and space-varying image restoration," in *Proc. ICAASP*, San Francisco,USA, 1992, pp. 169–172.

[26] Didier Calle and Annick Montanvert, "Super-resolution inducing of an image," in *Proc. of IEEE Int. Conf. on Image Processing*, Chicago, USA, 1998, pp. 742–746.

[27] R. R. Schultz and R. L. Stevenson, "A Bayesian approach to image expansion for improved definition," *IEEE Trans. on Image Processing*, vol. 3, no. 3, pp. 233–242, May 1994.

[28] Russel C. Hardie, K. J. Barnard and Ernest E. Armstrong, "Joint MAP registration and high resolution image estimation using a sequence of undersampled images," *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1621–1633, December 1997.

[29] Russel C. Hardie, K. J. Barnard, J. G. Bognar, E. E. Armstrong and E. A. Watson, "Joint high resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system," *Optical Engineering*, vol. 37, no. 1, pp. 247–260, January 1998.

[30] Simon Baker and Takeo Kanade, "Limits on super-resolution and how to break them," in *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, South Carolina, USA, June 2000.

[31] P. Cheeseman, Bob Kanefsky, Richard Kraft, John Stutz, and Robin Hanson, "Super-resolved surface reconstruction from multiple images," Tech. Rep. FIA-94-12, NASA Ames Research Center, Moffet Field, CA, December 1994.

[32] Hassan Shekarforoush, Marc Berthod, Josiane Zerubia and Michael Werman, "Sub-pixel bayesian estimation of albedo and height," *International Journal of Computer Vision*, vol. 19, no. 3, pp. 289–300, 1996.

[33] Hassan Shekarforoush, Marc Berthod, and Josiane Zerubia, "3D super-resolution using Generalized Sampling Expansion," in *Proc. Int. Conf. on Image Processing*, Washington D.C., 1995, pp. 300–303.

[34] A. K. Jain, *Fundamentals of digital image processing*, Prentice-Hall, New Jersey, 1989.

[35] R. W. Gerchberg, "Super-resolution through error energy reduction," *Opt. Acta*, vol. 21, pp. 709–720, 1974.

[36] D. O. Walsh and P. Nielsen-Delaney, "Direct method for super-resolution," *Journal of Optical Soc. of America, Series A*, vol. 11, no. 5, pp. 572–579, 1994.

[37] P. Sementilli, B. Hunt, and M. Nadar, "Analysis of the limit to super-resolution in incoherent imaging," *Journal of Optical Society of America, Series A*, vol. 10, pp. 2265–2276, 1993.

[38] Deepu Rajan, *Some new approaches to generation of super-resolution images*, Ph.D. thesis, School of Biomedical Engineering, Indian Institute of Technology, Bombay, 2001.

[39] S. Chaudhuri and A. N. Rajagopalan, *Depth from defocused images: A real aperture imaging approach*, Springer-Verlag, New York, 1999.

[40] Michal Irani and Shmuel Peleg, "Motion analysis for image enhancement : resolution, occlusion and transparency," *Journal of Visual Communication and Image Representation*, vol. 4, no. 4, pp. 324–335, December 1993.

[41] B. Bascle, A. Blake, and A. Zissermann, "Motion deblurring and super-resolution from an image sequence," in *Proc. of European Conf. on Computer Vision, Cambridge, UK*. Springer-Verlag, 1996.

[42] R. R. Schultz and R. L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. on Image Processing*, vol. 5, pp. 996–1011, June 1996.

[43] Andrew J. Patti, M. Ibrahim Sezan, and A. Murat Tekalp, "Super-resolution video reconstruction with arbitrary sampling lattices and nonzero aperture time," *IEEE Trans. on Image Processing*, vol. 6, no. 8, pp. 1064–1076, August 1997.

[44] P. E. Eren, M. I. Sezan, and A M. Tekalp, "Robust, object based high resolution image reconstruction from low resolution video," *IEEE Trans. on Image Processing*, vol. 6, pp. 1446–1451, Oct. 1997.

[45] N. R. Shah and Avideh Zakhor, "Resolution enhancement of color video sequences," *IEEE Trans. on Image Processing*, vol. 8, no. 6, pp. 879–885, June 1999.

[46] Min-Cheol Hong, Moon Gi Kang, and Aggelos K. Katsaggelos, "A regularized multichannel restoration approach for globally optimal high resolution video sequence," in *Proc. of Conf. on Visual Communications and Image Processing*, San Jose, CA, USA, 1997, pp. 1306–1313.

[47] Simon Baker and Takeo Kanade, "Super-resolution optical flow," Tech. Rep. CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, USA, 1999.

Chapter 2

IMAGE ZOOMING: USE OF WAVELETS

Narasimha Kaulgud
Department of Electronics and Communications
SJ College of Engineering
Mysore 570 006, India
kaulgud@sjce.ac.in

U. B. Desai
Department of Electrical Engineering
Indian Institute of Technology, Bombay
Mumbai 400 076, India
ubdesai@ee.iitb.ac.in

Abstract Here we propose a method to zoom a given image in wavelet domain. We use ideas from multiresolution analysis and zerotree philosophy for image zooming. Wavelet coefficient decay across scales is calculated to estimate wavelet coefficients at finer level. Since this amounts to adding high frequency component, proposed method does not suffer from smoothing effects. Zoomed images are (a) sharper compared to linear interpolation, and (b) less blocky compared to pixel replication. Performance is measured by calculating signal to noise ratio (SNR), and the proposed method gives much better SNR compared to other methods.

Keywords: Wavelets, Multiresolution, Zooming, Zerotree

1. Introduction

Image interpolation or zooming or generation of higher resolution image is one of the important branch of image processing. Much work is being done in this regard even now. The recent IEEE conference on Image Processing (ICIP-2000) had a full section on interpolation. Classical methods include linear interpolation and pixel replication. Linear inter-

polation tries to fit a straight line between two points. This technique leads to blurred image. Pixel replication copies neighboring pixel to the empty location. This technique tends to produce *blocky* images. Approaches like spline and sinc interpolation are proposed to reduce these two extremities. Spline interpolation is inherently a smoothing operation, while sinc produces ripples (the Gibbs phenomenon) in the output image. Researchers have proposed different solutions for the interpolation problem. Schultz and Stevenson [21] propose a Bayesian approach for zooming. In the super-resolution domain, Deepu and Chaudhuri [19] proposes physics based approach. Knox Carey *et al.* [25] proposed wavelet based approach. Jensen and Anastassiou [6] proposes a nonlinear method for image zooming. In this paper, we propose a simple method to estimate high frequency wavelet coefficients to avoid smoothing of the edges. We use the ideas of zerotree coding [22] and multiscale edge characterization [12].

This article is organized as follows: Section 2 gives some background on wavelets, multiresolution analysis (MRA) and KL transform. In Section 3 we overview some of the existing methods. Section 4 discusses the proposed method using MRA, Karhunen Loéve (KL) transform and scaling function based interpolations. Section 5 extends Section 4 to color images. Section 6 presents discussion on simulation results to illustrate superiority of the proposed method. Section 7 provides some concluding remarks.

2. Background

Here we present some background on multiresolution (wavelet) analysis of signals and KL transform.

2.1. Wavelets

Let $L^2(\mathbb{R})$ be the space of all square integrable functions. Then, it has been shown [1]-[2] that there exist a multiresolution analysis of the form: $L^2(\mathbb{R}) = \bigcup_{j\in\mathbf{Z}} V_j$, ($\mathbf{Z}$ is set of integers) where, the subspaces $\{V_j\}$ have the following properties: [1]

1 $V_j \subset V_{j+1}, \qquad V_{-\infty} = \{0\}, \qquad V_{\infty} = L^2(\mathbb{R})$

2 Let $\phi(t) \in L^2(\mathbb{R})$ be a scaling function; then, $V_j = span_k\{\phi_{j,k}(t)$ $k \in \mathbf{Z}\}$ and $\phi_{j,k} = 2^{j/2}\phi(2^j t - k)$ for all $j \in \mathbf{Z}$. As a consequence, $f(t) \in V_j \Leftrightarrow f(2t) \in V_{j+1}$

3 According to *property-2*, we see that $\phi(t) \in V_0$, then $\phi(2t) \in V_1$. Moreover, $V_0 \subset V_1$, therefore, $\phi(t) \in V_1$. Consequently, $\phi(t)$ can

be written as

$$\phi(t) = \sum_k h(k)\sqrt{2}\phi(2t - k) \tag{2.1}$$

where $h(k)$ is the scaling function coefficient

4 Let the orthogonal direct sum decomposition of V_j be $V_j = V_{j-1} \oplus W_{j-1}$. Then we can write

$$L^2(\mathbb{R}) = V_0 \oplus W_0 \oplus W_1 \oplus \ldots \tag{2.2}$$

Moreover, there exists a function $\psi(t)$ (referred to as the wavelet) such that $W_j = span\{\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k),\ \ k \in \mathbf{Z}\}$

5 Since $W_0 \subset V_1$ we can express $\psi(t)$ as

$$\psi(t) = \sum_k h_1(k)\sqrt{2}\phi(2t - k) \tag{2.3}$$

6 Finally, for any $g(t) \in L^2(\mathbb{R})$ we have the decomposition

$$g(t) = \sum_k c_0(k)\phi_{0,k}(t) + \sum_{j=0}^{\infty}\sum_{k=-\infty}^{\infty} d_j(k)\psi_{j,k}(t) \tag{2.4}$$

Coefficients $c_0(k)$ and $d_j(k)$ are calculated (inner product) as:

$$\begin{aligned} c_0(k) &= < g(t), \phi_k(t) > \\ d_j(k) &= < g(t), \psi_{j,k}(t) > \end{aligned}$$

Apart from these, we make use of the following two properties [10]. Let A_2^j be the operator which approximates a signal at a resolution 2^j. Then:

- The approximation signal at a resolution 2^{j+1} contains all the necessary information to compute the same signal at a lower resolution 2^j. This is the causality property.

- The approximation operation is similar at all resolutions. The spaces of approximated functions should thus be derived from one another by scaling each approximated function by the ratio of their resolution values. That is,

$$\forall j \in \mathbf{Z}\ \ f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}$$

2.2. The KL Transform

Let $\{x(n), 1 \leq n \leq N\}$ be a complex random sequence whose auto correlation matrix is **R.** Moreover, let

$$\mathbf{R}\phi_{\mathbf{k}} = \lambda_k \phi_{\mathbf{k}} \qquad 1 \leq k \leq N$$

Then the Karhunen-Loéve transform (KL transform of $x = [x(1), \cdots, x(N)]^T$ is defined as [5]

$$y = \mathbf{\Phi}^{*T} x \qquad \mathbf{\Phi} = [\phi_1, \ldots, \phi_N] \tag{2.5}$$

Moreover the inverse transform is

$$x = \mathbf{\Phi} y = \sum_{k=1}^{N} y(k)\phi_k \tag{2.6}$$

where $y(k)$ is the $k - th$ element of the vector y.

The KL transform orthogonalizes the data, namely,

$$\begin{aligned} E[yy^{*T}] &= \mathbf{\Phi}^{*T}\{E[xx^{*T}]\}\mathbf{\Phi} = \mathbf{\Phi}^{*\mathbf{T}}\mathbf{R}\mathbf{\Phi} = \mathbf{\Lambda} \\ E[y(k)y^*(l)] &= \lambda_k \delta(k-l) \end{aligned} \tag{2.7}$$

If **R** represents the covariance matrix rather than the auto correlation matrix of x, then the sequence $y(k)$ is uncorrelated. The unitary matrix $\mathbf{\Phi}^{*T}$ is called the KL transform matrix and $\mathbf{\Phi}$ is an $N \times N$ unitary matrix, that reduces **R** to its diagonal form. One dimensional KL transform can easily be extended to two dimensions.

3. Some Existing Methods

Interpolation involves filling intermediate values. Most commonly used methods involve placing zeros in the intermediate samples, and then passing through a filter. Different methods of interpolation are attributed to different types of filters. Here we mention some of the popular interpolation methods, paying special attention to wavelet based interpolation techniques. We compare our method with some of these methods.

3.1. Spatial domain methods

- *Pixel replication:* It is a zero-order hold method, where, each pixel along a scan line is repeated once and then each scan line is repeated. Or equivalently, pad rows and columns with zeros and then convolve with the mask

$$\mathbf{H} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Due to replication of pixels, it gives a blocky image.

- *Linear interpolation:* This is basically a first order hold method. Here a rows and columns of the low resolution images are first interleaved with zeros. Then, a straight line is fit along rows, followed by straight line fit along columns. The straight line fits are equivalent to convolving the image with the mask

$$\mathbf{H} = 1/4 \begin{pmatrix} 1/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{pmatrix}$$

origin being the center of the mask. Since it is an averaging filter, linear interpolation tends to produce a blurred image.

- *Spline interpolation:* Spline (cubic spline) interpolation [14],[24] is one of the most popular techniques. The goal is to get an interpolation that is smooth in the first derivative, and continuous in the second derivative. We have used the spline routine of [26].

3.2. Transform domain methods

- *Sinc interpolation:* The sinc interpolation is basically a Fourier transform (FT) based interpolation method. It assumes the signal under consideration to be band limited and thus can be carried out by zero extension of the FT. That is, we take the N point Discrete FT (DFT) of the original sequence, pad it with zeros for $N + 1$ to $2N$, and take $2N$ point inverse DFT. This results in a sequence of $2 \times$ length.

- Instead of choosing the DFT, we can choose Discrete Sine Transform (DST) or Discrete Cosine Transform (DCT)[1] to perform interpolation. Method will be similar to that of using DFT. Martucci [13] proposes a new set of basis for DST and DCT approaches and use convolution-multiplication property of DSTs and DCTs.

3.2.1 Wavelet techniques. Since our focus is on wavelet based interpolation, we will give more emphasis to wavelet based interpolation techniques

In the recent past, wavelets are used for modeling images - particularly the smooth regions [9] [12]. Extension of these works particularly for image zooming can be found in [3][15][16][17][25].

[1] JPEG (http://www.jpeg.org) uses DCT

Grouse *et al* [9] proposes the use of Hidden Markov Model (HMM) for predicting wavelet coefficients over scales. In the training phase, HMM is trained using an image database. They predict *exact* coefficient from the *observed* coefficient of a noisy image, for denoising application. Principle used here is that the coarser scale coefficients are less affected by noise, while the detail coefficients contain most of the noise. The same idea is be extended to image zooming also.

Carey *et al* [25] use the Lipschitz property, namely, near sharp edges, the wavelet coefficients decay exponentially over scale [11] [12]. At each index, an exponential fit over scale was used for wavelet coefficients. If the fit was close enough to exponential, then it was used to predict the detail signal at the finer scale, else data was left unmodified. On a similar basis, Chang *et al* [3] extrapolates the features in textured region as well. Their method extrapolates wavelet transform extrema across scales, and, important singularities are selected. Corresponding extrema across the scales are associated using least squares error criterion.

Nicolier *et al* [17] uses zero-crossings to predict the high frequency coefficients in the wavelet domain. Using 2^{nd} order type wavelets, zero-crossings in the detail coefficients are produced at the location of a step signal. They have shown the result using 9th order B-spline wavelet for the purpose.

3.3. Scaling function based

This method was proposed in [20]. Consider a wavelet ψ; now, if the wavelet in question generates a multiresolution analysis of $L^2(\mathbb{R})$ and $f \in V^j$ for some j, where $V^{n-1} \subset V^n$ is the MRA corresponding to ψ, then one can write

$$f(.) = \sum_{k=-\infty}^{\infty} c_{j,k}\phi_{j;k}(.) \tag{2.8}$$

where ϕ is the scaling function corresponding to the wavelet ψ. Thus, we can express f completely in terms of its scaling function coefficients $c_{j,k}$. Hence, from the given data samples if we can somehow estimate the scaling function coefficients at resolution j then we have solved the problem. Given a MRA of $L^2[0,1]$ with a compactly supported, p times differentiable scaling function ϕ, we want to estimate the scaling function coefficients of a smoothest function $\hat{f}$, at some resolution j. This scaling function passes through the samples of the given function $b_i = f(i/m)$, $i = 1, \ldots, m = 2^k$. We assume that both f and $\hat{f} \in V^j$, for some known $j \geq k$.

Let us denote the j^{th} scale scaling coefficient as $c_{j,l}$. From the above assumption, we can write

$$\hat{f}(.) = \sum_{l=0}^{2^j-1} c_{j,l}\phi_{j;l}(.) \tag{2.9}$$

Then, we have the following conditions on $\hat{f}$.

1

$$\hat{f}(i/2^k) = \sum_{l=0}^{2^j-1} c_{j,l}\phi_{j;l}(i) = f(i/2^k) = b_i \quad i = 1, \ldots, 2^k \tag{2.10}$$

2 $\hat{f}$ should be at least as smooth as f.

Once we have $c_{j,l}$, we can compute the value of $\hat{f}$ at any point using (2.9). Thus, the problem of estimating f is the same as that of estimating $c_{j,l}$. Next, if b and c are vectors such that $b^k = [b_1, b_2, \ldots, b_{2^j}]^T$ and $c^j = [c_{j,0}, c_{j,1}, \ldots, c_{j,2^j-1}]^T$, then (2.10) can be written as

$$b^k = \Phi_k^j . c^j \tag{2.11}$$

Where, Φ is a matrix, given by

$$\Phi_k^j = \begin{pmatrix} \Phi_{j;0}(1) & \Phi_{j;1}(1) & \ldots & \Phi_{j;2^j-1}(1) \\ \Phi_{j;0}(2) & \Phi_{j;1}(2) & \ldots & \Phi_{j;2^j-1}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{j;0}(2^k) & \Phi_{j;1}(2^k) & \ldots & \Phi_{j;2^j-1}(2^k) \end{pmatrix} \tag{2.12}$$

Consider the minimization problem $(\|c^j\|^2 = (c^j)^T c^j)$

$$\min_{\Phi_k^j c^j = b^k} \|c^j\| \tag{2.13}$$

The solution for 2.13 is well known and is given by

$$c^j = \Phi_k^{j*} \psi (\Phi_k^j . \Phi_k^{j*})^{-1} \psi b^k \tag{2.14}$$

Thus we have estimated scaling coefficients of a smoothest possible linear interpolation. The smoothness condition assures visual quality of the signal while interpolating images.

4. Proposed Method

In this section we present a Multiresolution analysis (MRA) based on the approaches for estimating the wavelet coefficients at higher (finer) scales. Some of these results have been reported earlier in[7]-[8].

4.1. MRA method

The basic strategy for zooming is depicted in Fig. (2.1), where X_{avl} is the available low resolution image, while X_{unk} is the (unknown) high resolution image. H and L are appropriate high pass and lowpass filters in the wavelet analysis. Using X_{avl}, we estimate the coefficients required for synthesizing the high resolution signal. Having estimated the coefficients, rest is a standard wavelet synthesis filter bank.

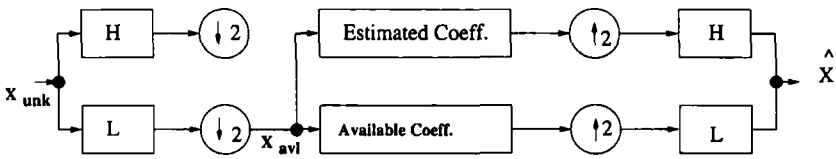


Figure 2.1. Basic zooming strategy

In order to illustrate the estimation of the coefficients in Fig. (2.1), consider Figure (2.2a). We assume that a wavelet transform of an $M \times M$ image composed of boxes 0, *I*, *II*, *IV*, *V*, *VII*, *VIII*, is available and we want to zoom it to size $2M \times 2M$. This would be possible if we can estimate the wavelet coefficients in boxes *III*, *VI* and *IX*. Having estimated these wavelet coefficients, we simply feed these along with the $M \times M$ image to the wavelet based image synthesis filter bank (Fig. 2.2(b)) and obtain the interpolated (zoomed) image of size $2M \times 2M$. We exploit ideas from zerotree concept [22] to estimate the wavelet coefficient in boxes *III, VI* and *IX*. The zerotree concept has the following properties:

- If a wavelet coefficient at a coarser scale is insignificant with respect to a given threshold T, then all wavelet coefficients of the same orientation in same spatial location at finer scales are likely to be insignificant with respect to that T.

- In a multiresolution system, every coefficient at a given scale can be related to a set of coefficients at the next coarser scale of similar orientation.

To estimate the wavelet coefficients at the next finer level, namely, in boxes *III*, *VI* and *IX*, we find the significant wavelet coefficients at two

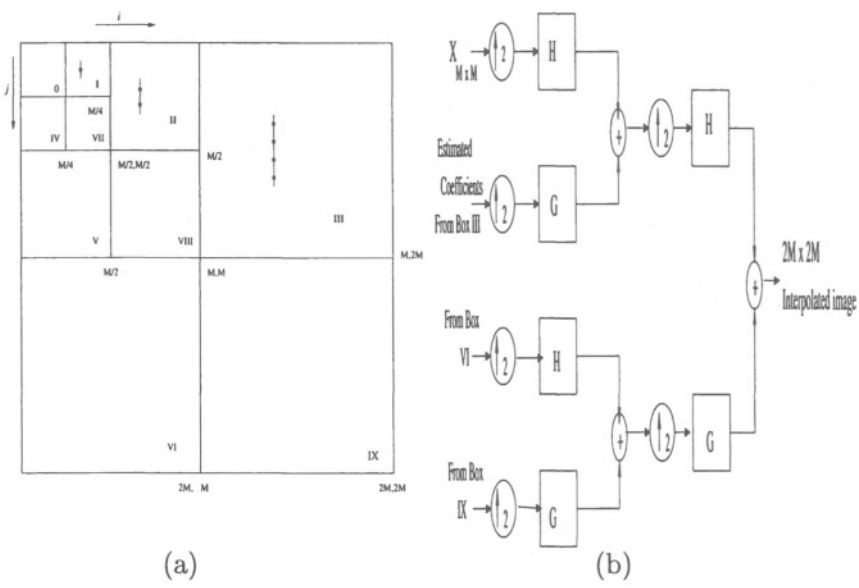


Figure 2.2. Zooming Process using estimated wavelet coefficients (a) estimated wavelet coefficients (b) synthesis filter

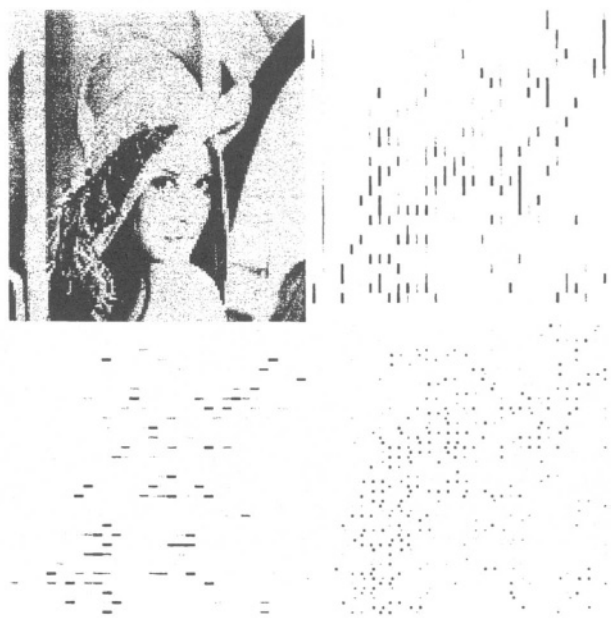
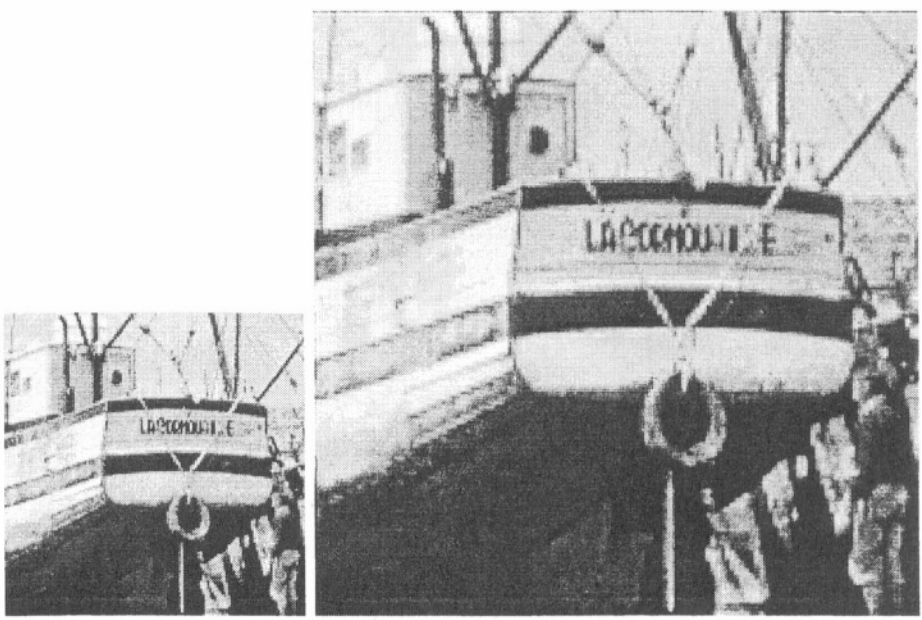


Figure 2.3. Estimated significant wavelet coefficients in box *VI,IX* and *XI* for the Lena image based on MRA zooming scheme

resolutions (namely, in boxes *I* – *II*, *IV* – *V*, *VII* and *VIII*). For example, consider boxes *I* and *II* of Fig(2.2a), with significant coefficients shown as dots. Denote coefficients in respective boxes as $d_1(i_1, j_1) \in I$ and $d_2(i_2, j_2) \in II$. Note that, i_1, j_1 satisfy $M/4 \leq i_1 \leq (M/2) - 1$ and $0 \leq j_1 \leq (M/4) - 1$. Also, i_1 and i_2 are related by $i_1 = \lfloor i_2/2 \rfloor$ ($\lfloor . \rfloor$ represents the floor operator); j_1 and j_2 are similarly related. We have found through empirical studies that the ratio of the coefficients of finer scale (box *II*) the the next coarser scale (box *I*) remains almost invariant. We define $D_{(.)}(i, j)$ as (between boxes *I* and *II*):

$$D_1(i, j) = \frac{d_2(i, j)}{d_1(\lfloor i/2 \rfloor, \lfloor j/2 \rfloor)} \tag{2.15}$$

$$D_2(i, j) = \frac{d_2(i, j+1))}{d_1(\lfloor i/2 \rfloor, \lfloor (j+1)/2 \rfloor)} \tag{2.16}$$



(a) (b)

Figure 2.4. (a) Original low-resolution Boat Image, and (b) MRA Based Zoomed Image.

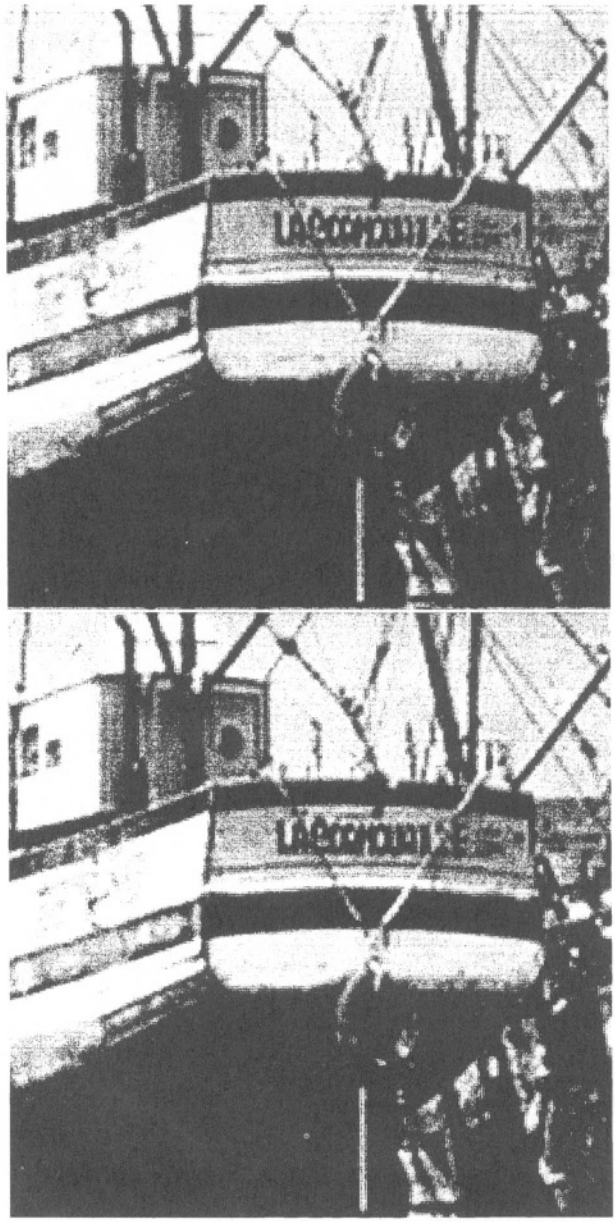


Figure 2.5. Sinc (top) and Spline Based (bottom) zoomed Boat Images

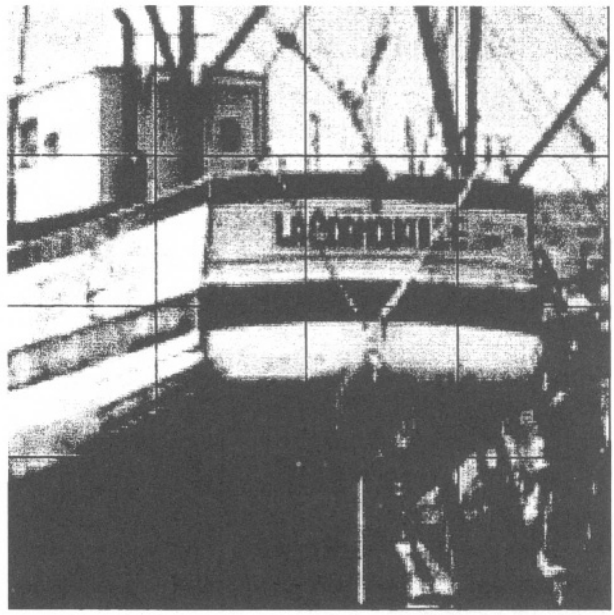


Figure 2.6. Scaling function based zoomed boat image

These $D_{(.)}(i,j)$ values are used to estimate coefficients $\hat{d}$ at the finer scale (box *III*).

$$\begin{aligned}\hat{d}(2i,2j) &= D_1(i,j)d_2(i,j)(1-l_{d(i,j)})\\ \hat{d}(2i,2j+2) &= D_2(i,j)d_2(i,j+1)(1-l_{d(i,j+1)})\end{aligned} \tag{2.17}$$

we set:

$$\begin{aligned}\hat{d}(2i,2j+1) &= \hat{d}(2i,2j)\\ \hat{d}(2i,2j+3) &= \hat{d}(2i,2j+2)\end{aligned} \tag{2.18}$$

$l_{d(i,j)}$ is an indicator function; $l_{d(i,j)}$ is set to zero, if $d(i,j)$ is significant, else to one. We define $d(i,j)$ to be significant if $|d(i,j)| > T$. Note that Eqn. 2.17 implies an exponential decay and this is consistent with what is reported in [3] [12],[25]. Thus, we refer to $D_{(.)}(i,j)$ as the decay parameter.

In principle, for each coefficient $d_2(i,j) \in II$ we should have four coefficients in box *III*. But, our experiments has shown that doing this leads to a "blocky" zoomed image. Hence, we generate only *two* coefficients in box *III* corresponding to $d_2(i,j) \in II$. Moreover, we know that the detail sub-images using the wavelet transform yields vertical lines in boxes *I, II* and *III*; horizontal lines in boxes *IV, V* and *VI*; and diagonal lines

in boxes *VII, VIII* and *IX*. We use this intuition and compute wavelet coefficients along vertical direction in box *III*, along horizontal direction in box *VI*, and along diagonal direction in box *IX* and that too only along alternate lines. For box *III* equation (2 17) and (2.18) hold good. Analogous expressions can easily be obtained for wavelet coefficient in box *IV* and *IX*. Now, the estimated $\hat{d}$'s and the original $M \times M$ image is fed to the wavelet based image synthesizer (Fig2.2b) to obtain the zoomed image which is of twice the size of the given image. In all our simulation the threshold T was selected as half the maximum coefficient in the respective boxes, namely boxes *II*, *V* and *VII*. Fig. (2.3) shows an example of the estimated wavelet coefficients that were used in zooming of the Lena image. DAUB4 wavelet was used for computing the discrete wavelet transform. Pseudocode for the above scheme is:

```
BEGIN
take two level wavelet transform of image(x) size M x M ;
  /* for box II of Fig. 2.2(a) */
      FOR i=M/2 TO M DO
      FOR j = 0 TO M/2 DO
          find the max coefficient in box II ;
      T= max/2 ;
      FOR i=M/2 TO M DO
      FOR j = 0 TO M/2 DO
      BEGIN
          IF (x[i][j] && x[i/2][j/2] > T )
              estimate wavelet coefficients for box II
      END FOR
      /*  Repeat for boxes V and VII **/
      take 3-level inverse wavelet transform of x
                  to get 2M x 2M image ;
END.
```

Results from individual methods shown in the figures 2.4-2.9.

5. Color Images

The proposed method was extended to color images as well. Many color coordinates are reported in literature and are in use [4][18][23][27]. Most of the color coordinates are obtained as a transformation of other coordinate, usually the RGB. Here we discuss the results obtained from YIQ color coordinates. The *Y* component of the image is considered as simple gray scale image and the proposed MRA algorithm was run on it.

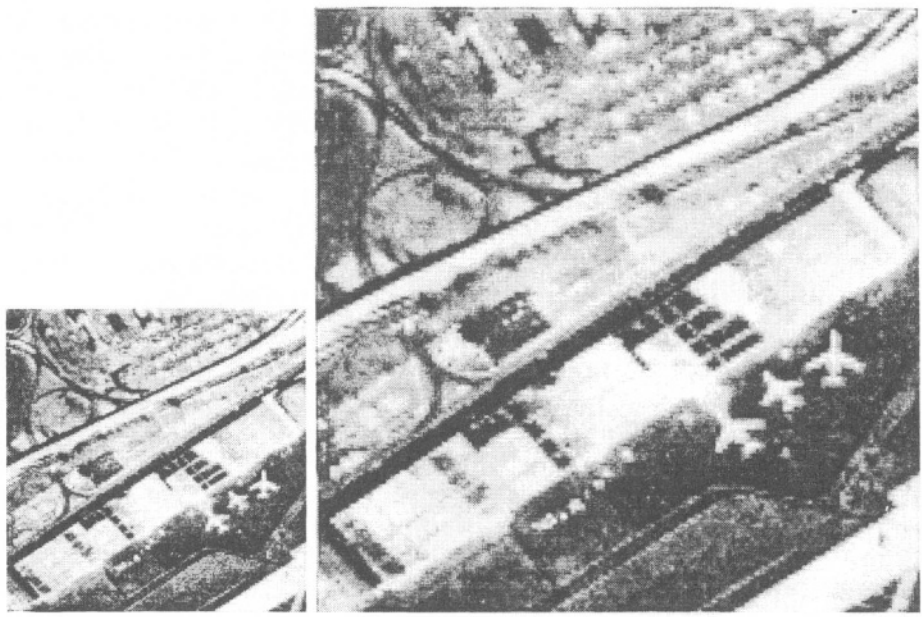


Figure 2.7. (a) Original Airport Image and (b) MRA based zoomed image

For I and Q components, pixel replication and linear interpolation give similar results as any other methods. This is due to the fact that these components are comparatively smooth. Keeping computational complexity in mind, we opted for linear interpolation for zooming I and Q components and MRA based method for zooming Y component. Results for Suzie image is shown in Fig.(2.10 - 2.14). We compare the proposed MRA method with spline interpolation method. For both the methods, I and Q components are linearly interpolated and Y component alone is zoomed by the appropriate methods. Resulting images are shown in (Fig.2.11-2.14). We can readily observe the proposed performing better than spline interpolation method (it is not apparent in the monochrome versions of images, but, for color images, difference is apparent). Image quality may further be improved by pre and post-processing techniques.

5.1. K-L Transform

The KL transform method is used for multi spectral images - color images, in our case. With color image having RGB components, X of equation (2.5) is three-dimensional. Then, the co-variance and $\mathbf{\Phi}$ matrices will be of size 3×3. The covariance matrix $\mathbf{C_x}$ is given by

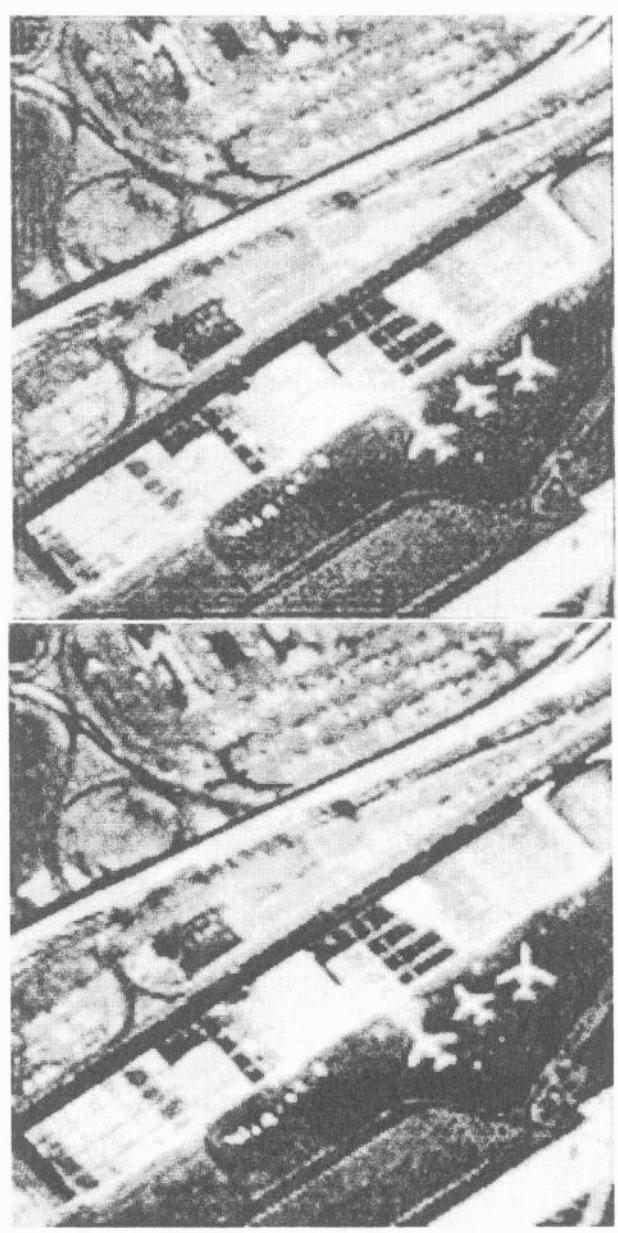


Figure *2.8.* Sinc (top) and Spline based (bottom) zoomed Airport images

Figure 2.9. Scaling function based Zoomed airport images

Figure 2.10. *Y* component of Suzie image: Original and zoomed using MRA

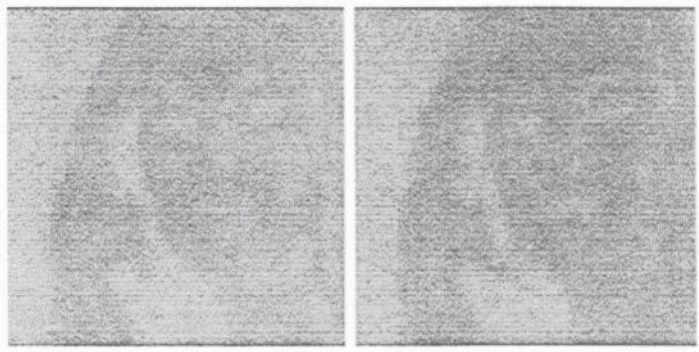


Figure 2.11. *I* component of Suzie image: Original and zoomed using linear interpolation

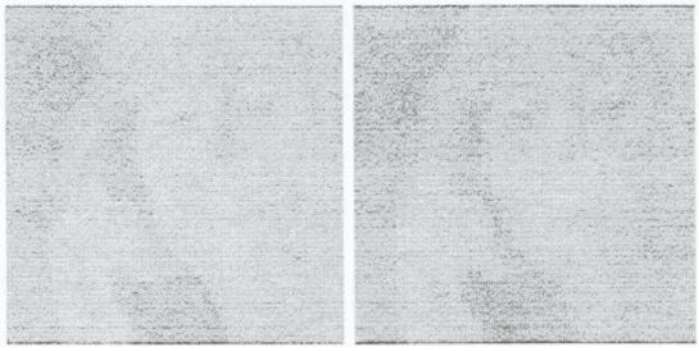


Figure 2.12. *Q* component of Suzie image: Original and zoomed using linear interpolation

Figure 2.13. Spline interpolated Suzie image *Y* component

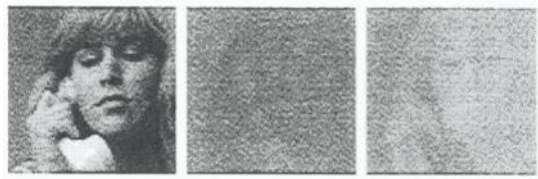


Figure 2.14. Low resolution Suzie image: Y ,I and Q components

$\mathbf{C_x} = \frac{1}{N}\sum_{k=1}^{N} X_k X_k^T - m_x m_x^T$. Where, m_x is the vector corresponding to the means of the individual spectra. Now we have a monochrome image Y generated from the color image X as per Eqn.(2.5). Once we have generated a monochrome image Y, as the principle component of color image X, interpolation of this Y is carried out as described in section (4.1). We can interpolate and retain only the principle component which results in a interpolated monochrome image from a multi-spectral image which is shown in the figure[2](2.15,2.16). Alternatively, we can take the inverse KL transform and combine the three components to get color interpolated image. In such a method, we see that there is a slight deterioration in the contrast of the image. This has to be taken care by proper post processing.

The pseudo code for this method is:

```
BEGIN
    read the RGB component of the image ;
    evaluate the mean for R, G and B component ;
    evaluate the co-variance matrix, and -
       - eigenvectors of co-variance matrix ;
    estimate the principle component as y
    interpolate the principle component y using MRA
        (section 4.1) ;
END.
```

6. Results and Discussion

Results of individual methods are illustrated for two image samples in Figures (2.4-2.9). First set of images Fig. (2.4-2.6) show the results for boat images and the second set namely, Fig.(2.7-2.9), are for airport image. As expected, sinc method produces ripples at the image boundaries, evident from Fig. (2.5a) and (2.8a). According to Fig. (2.5b)

[2]Image courtesy and ©SAC, Ahmedabad, India

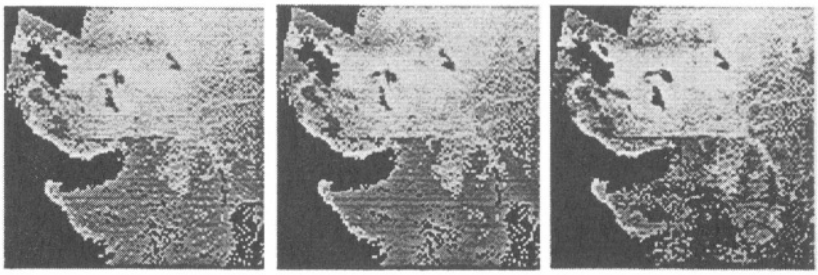


Figure 2.15. Original Gujarat Coast Line - RGB components (enhanced for display purposes).

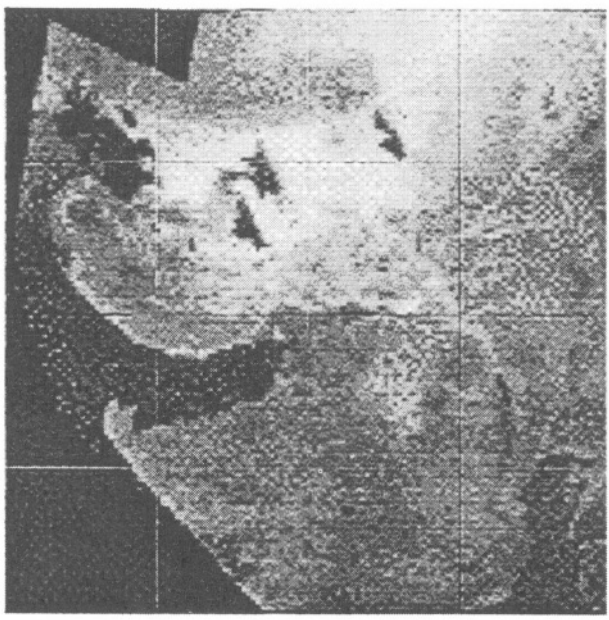


Figure 2.16. Zoomed Gujarat coast image, Using KL Transform (Principal Component)

and (2.8b), visual qualities of spline and scaling function based methods are comparable, and seem to perform better. However, both of them smooth out some sharp edges of original low-resolution images. This can be observed at the rope which is tied to the tyre in boat image (Fig. 2.4b) and the road outside the airport (2.7b).

We evaluate the performance of different techniques by calculating Peak Signal to Noise Ratio (PSNR), which is defined as:

$$PSNR = 10\ log\ \frac{255^2}{(X - \hat{X})^2}$$

where, X is the original image and $\hat{X}$ is the zoomed image. The PSNR values are tabulated in Table(2.1)

Note: To calculate PSNR, a low resolution version of the high resolution image is zoomed. A low resolution image is generated according to the method proposed in [21].

Table 2.1. Comparison of PSNR Values for different methods

Image	*Spline*	*Sinc*	*MRA*	*Scal.fn.*
Boat.	24.97	24.49	29.21	25.72
Airport	23.82	22.88	26.98	24.44
Lena	25.73	24.14	29.80	23.49
Bird	30.89	20.00	33.25	21.41
Einstein	28.28	19.18	30.17	24.51

We observe that the proposed method performs well for various classes of images. This is evident from the PSNR improvements, as per Table (2.1). Even though a little amount of staircase effect is observed in the zoomed image from the proposed method, overall quality of zoomed image is good, as sharp edges of original image are retained (ropes in boat image). To retain these sharp edges, we have used Daubechies DAUB4 wavelet. It is observed that higher wavelets, like DAUB6, tends to smooth the edges and Harr wavelet produces more staircase effect.

For color images, it is seen that color contrast deteriorates slightly while operating on RGB color coordinates. This has to be taken care of by suitable post-processing. In this regard, YIQ color coordinate gives satisfactory results, without the need for pre or post-processing.

We compare the wavelet coefficients obtained by the proposed MRA method with those obtained by the scaling function based method. Results of obtaining coefficients from these two methods is shown in Fig. (2.17). It is observed that coefficients obtained from both the methods are almost the same, and hence, the proposed method is justified. The

proposed method is computationally less taxing than the scaling function based method and hence faster. Visually, scaling function based method performs slightly better.

For zooming up to 8 times (8×), the output suffers from staircase effect. For such cases (8×), spline does a better job. Another limitation with the proposed technique is presence of spurious edges, when there is a rapid change in the gray levels. For checkerboard kind of images, these spurious edges are more predominant. These spurious edges can be attributed to the insertion of high frequency components in the image. This effect can be overcome by some post-processing techniques. For a relatively smooth image, the proposed method performs very well.

7. Conclusion

We have reviewed some of the techniques for image interpolation, paying special attention to the wavelet based zooming methodologies. Basic idea of image zooming in wavelet domain is to estimate the coefficients at the finer scale. We have overviewd some of the techniques reported in literature, to estimate these coefficients.

We have proposed a simple scheme, which is computationally fast. As the proposed scheme is efficient, it can be used for color images and in real-time applications also. We have mentioned the advantages and limitations of the proposed scheme.

Comparing the performance of the proposed technique with some of the conventional approaches, we observe that output images are sharper and there is a good improvement of PSNR with about 3 dB. Thus, the proposed method performs better than conventional approaches, both visually and numerically.

Acknowledgments

This work was supported by a grant from ISRO (India) and IIT-Bombay Cell.

References

[1] C. Sidney Burrus, Ramesh A. Gopinath and Hairao Guo. *"Introduction to wavelets and wavelet transforms"*. Prentice-Hall, New Jersy, 1998.

[2] I. Daubechies. *Ten lectures on wavelets.* SIAM, Philadelphia, Pennsylvania, 1992.

[3] G. Grace Chang, Zoran Cvetkovicand Martin Vetterli. "Resolution enhancement of image using wavelet transform exterma interpola-

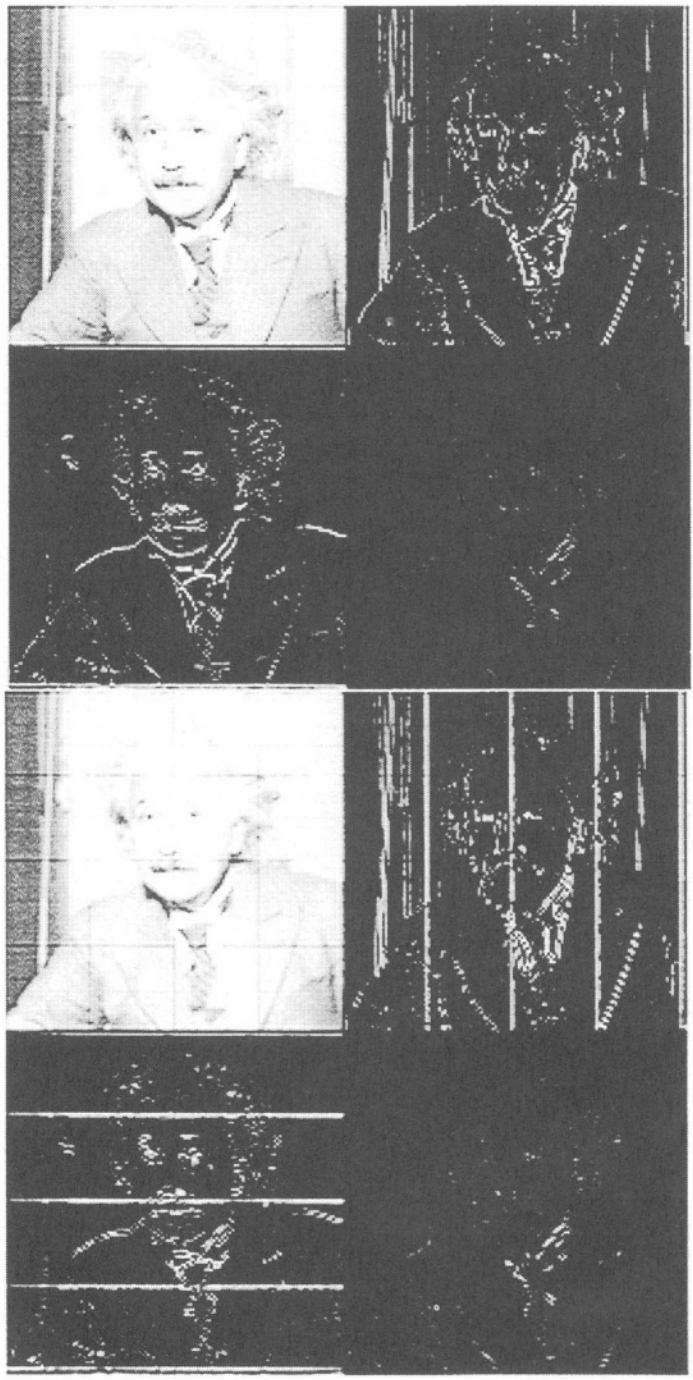


Figure 2.17. Estimated wavelet coefficients: (a) MRA method (b) Scaling function based method

tion. In *"IEEE-ICASSP"*, pages 2379–2383, 1995.

[4] Keith Jack. *"Video demystified"*. High Text, San Diego, 1996.

[5] A. K. Jain. *Fundamentals of Digital Image processing*. PHI, New Delhi, 1995.

[6] Kris Jensen and Dimitris Anastassiou. "Spatial resolution enhancement om images using nonlinear interpolation". In *International Conference on Acoustics and Speech Signal Processing*, pages 2045–2048, 1990.

[7] Narasimha Kaulgud and U. B. Desai. "Wavelet based approaches for image interpolation". *To appear in International Journal of Imaging and Graphics IJIG*.

[8] Narasimha Kaulgud and U. B. Desai. "Joint MRA and MRF based image interpolation". In *"Proceedings of National Conference on Communications NCC-2000,New Delhi, India"*, pages 33–36, Jan. 2000.

[9] M. S. Crouse, R. D. Nowak and R. G. Baranuik. "Wavelet based signal processing using hidden markov models". *IEEE Transactions on Signal Prpcessing*, 46, Sept. 1998.

[10] Stephen G. Mallat. "A theory for multiresolution signal decomposition:The wavelet representation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, July 1989.

[11] Stephen G. Mallat and W. Hwang. "Singularity detection and processing with wavelets". *IEEE Transactions on Information Theory*, 38:617–643, Mar 1992.

[12] Stephen G. Mallat and Sifen Zhong. "Characterization of signals from multiscale edges". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7):710–732, July 1992.

[13] Stephen A. Martucci. "Interpolation in the dst and dct domains ". In *"International Conference on Image Processing ICIP-2000"*, 2000.

[14] Michael Unser, Akram Aldroubiand Murray Eden. "Color information for region segmentation". *IEEE Tx on Image Processing*, 4(3):247–258, March 1995.

[15] D. Darian Muresan and Thomas W. Parks. "Predection of image detail". In *"International Conference on Image Processing ICIP-2000"*, 2000.

[16] Nhat Nguyen and Peyman Milanfar. "An efficient wavelet based algorithm for image superresolution". In *"International Conference on Image Processing ICIP-2000"*, 2000.

[17] Fred Nicolier and Fred Truchetet. "Image maginfication using decimeted orthogonal wavelet transform". In "*International Conference on Image Processing ICIP-2000*", 2000.

[18] Y. Ohta, T. Kanade, and T. Sakai. "Color information for region segmentation". *CVGIP*, 13:222–241, 1980.

[19] Deepu Rajan and S. Chaudhuri. "Physics based approach to generation of super resolution images". In *International Conferance on Vision, Graphics and Image Processing, New Delhi*, pages 250–254, 1998.

[20] Uday Savagoankar. "Improving image resolution by scaling function based interpolation". Master's thesis, EE Dept., IIT, Bombay, India, 1998.

[21] Richard R. Schultz and R. L. Stevenson. "Bayesian approach to image expansion for improved definition". *IEEE Transactions on Image Processing*, 3(3):234–241, May 1994.

[22] Herome M. Shapiro. "Embedded image coding". *IEEE Transactions on Signal Processing*, 41(12):3445–3462, Dec. 1993.

[23] David Travis. *Effective Color Displays: Theory and Practice.* Academy Press, London, 1991.

[24] Michael Unser. "Splines - A perfect fit for signal and imageprocessing". *IEEE Signal Processing Magazine*, pages 22–38, 1999.

[25] W. Knox Carey, Daniel B. Chuang and Sheila S. Hemami. "Regularity preserving image interpolation ". *IEEE Transactions on Image Processing*, 8(9):1293–1297, Sept. 1999.

[26] William Press, Saul Teukolsky, William Vetterling and Brian Flannery. "*Numerical Recipes in C*". Cambridge Univ. press, New Delhi, 1993.

[27] G. W. Wyzecki and W. S. Stiles. *Color Science*. John Wiley, New York, 1967.

Chapter 3

GENERALIZED INTERPOLATION FOR SUPER-RESOLUTION

Deepu Rajan*
School of Biomedial Engneerng
Indian Institute of Technology-Bombay
Powai, Mumbai-400 076. India.
dr@doe.cusat.edu

Subhasis Chaudhuri
Department of Electrical Engineering
Indian Institute of Technology-Bombay
Powai, Mumbai-400 076. India.
sc@ee.iitb.ernet.in

Abstract In this chapter, we present a generalized interpolation scheme for image expansion and generation of super-resolution images. The underlying idea is to decompose the image into appropriate subspaces, to interpolate each of the subspacs individually and finally, to transform the interpolated values back to the image domain. This method is shown to presere various optical and structural properties of the image, such as 3-D shape of an object, regional homogeneity, local variations in scene reflectivity, etc. The motivation for doing so has also been explained theoretically. The generalized interpolation scheme is also shown to be useful in perceptually based high resolution representation of images where interpolation is done on individual groups as per the perceptual necessity. Further, this scheme is also applied to generation of high-resolution transparencies from low resolution transparencies.

Keywords: Interpolation, Super-resolution, Perceptual grouping

*On leave from Department of Electronics, Cochin University of Science and Technology, Cochin - 682 022. India.

1. Introduction

In several applications like medical imaging, consumer electronics and aerial surveillance, interpolation of images is done to zoom into a particular region of the image or to increase the resolution of the image as a whole. The standard procedure for interpolation is to fit the data with a continuous function and resample the function at finer intervals as desired. The difference among various approaches lies in the interpolation model that is chosen [1, 2]. The simplest interpolating function is the nearest neighbor function where the value at the interpolating point is the same as the value at the grid point closest to it. In linear interpolation, the unknown point is interpolated linearly between the grid points. Bilinear and cubic spline interpolators are the standard techniques preferred in commercial image processing tools due to their improved performance and moderate complexity. Enlargement and reduction of digital images through interpolation in the intensity domain have been extensively studied; see e.g. [1, 2, 3, 4]. It is well known that lower order interpolation methods are the simplest and the fastest, but they produce artifacts. Higher order interpolation methods, particularly smoothing splines perform better though they tend to blur the image. Recently, wavelets have emerged as a promising tool for interpolation, as discussed in the previous chapter.

In many engineering applications, we come across situations where a problem is transformed from its original domain to another through an operator in order to develop elegant solutions. A classic instance is the pattern classification problem where separability, linear or polynomial, of random patterns is crucial. According to Cover's theorem, a complex pattern classification problem cast in high-dimensional space nonlinearly is more likely to be linearly separable than in a low-dimensional space [5]. In some other cases, the problem space is decomposed into appropriate subspaces which possess certain properties such that the functions/elements contained in them can be represented in a mathematically tractable form as well as processed effectively. Consider the case of sampling and reconstruction of a function that is not bandlimited; it is customary to use an ideal low-pass filter before the sampler, to suppress aliasing [6, 7]. Based on this observation, the sampling process is viewed as an approximation procedure leading to the formulation of least square sampling theories for a class of functions generated from a generating kernel $\phi(x)$ and its integer translates [8]. In this chapter, we describe a method of generalized interpolation [9] where the space containing the original function values is decomposed into appropriate subspaces. These subspaces are chosen so that the rescaling operation

preserves those properties of the original function, some of which might not even be evident in the original space. On combining the rescaled "sub-functions", we get back to the original space containing the enlarged function, possibly with less information loss compared to direct interpolation in the original space. Those properties that were concealed earlier could also be revealed in this fashion. Such an interpolation scheme is called the *generalized interpolation* in this chapter.

The primary focus of researchers trying to solve the image (or scattered data) interpolation problem is in treating it as an approximation problem subject to some standard continuity conditions. Even though there has been a spate of developments in this area, what is often overlooked is that the quality of the interpolated image is judged by the way one perceives it. A particular interpolation scheme may perform quite well for an image if it contains objects of nearly similar textures. In presence of dissimilarly textured regions, it affects the intervening region. The motivation here is to preserve the output from introducing any perceptual artifacts, mostly at places of object boundaries. In the literature there has been very little effort in developing a perceptually motivated interpolation method. Ramponi [10] has proposed a space variant linear image interpolation scheme wherein the definition of distance (as in the bilinear interpolation) is locally changed to take care of the neighborhood structure at a grid point. Although it is difficult to analyze the result mathematically, it may have a weak relevance to the nature of visual perception, namely the locality of the dynamic range. Thurnhoffer and Mitra [11] have proposed a non-linear interpolation scheme based on a polynomial operator wherein perceptually relevant features (say, edges) are extracted and zoomed separately. Usually the edges play an important role in perception and the authors put due emphasis on this fact while interpolating the image. It should be possible to extend the idea further by perceptually grouping these features and treating them individually as per their perceptual relevance.

In the next section, we develop the mathematical theory of the generalized interpolation. Section 3 illustrates the proposed technique through several practical applications. We pick a few applications like generation of 3-D structure and reflectance preserving method of image interpolation using photometric stereo, super-resolution rendering and perceptual-grouping based interpolation and show results of the proposed scheme in Section 4. Conclusions are presented in Section 5.

2. Theory of Generalized Interpolation

Let us first review the existing method of image (or scattered data) interpolation. The real valued multivariate interpolation problem can be stated as follows : Given N different observation points $\mathbf{x_i} \in \Re^n$, $i = 1,\ldots, N$ and N real observations $\{y_i \in \Re, \quad i = 1,\ldots,N\}$, find a mapping $f : \Re^n \rightarrow \Re$ satisfying the interpolation conditions $f(\mathbf{x_i}) = y_i$, $i = 1,\ldots, N$. Similarly, if we want to approximate a set of data $S = \{(\mathbf{x_i}, y_i) \in \Re^n \times \Re | i = 1,\ldots,N\}$ with a function f, one needs to minimize the following cost function,

$$H = \sum_{i=1}^{N} (y_i - f(\mathbf{x_i}))^2. \tag{3.1}$$

Obviously, in either case, the solution is ill-posed as the interpolation function can take arbitrary values $f(\mathbf{x})$ where the function is not defined. This calls for the use of a regularizing term, and one solves the following cost function :

$$H = \sum_{i=1}^{N} (y_i - f(\mathbf{x_i}))^2 + \lambda \|Pf\|^2 \tag{3.2}$$

where P is the constraint operator, also called a stabilizer, and λ is a positive real number called the regularization parameter. Duchon [12] and Meinguet [13] consider stabilizers of the form

$$\|Pf\|^2 = \sum_{i_1 \ldots i_m = 1}^{\infty} \int_{\Re^n} d\mathbf{x} (\partial_{i_1 \ldots i_m} f(\mathbf{x}))^2 \tag{3.3}$$

where $\partial_{i_1 \ldots i_m} = \frac{\partial^m}{\partial x_1 \ldots \partial x_m}$ and $m \geq 1$. For a bivariate interpolation scheme, a common choice of the regularization term is [14]

$$\|Pf\|^2 = \int_{\Re^2} [(\frac{\partial^2 f}{\partial x^2})^2 + 2(\frac{\partial^2 f}{\partial x \partial y})^2 + (\frac{\partial^2 f}{\partial y^2})^2] dx dy. \tag{3.4}$$

coplanar points. The surface that minimizes this expression is referred to as *thin plate spline* since it relates to the energy in a thin plate forced to interpolate the data [14].

According to Lagrange's Theorem, if a function $F(x)$ possesses the $(n+1)^{th}$ derivative $F^{(n+1)}(x)$ at all points of an interval containing the point x_0, the remainder $R_n(x)$ is representable in the form

$$R_n(x) = F^{(n+1)}(\xi) \frac{(x - x_0)^{(n+1)}}{(n+1)!}$$

for every point x of this interval where ξ is a number lying between x_0 and x. Using this, it can be easily shown that for a k^{th} order polynomial approximation of the original (unknown) function $\tilde{f}$ at a point $\delta\mathbf{x}$ away from the nearest grid point, the approximation error is bounded by [15]

$$|f - \tilde{f}| < \frac{|\delta\mathbf{x}|^{k+1}}{(k+1)!} \max_{\mathbf{x}} |(\frac{\partial}{\partial x} + \frac{\partial}{\partial y})^{k+1} f(\mathbf{x})||. \tag{3.5}$$

For a thin plate fitting spline over a square grid of size h, the maximum error is bounded by [16]

$$|f - \tilde{f}| \leq ch\sqrt{|\log h|}\ \|Pf\|, \tag{3.6}$$

where c is a positive number given by $[(32\pi)^{-1}(3\log 2)]^{\frac{1}{2}}$.

Let us now consider the following abstract parametric decomposition of the function $\tilde{f}(\mathbf{x})$.

$$\tilde{f}(\mathbf{x}) = g(a_1(\mathbf{x}), a_2(\mathbf{x}), \ldots, a_m(\mathbf{x})), \tag{3.7}$$

where $a_i(\mathbf{x})$, $i = 1, 2, \ldots, m$ are different functions of the interpolating variable $\mathbf{x}$ and when they are combined by an appropriate m-variate function g, one recovers the original function f. We simply assume these functions $a_i(\mathbf{x})$ and g to be arbitrary, but continuous (i.e., in C^k). The rationale for such a parametric decomposition will be explained in the next section. We can now interpolate the individual functions $a_i(\mathbf{x})$ and combine them using Eq. (3.7) to obtain a rescaled $\tilde{f}(\mathbf{x})$. We call this operation as *generalized interpolation.*

The interpolation error at a point $\mathbf{x}$ can be written as

$$\begin{aligned} |f_g - \tilde{f}| &= |g(a_1(\mathbf{x}) + \epsilon_1, \ldots, a_m(\mathbf{x}) + \epsilon_m) \\ &\quad - g(a_1(\mathbf{x}), a_2(\mathbf{x}), \ldots, a_m(\mathbf{x}))| \\ &\approx |\epsilon_1 \tfrac{\partial g}{\partial a_1} + \cdots + \epsilon_m \tfrac{\partial g}{\partial a_m}| \end{aligned} \tag{3.8}$$

where f_g represents the result of generalized interpolation. Here ϵ_i, $i = 1, 2, \ldots, m$ are the interpolation errors at the same point $\mathbf{x}$ for the associated interpolant $a_i(\mathbf{x})$. For a mathematically tractable evaluation of equation (3.8), it is necessary to have some knowledge of the functions g and $a_i(\mathbf{x})$. In order to get a feel for the behavior of the error function, we consider g to be a linear function, i.e.,

$$g(a_1(\mathbf{x}), a_2(\mathbf{x}), \ldots, a_m(\mathbf{x})) = \sum_{i=1}^{m} \alpha_i a_i(\mathbf{x}),\ \alpha_i \geq 0\ \forall i. \tag{3.9}$$

From Eq. (3.9), the interpolation error using a *kth* order polynomial at a point δ away from a grid point $\mathbf{x}$ is given by

$$f_g(\mathbf{x}) - \tilde{f}(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \epsilon_i,$$

$$\text{or} \quad |f_g(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq \sum_{i=1}^{m} \alpha_i |\epsilon_i|,$$

i.e.,

$$|f_g(\mathbf{x}) - \tilde{f}(\mathbf{x})| < \frac{|\delta|^{(k+1)}}{(k+1)!} \sum_{ii=1}^{m} \alpha_i \max_{\mathbf{x}} |(\frac{\partial}{\partial x} + \frac{\partial}{\partial y})^{k+1} a_i(\mathbf{x})|. \tag{3.10}$$

On the other hand, if one performs a *kth* order polynomial interpolation at the same location on the scattered data $f(\mathbf{x}_i)$ itself, the corresponding error bound is

$$|f(\mathbf{x}) - \tilde{f}(\mathbf{x})| < \frac{|\delta|^{(k+1)}}{(k+1)!} \max_{\mathbf{x}} |(\frac{\partial}{\partial x} + \frac{\partial}{\partial y})^{k+1} f(\mathbf{x})|. \tag{3.11}$$

We need to determine whether we gain anything by individually interpolating the constituent functions of g instead of interpolating the function $f(\mathbf{x})$ directly? In order to prove that there is, indeed, some gain, one should compare Eq. (3.10) and (3.11) and must prove that

$$\sum_i \alpha_i \max_{\mathbf{x}} |\frac{\partial^{k+1} a_i(\mathbf{x})}{\partial \mathbf{x}^{k+1}}| \leq \max_{\mathbf{x}} |\frac{\partial^{k+1} f(\mathbf{x})}{\partial \mathbf{x}^{k+1}}| \tag{3.12}$$

Similarly, for a thin plate spline interpolation, it can be shown that if one were to achieve a lower approximation error using the parametrically decomposed generalized method, we must have

$$\sum_{i=1}^{m} \alpha_i \|P a_i(\mathbf{x})\| \leq \|P f(\mathbf{x})\|. \tag{3.13}$$

Similarly, if the function g has the product form

$$g = \alpha \prod_{i=1}^{m} a_i(\mathbf{x}),$$

the corresponding relationship to be satisfied can be derived to be

$$\sum_{i=1}^{m} \frac{\|P a_i(\mathbf{x})\|}{|a_i(\mathbf{x})|} \leq \frac{\|P f(\mathbf{x})\|}{|f(\mathbf{x})|}. \tag{3.14}$$

Unfortunately, all these above relationships are not valid when g is a rational function of polynomials. Thus, a direct interpolation of the function $f(\mathbf{x})$ seems to be a better option instead of the indirect one. However, although the authors are not aware of any proof, except for what has been proved in the theorem given below, there may exist some arbitrary functions g for which the inequality $|f_g(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq |f(\mathbf{x}) - \tilde{f}(\mathbf{x})|$ may hold (here f_g represents the result of generalized interpolation), justifying the interpolation of the constituent functions. Our experimental results in Section 4 support this. Furthermore, we know that the solution for a thin plate spline interpolation is of the form [12]

$$s(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i \phi(\|\mathbf{x} - \mathbf{x}_i\|_2) + \lambda_{n+1} + \lambda_{n+2}x + \lambda_{n+3}y, \tag{3.15}$$

where the function ϕ is defined as

$$\phi(r) = r^2 \log r, \quad 0 \leq r < \infty,$$

and the parameters $\{\lambda_i : i = 1, 2, \ldots, n+3\}$ satisfy the equations

$$\sum_{i=1}^{n} \lambda_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} \lambda_i \mathbf{x}_i = 0.$$

Hence, if the functions $a_i(\mathbf{x})$ actually have a form similar to that of $s(\mathbf{x})$ given in Eq. (3.15), the product form decomposition of the image $f(\mathbf{x})$ definitely reduces the approximation error.

We refrain from attempting, in this study, to find out the exact mathematical conditions under which the generalized interpolation scheme would outperform the image based interpolation method as this is an unsolved problem. But we do prove below the superiority of the proposed scheme for a particular case.

Theorem. If all functions $\{a_i(\mathbf{x})\}$ given in Eq. (3.7) are bandlimited to a maximum frequency of W, and the functional $f(\mathbf{x})$ is sampled uniformly with a sampling frequency $f_s \geq 2W$, then

$$0 = |f_g(\mathbf{x} \uparrow n) - \tilde{f}(\mathbf{x} \uparrow n)| \leq |f(\mathbf{x} \uparrow n) - \tilde{f}(\mathbf{x} \uparrow n)| \tag{3.16}$$

where the symbol $(\uparrow n)$ denotes upsampling by a factor of n, and when the sinc interpolation is used for upsampling. The equality holds only when the function $g(\cdot)$ is such that the bandwidth of the resulting function $\tilde{f}(\mathbf{x}) \leq W$.

Proof: The proof is quite simple. Since all $a_i(\mathbf{x})$'s are band-limited and since a sinc interpolation is used, all interpolations are exact, i.e.,

$$\tilde{a}_i(\mathbf{x} \uparrow n) - a_i(\mathbf{x} \uparrow n) = 0, \forall i.$$

Hence the generalized interpolation $f_g(\mathbf{x} \uparrow n)$ yields zero error as given in Eq. (3.16).

Now the function $\tilde{f}(\mathbf{x})$ given in Eq. (3.7) need not be band-limited to the frequency W. This is true when the function $g(\cdot)$ is non-linear. Under such circumstances, the function $f(\mathbf{x})$ suffers from aliasing, when upsampling does not help, as aliased components cannot be recovered. However, the generalized interpolation scheme can recover the original, unaliased signal exactly. It may be noted that one can construct some pathological examples when the bandwidth of $f(\mathbf{x})$ is less than that of $a_i(\mathbf{x})$. For an example, consider $f(\mathbf{x}) = a_1(\mathbf{x})a_2(\mathbf{x})$ where $a_1(\mathbf{x}) = a_{10}(\mathbf{x})/a_0(\mathbf{x})$ and $a_2(\mathbf{x}) = a_{20}(\mathbf{x})a_0(\mathbf{x})$. Because of cancellation of the common factor, the bandwidth of $f(\mathbf{x})$ gets reduced. It should be noted here that any bandlimited function $a_i(\mathbf{x})$ can be written in terms of Nyquist pulses and the samples values

$$a_i(\mathbf{x}) = \sum_{-\infty}^{+\infty} a_i(\frac{k}{2W}) \frac{\sin[\frac{\pi}{2W}(x - \frac{k}{2W})]}{\frac{\pi}{2W}(x - \frac{k}{2W})}. \tag{3.17}$$

The same argument is valid for $a_i(\mathbf{x} \uparrow n)$. However, $\tilde{f}(\mathbf{x})$ not being bandlimited, the above representation (equation (3.17)) is neither valid for $\tilde{f}(\mathbf{x})$ nor $\tilde{f}(\mathbf{x} \uparrow n)$. The function $\tilde{f}(\mathbf{x} \uparrow n)$ will have aliasing above the frequency nW. By selecting the upsampling factor n, it is possible to recostruct $\tilde{f}(\mathbf{x} \uparrow n)$ perfectly.

Illustration. Consider a simple function

$$f(n) = a_1(n)a_2(n) = \cos(2\pi f_1 n + \phi_1)\cos(2\pi f_2 n + \phi_2).$$

We take $f_1 = 0.3750$ and $f_2 = 0.2500$. The function $f(n)$ as shown in Fig. 3.1(a) suffers from aliasing. When $f(n)$ is upsampled using a sine interpolator, $f(n \uparrow 2)$ as shown in Fig. 3.1(b) suffers from the same kind of aliasing. In Fig. 3.1(c), we plot $a_1(n \uparrow 2) \cdot a_2(n \uparrow 2)$, the result of generalized interpolation, and this is free from any aliasing.

The theory of generalized interpolation scheme is also shown to be suitable for interpolation based on perceptual grouping. According to Lowe [17] : "Perceptual organization refers to a basic capability of the human visual system to derive relevant groupings and structures from an image without prior knowledge of its contents." Several researchers have addressed the problem of perceptual grouping in computer vision. Sarkar and Boyer [18] modify the formalism of Bayesian networks to aid in an integration of the top-down and bottom-up approaches for visual processing. Clemens and Jacobs [19] illustrate the importance of grouping to indexing based recognition strategies. Raman *et al.* [20]

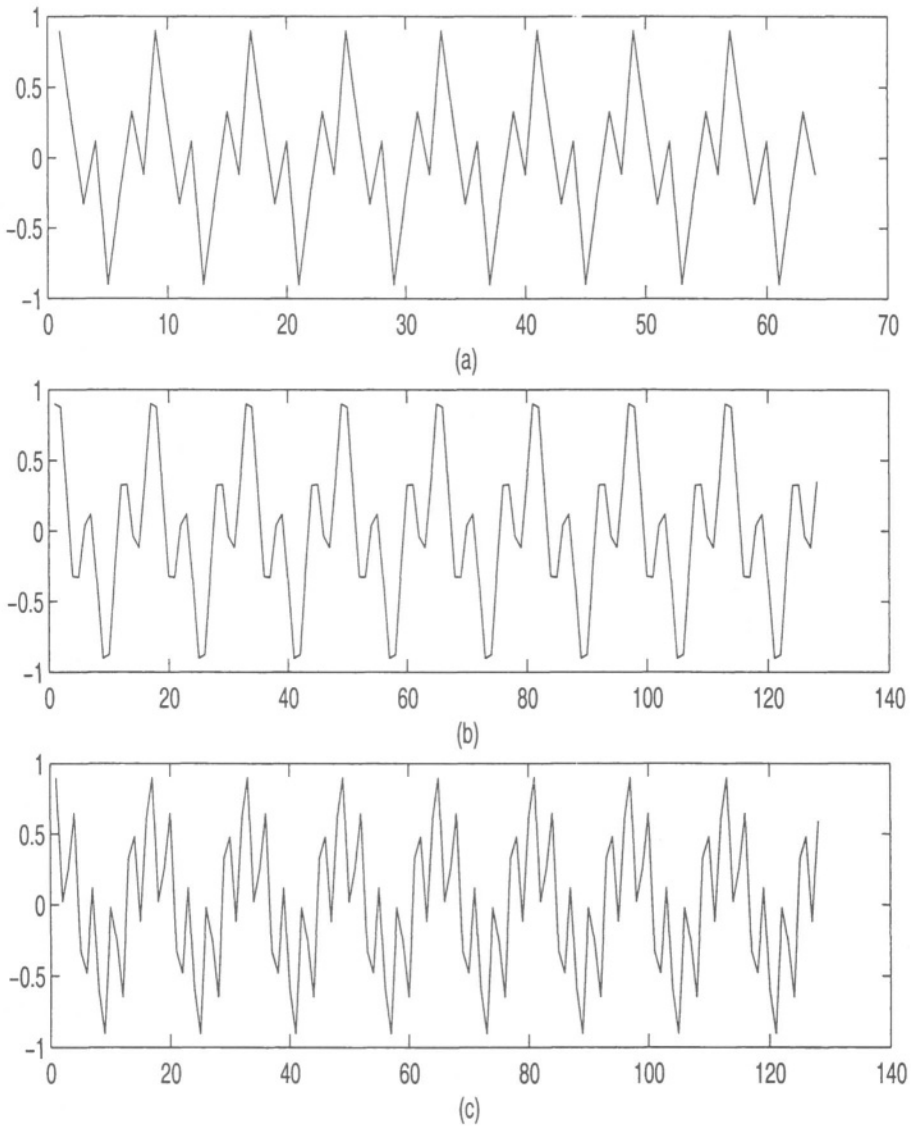


Figure 3.1. (a) The original signal, (b) Sinc interpolation of (a) wherein the aliased component remains, (c) Generalized interpolation of (a) which is free from aliasing.

describe a method for automatic generation of educated guesses regarding structures present in medical images. Liou *et al.* [21] develop a signal level perceptual organization(SLPO) to generate abstract representations for partitioning and identification of regions. The fact that low level features like edges or regions group together to form higher

level inxgeometric features suggest that such geometric arrangements are mutually dependent. Thus, perceptual grouping can be viewed as a inxbottom-up process in which image features from a single object tend to cluster together.

Our contention is that the interpolation procedure should not be independent of the perceptual grouping present in an image. Rather, it should be such that the properties of the respective perceptual groups are retained. To this end, we assume that the perceptual grouping has already been done on an image. Cues that could be used to achieve this may include texture, directionality, edges, contours, depth related defocus, symmetry, etc. For the proposed technique the functions $a_i(\mathbf{x})$ can be defined based on grouping as is explained in the next section. For example, Syeda-Mahmood [22] does grouping based on salient textural features. These are examples of tangible cues based on some physical properties of the image. In this paper we allow the grouping to be based on abstract cues also. In general, the grouping results in disjoint partitions of the image, where each region can be interpolated separately, leading to a fairly trivial solution. Interestingly enough, the grouping does not have to be disjoint in the image space. Take the case of image transparency when objects at different depths are superposed together on the same image plane. One can use a method proposed in [23] to separate the corresponding transparency layers. Instead of doing interpolation on the combined image (when edges at different layers appear at different locations and they tend to affect the gray level in other layers), one must perform the interpolation on the perceptually grouped data (i.e., different layers) and combine them meaningfully. In yet another example, we may want to group the objects in an image in terms of its 3-D shape and/or variations in albedo. These are all intrinsic properties of a scene and they get very easily changed whenever any processing is done on the image, thus losing their significance. By employing the generalized interpolation scheme described above, we can handle all these above categories of perceptual grouping problems in a unified framework. Indeed, all kinds of groupings such as disjoint, overlapped and abstract, are allowed under the proposed method.

3. Some applications of Generalized Interpolation

In this section, we illustrate some applications of the proposed generalized interpolation scheme.

3.1. Structure preserving super-resolution

The proposed generalized interpolation scheme can be applied to generation of super-resolved images. The question we pose is - how can the 3D structural information of an object present in the scene be exploited to generate super-resolved images? A super-resolution technique based on the surface reflectance properties of objects in an ensemble of images captured under different distributions of light source positions is presented in [24]. Given such an ensemble the task is to obtain a super-resolved image not only with a source position already present in one of the images in the ensemble, but also with an arbitrary source position. This is achieved through interpolation in the structural domain (i.e., surface normals at different points on the object) and on the albedo of the surface. The key idea here is to preserve the 3D surface properties of the object rather than the observed image property during the up-sampling process. The importance of this has already been suggested by Peleg and Ron in [25] during the downsampling process. Assuming all sources to be point sources and surfaces to be Lambertian (note that the method is valid for any other reflectance model) and that the images are formed by orthographic projection, we use photometric stereo to recover local surface orientation (p,q) and albedo $\rho(x, y)$ [26]. Two images taken with different lighting yield two irradiance equations:

$$R_1(p, q) = E_1 \quad \text{and} \quad R_2(p, q) = E_2 \tag{3.18}$$

where p and q are the surface normals and E_i is the intensity at (x,y).. where p and q are the surface normals and E_i is the intensity at (x,y). However, in order to recover albedo simultaneously, we need another equation and hence a third image with a different source position. Suppose the unit vectors in the directions of three source positions are, respectively, given by

$$\hat{s_i} = (-p_i, -q_i, 1)^T / \sqrt{1 + {p_i}^2 + {q_i}^2}, \quad i = 1, 2, 3$$

and the unit surface normal at a location (x,y) is given by

$$,, \hat{n} = (-p, -q, 1)^T / \sqrt{1 + p^2 + q^2}, \tag{3.19}$$

then $E_i = \rho(\hat{s_i}.\hat{n})$ where ρ is the albedo $(0< \rho <1)$. Writing this expression in matrix form, we get

$$\mathbf{E} = \rho S \hat{n} \tag{3.20}$$

where the rows of S are the source directions and the components of the vector **E** are the three brightness measurements. Assuming non-

singularity of S,

$$\rho\hat{n} = S^{-1}\mathbf{E}. \tag{3.21}$$

The length of the vector $\rho\hat{n}$ in Eq. (3.21) gives the albedo. For the case of m brightness measurements, a least squared solution for $\hat{n}$ is found from Eq. (3.21) by taking the pseudo-inverse of S.

Having obtained the surface normals for, say, an $M \times N$ image, bicubic spline (or any other suitable) interpolation is carried out individually in the p, q and albedo (ρ) spaces to yield magnified normal and albedo spaces of dimension $Mb \times Nb$, where b is the magnification factor. Such a spline interpolation technique is commonly known as *vector spline interpolation* [27, 28]. The interpolated normals are now used to reconstruct the image according to the image irradiance equation. Note that this reconstruction is general in the sense that any source position can be used to generate the corresponding shaded image. No matter which source directions were used in the estimation of normals and albedo, a new view can always be generated. In comparison to interpolation in the image domain $f(x, y)$, we note that here the interpolants $a_1(x, y) = p(x, y)$, $a_2(x, y) = q(x, y)$ and $a_3(x, y) = \rho(x, y)$ are non-parametric functions of (x, y) and the equivalent function g in Eq. (3.7) is an irrational function of the interpolants a_1, a_2 and a_3. We further note that in order to obtain the inverse g^{-1}, one requires several photometric measurements. However, the advantage is that one does not require to establish the sub-pixel registration of different observations, as most of the existing methods in the literature do, (e.g., [29] and [1]), while generating the super-resolution image.

3.2. Object-based grouping

This is a very simple example to illustrate the concept of generalized interpolation. Consider an image f(**x**) consisting of an object and a background. Let χ_A denote the characteristic function of set A. Then

$$f(\mathbf{x}) = f_{bac}(\mathbf{x})(1 - \chi_{obj}(\mathbf{x})) + f_{obj}(\mathbf{x})\chi_{obj}(\mathbf{x})$$

where subscripts *bac* and *obj* stand for background and object areas, respectively. Often the significant part of the information in an image is contained in the foreground while little or irrelevant information is present in the background. A typical instance is that of medical images like CT and MRI scans. Under such conditions, the computational cost of interpolation can be minimized by interpolating the object and the foreground separately and differently, as per the demand of visual perception, e.g., the foreground could be interpolated with B-splines for higher accuracy while a simple bilinear interpolation will suffice for the

background. One, however, must know $\chi_A(\mathbf{x})$ in order to proceed with the interpolation. The characteristic function χ can be obtained through a proper scene segmentation.

3.3. Super-resolution Rendering

At the core of any 3D graphics system is a sub-system that renders objects represented by a set of polygons. The usual approach to rendering of 3D objects is to build a basic renderer that incorporates a local reflection model like the Phong model [31] into an incremental shader and then add on various enhancements. Gouraud shading and Phong shading are two algorithms that have contributed to the growth of polygon based renderers. While Gouraud shading calculates intensities at polygon vertices only, using a local reflection model, and interpolates these for pixels within the polygon, Phong shading interpolates vertex normals and applies a local reflection model at each point.

We consider the case of image rendering at a higher resolution, which incorporates the shading effects of multiple illumination sources. Hence, the problem is to generate a high-resolution image wherein the shading effects due to multiple images are integrated. In comparison to ordinary image rendering techniques, we do not have the shape of the object *apriori*; we find the shape at a low resolution using photometric stereo. Two approaches to the solution are suggested. The first one involves interpolation in the surface normals and albedo space and the other involves interpolation in the spatial domain itself.

In the first method, i.e., Phong shading which is an existing usage of generalized interpolation, the surface normals (p,q) and albedo(ρ) are determined from photometric stereo, followed by cubic spline interpolation in the p and q spaces, as was done in the section 3.1. The albedo space is also separately interpolated. If there are m sources, each having a corresponding weight β_i, then

$$E(x,y) = \sum_{i=1}^{m} \beta_i \, \rho \, \hat{n} \, . \, \hat{s}_i = \rho \, \hat{n} \, . \, (\sum_{i=1}^{m} \beta_i \, \hat{s}_i). \tag{3.22}$$

The weights are normalized so that $\sum_{i=1}^{m} \beta_i = 1$. The interpolated values of normals and albedo are used to reconstruct the super-resolved image using Eq. (3.22). Further modifications to the rendered image can be done by changing the high resolution albedo, causing new textures, soft shadows, etc. to be introduced.

The second approach (Gouraud shading combined with generalized interpolation) is to carry out interpolation in the spatial domain itself, but taking into account the differences in illuminated (*il*) and dark (*da*)

regions in the image. This is similar to the example in Section 3.2, where the dichotomy was between object and background regions. Let $f_i(\mathbf{x})$ be the shaded image obtained from the ith source S_i. Following the notation used earlier

$$f_i(\mathbf{x}) = f_{i,il}(\mathbf{x})(1 - \chi_{i,da}(\mathbf{x})) + f_{i,da}(\mathbf{x}))\chi_{i,da}(\mathbf{x}). \tag{3.23}$$

For each source and for each region of the image, one can perform the corresponding Gouraud shading differently using appropriate interpolation schemes (for example, some of these schemes could include wavelet based interpolation, Markov random field-based interpolation, traditional spline fitting or bilinear interpolation). The corresponding rendered image would be of the form

$$f(\mathbf{x} \uparrow 2) = \sum_i \beta_i f_{i,il}(\mathbf{x} \uparrow 2)(1 - \chi_{i,da}(\mathbf{x} \uparrow 2)) + f_{i,da}(\mathbf{x} \uparrow 2)\chi_{i,da}(\mathbf{x} \uparrow 2).$$

3.4. Grouping based on Transparency

In a typical outdoor or indoor scene, there exist many transparent objects with glossy surfaces such as glass and water. We see the reflected image of an object on such transparent surfaces together with other objects behind them. Thus, there is an overlapping of objects behind and in front of the transparent layer. A typical instance is the case of a person in a car who is not clearly visible from outside due to reflection of the surrounding scene on the window glass. Separation of real components from reflected components in an overlapping image is useful for high quality TV camera images, transparent object avoidance for mobile robots and detection of objects outside the camera's field of view. Although separation of transparent layers seem to be a very difficult task computationally, the human visual system is able to group the layers very effectively. A few techniques have been proposed for separating real and virtual objects based on the polarization of light reflected on a specular surface using a series of images captured by rotating a polarizing filter [32, 33]. Schechner *et al.*, [23] use the depth from focus cue for detection and depth estimation of transparent layers. The overlapping image is represented as

$$I(x, y) = I_A(x, y) + I_B(x, y)$$

where $I_A(x, y)$ is due to object A behind the transparent object and $I_B(x, y)$ is due to object B in front and on the same side as the camera. The images I_A and I_B may have different structural properties and hence, it is not a good idea to interpolate the combined image I as a single entity. Hence these layers must first be grouped separately and then

processed. The constituent images I_A and I_B should be separately interpolated using appropriate interpolation techniques and then combined. The additional information one requires here is the separate availability of individual layers or the cue needed to split the image into different layers.

4. Experimental Results

In this section, we present some experimental results to demonstrate the performance of the proposed generalized interpolation scheme. First, the case of structure preserving super-resolution imaging is illustrated through simulations on the synthetic "sphere" image of size 64 × 64. The albedo varies from 1 at the center of the sphere to 0.75 at the occluding boundary according to $\rho = 1 - \frac{r}{4 \times r_o}$ where r_o is the radius of the sphere. Figure 3.2(a) shows one such low-resolution image with source position at (0,0), out of an ensemble of 9 input images. Various super-resolved images of magnifications 2,3 and 4 were generated using the technique described. Figure 3.2(b) is the super-resolved image of Figure 3.2(a) with a magnification 2. In this case, surface normals and albedo were estimated using 3 sources. The intensity domain interpolated sphere image is shown in Figure 3.2(c). It is quite apparent from the figures that both techniques yield nearly identical results as the image intensity and the (p, q, ρ) functions are all very smooth. Any difference between these two reconstructions can be captured through the computation of root mean square (RMS) errors (with respect to the original image). Figure 3.3 shows the RMS error in the reconstruction of a twice enlarged image, as the number of sources is increased. The source positions (p_s, q_s) used are indicated at the bottom of the figure in serial order. (In all the simulations, sources were used in this order, e.g., 6 sources used to estimate normals/ albedo implies that the first six source positions were used). For the sake of clarity, we have included plots for only 4 out of 9 reconstructions. Each curve in the graph corresponds to reconstruction with a particular source position, e.g., curve labeled "e9" corresponds to reconstruction of the 9th image whose source position is (0.3, -0.4).

We now add zero mean white Gaussian noise to the input images to study how the technique works for noisy images. Reconstruction errors for a two-fold magnified image is shown in Figure 3.4 for the same experimental data and, as expected, the error goes down with more number of observations. Comparisons with cubic spline interpolation in the intensity domain are presented for the noise-free and noisy cases in Figures 3.5 and 3.6, respectively. One can clearly see the benefit of using the generalized interpolation scheme as the RMS error is less for

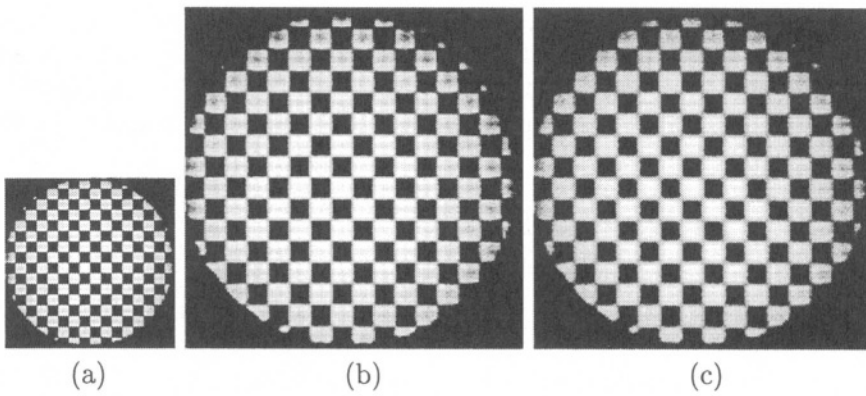


Figure 3.2. (a) Example of a low-resolution image of a sphere, (b) super-resolved image of (a) using the proposed technique, and (c) intensity domain interpolation of (a). (©Elsevier Science)

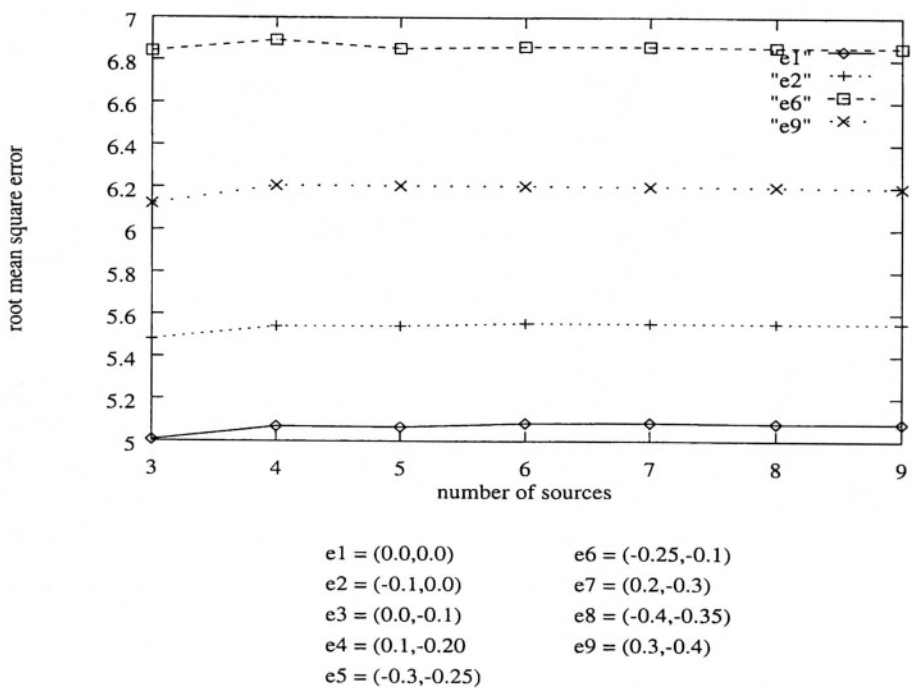


Figure 3.3. RMS errors in reconstruction of the spherical object with magnification 2. The corresponding RMS error for intensity domain interpolation for source position e2 is 7.613, which is more than what is obtained with generalized interpolation.

the proposed method at all magnifications. Figure 3.7 illustrates the effect of increasing noise on the reconstruction error. This figure also

shows how additional information as the number of sources is increased, reduces error in the super-resolved image. In the next example, we take the case of a real image where we obtain perceptually improved interpolation results using the generalized interpolation scheme.

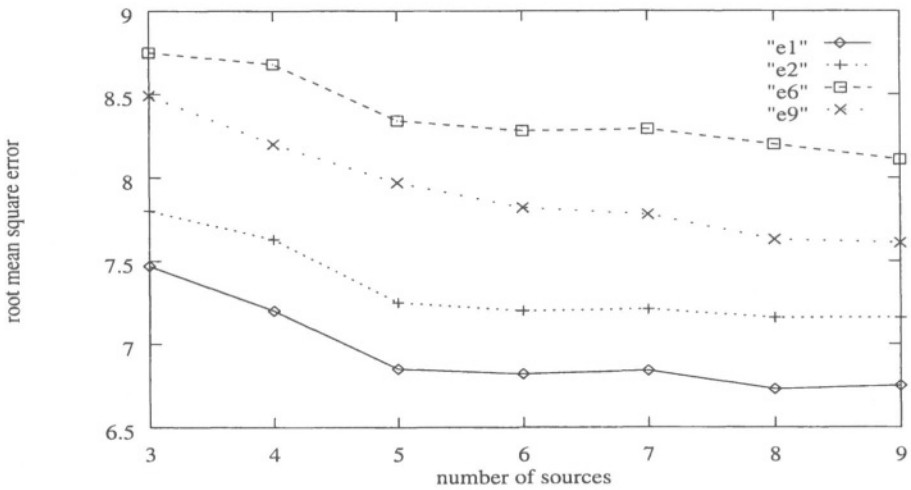


Figure 3.4. RMS errors in reconstruction of noisy image with magnification 2 (Noise variance $\sigma^2 = 6$) for 4 instances of source positions. The corresponding RMS error for intensity domain interpolation for source position e2 is 10.691. (©Elsevier Science)

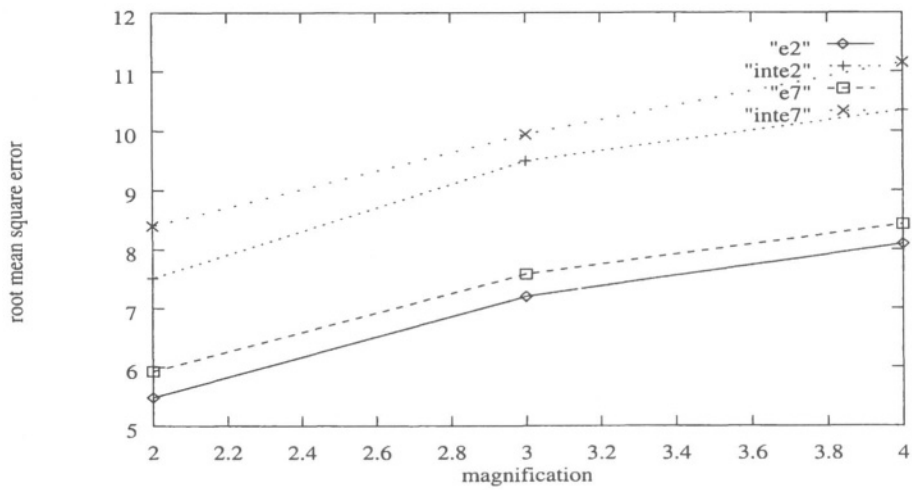


Figure 3.5. Comparison between interpolation in surface normals/albedo domain (ei's) and intensity domain (intei's) for 2 instances of source positions, in absence of observation noise.

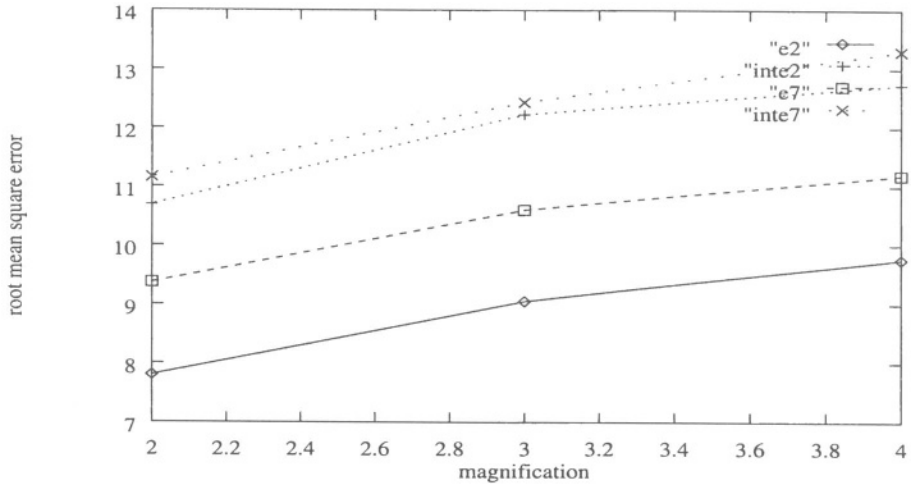


Figure 3.6. Comparison between interpolation results in surface normals/albedo domain (e*i*'s) and intensity domain (Inte*i*'s) for 2 instances of source positions with noise variance $\sigma^2 = 6.0$. (©Elsevier Science)

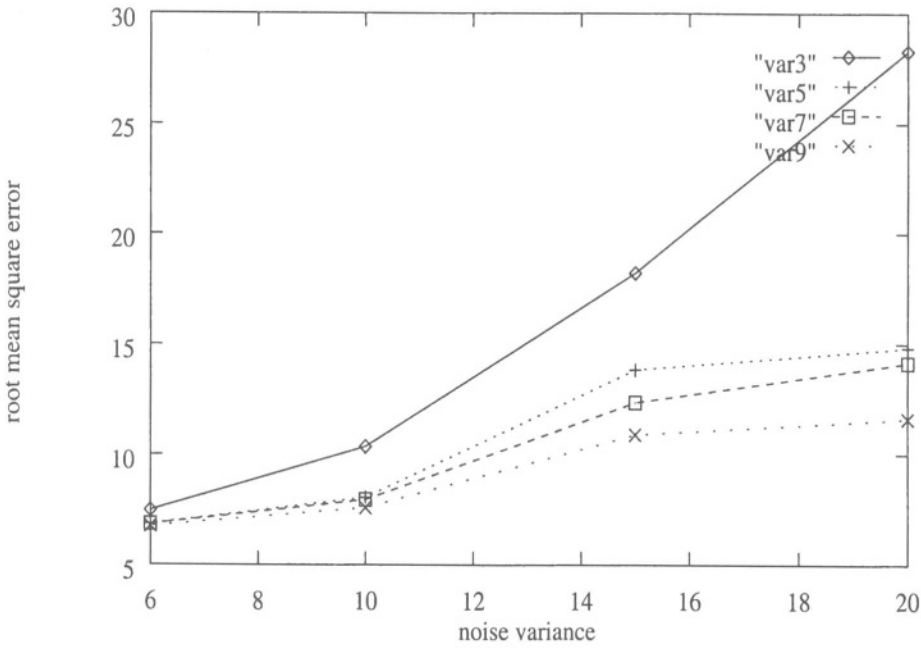


Figure 3.7. Effect of increased noise on reconstruction error (var*i* indicates *i* sources used for reconstruction). (©Elsevier Science)

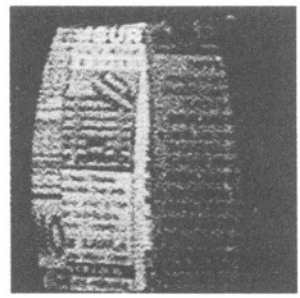


Figure 3.8. One of the low-resolution images of the pen-stand. (©Elsevier Science)

Here, we consider a real example where most of the information in the image is contained in the foreground while there is little or no information in the background. Images of an object with matte surface were captured under controlled illumination. Eight low resolution images of size 235 × 235 with different distributions of source were used to generate a super-resolved image of size 470 × 470. One such low resolution image with source directions (0.2352, 0.4778) is shown in Figure 3.8. The super-resolved image for an arbitray source direction, i.e., one which is not contained in the original ensemble, is shown in Figure 3.9. The clarity of the script printed on the pen-stand after the interpolation is very good compared to its low resolution version given in Figure 3.8 or the intensity domain interpolated image given in Figure 3.10. Furthermore, the right rim of the object which was almost lost in the input image has been recovered quite well in the proposed method. Thus, if the available source positions are insufficient to illuminate a particular part of an object, our method is suitable to reveal details therein. The super-resolved image of Figure 3.8 using the same source position as in that figure is shown in Figure 3.11. Comparison of Figure 3.11 with Figure 3.10 clearly shows the superiority of the proposed method since the letters in the former are much clearer than those in the latter image. While interpolating the (p, q) vectors, each component was interpolated independently as this is much easier to accomplish. Hence the interpolated $(p(\mathbf{x} \uparrow 2), q(\mathbf{x} \uparrow 2))$ function may not satisfy the integrability constraint, i.e., $p_y = q_x$. We also ran experiments where the above constraints were explicitly enforced, but no visible improvement in image quality was observed.

We illustrate the super-resolution rendering concept on the "pen-stand" image. Figures 3.12(a) and 3.12(b) show the result of rendering with three light sources each having different positions and weights. Note that in the absence of the weights for each source, the object would have

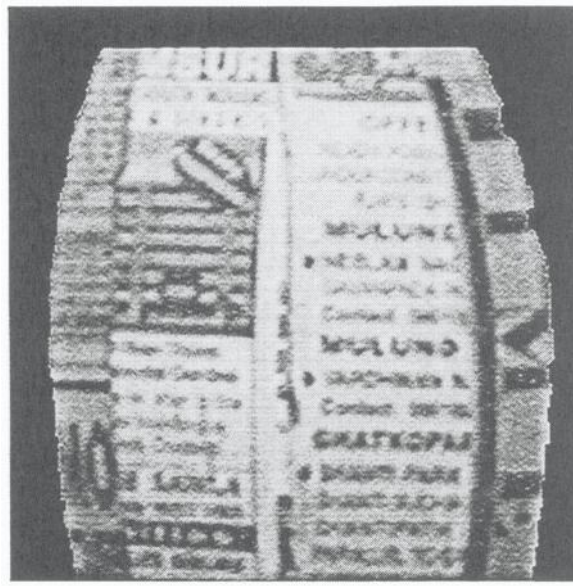


Figure 3.9. The "pen-stand" image with a magnification of two using the proposed technique.

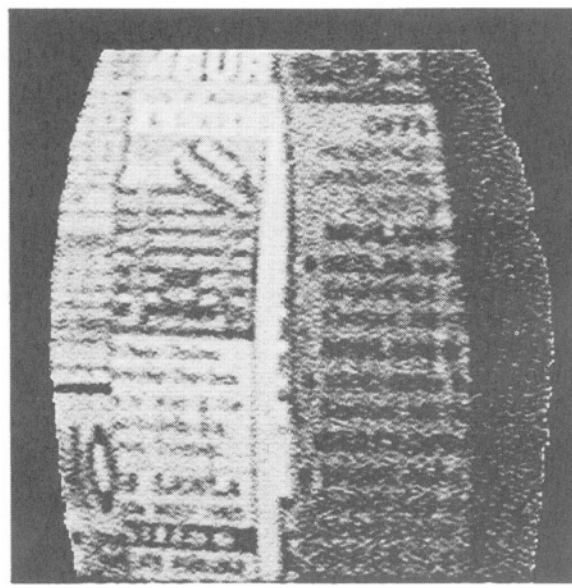


Figure 3.10. The "pen-stand" image interpolated in the intensity domain.

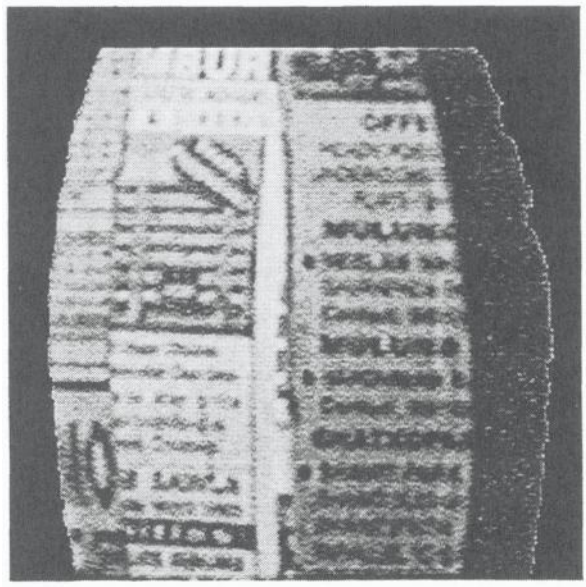


Figure 3.11. The "pen-stand" image super-resolved using the proposed technique with the same source position as in the low resolution observation of Figure 3.8. (©Elsevier Science)

been illuminated uniformly from all the directions and thus no effect of rendering would have been achieved. The utility of this method arises in those cases where the sources cannot be moved around, and yet we would like a rendering of the object that illuminates a particular portion of it which, otherwise, would not be possible with any combination of sources "switched" on.

In order to illustrate the interpolation scheme using the object based grouping, consider the "collage" image given in Figure 3.13 which consists of 5 parts, each having its own perceptual details, e.g. there are more features contained in the "cameraman" portion of the figure compared to the smooth surface of the "vase" (bottom left). Hence, depending on the type of features present in each object, appropriate image expansion techniques can be used to maintain the perceptual quality of the scaled image. In this example, the cameraman and lena portions are interpolated using bicubic splines, the vase and the text image portions are expanded by zero order hold replication of the pixels over a neighborhood while a bilinear interpolation is carried out over the central part of the image. The top half needs a better clarity during the interpolation and hence a spline is used. However, a spline smoothens the textual part. A simple zero-order hold offers a better perceptual

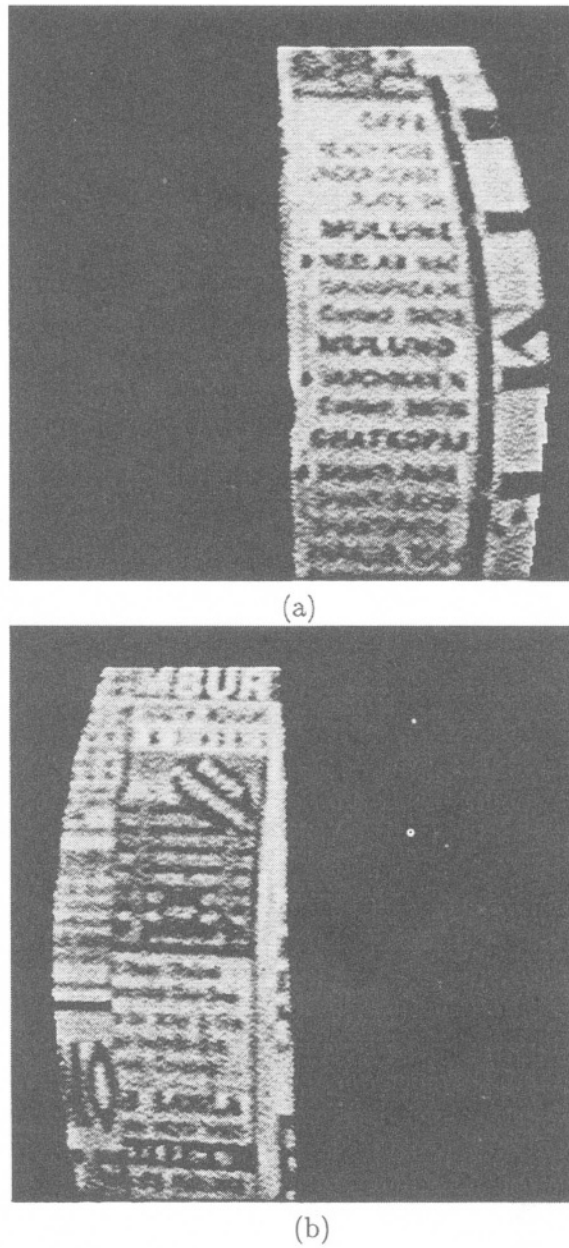


Figure 3.12. Super-resolution rendered "pen-stand" images with multiple weighted sources.

quality, apart from savings in computation. For the vase part of the image, any interpolation scheme does a fairly good job, and the zero-order hold is used from the computational point of view. The results of both the grouping-based interpolation and the existing bi-cubic spline interpolation over the entire image are shown in Figure 3.14(a) and Figure 3.14(b), respectively. We observe that the proposed technique offers better results in terms of (a) preserving the distinctness of the regional boundaries, (b) preventing over smoothening of the textual component, and (c) savings in computation.

Figure 3.13. Original "collage" image. (©Elsevier Science)

(a) (b)

Figure 3.14. Interpolation (a) using object based grouping and (b) without taking grouping into account. (©Elsevier Science)

Next, we illustrate the application of the generalized interpolation scheme to grouping based on transparency. Figure 3.15 shows a transparency image consisting of text pasted on the glass window of a laboratory. Here the "text" region is in the foreground and the "lab" forms the background. The camera is set to focus on the text layer while the

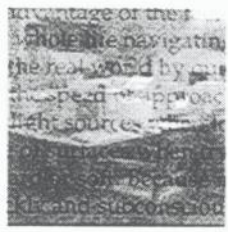


Figure 3.15. Low resolution transparency image containing text pasted on the glass window of a laboratory. (©Elsevier Science)

"lab" layer is defocused. Assuming the layers are separated, using an algorithm as in [23], we have two objectives at hand : the first is to generate a high resolution transparency of the given image and the second is to reconstruct a high resolution transparency where the focus is now set to the background "lab" layer while the "text" layer in the foreground is defocused. Since the background contains plenty of detail, bilinear interpolation is carried out there while zero order hold interpolaton is done over the more sparse foreground. Figure 3.16(a) shows the result of applying the generalized interpolation scheme. For comparison, we have also illustrated in Figure 3.16(b), the case in which the transparency image is enlarged without taking grouping into account; here bilinear interpolation is done over the entire low-resolution transparency. Figure 3.17 shows the reconstructed high resolution transparency image in which the foreground "text" layer is out of focus while the background "lab" layer is focused. This assumes that the blur parameters are known so that the different layers can be manipulated accordingly. We can clearly see the textured pattern of the shirt on the person standing in the lab.

5. Conclusions

We have proposed a generalized interpolation scheme for image resizing and super-resolution applications. There are two facets to our technique. Firstly, the image is decomposed into appropriate subspaces and interpolation is done in each of the subspaces, followed by an inverse transformation of the interpolated space back to the original domain. This allows us to better preserve the structural properties of the object(s) and other reflectance properties in the scene after the interpolation. Secondly, the generalized interpolation scheme is applied to high resolution perceptual grouping in images. The intensity distribution of the image is exploited to determine which areas in an image require a finer and more accurate interpolation technique and in which areas a

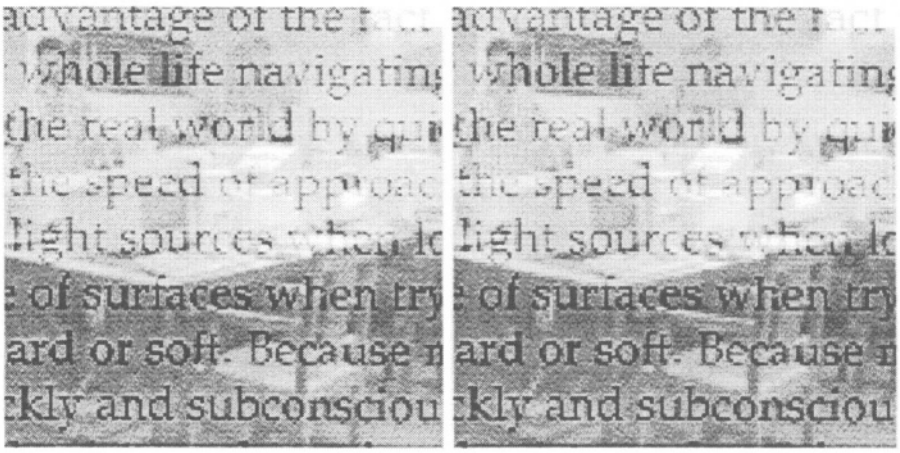


Figure 3.16. Transparency image Figure 3.15 enlarged (a) using the proposed scheme and (b) without taking grouping into account. (©Elsevier Science)

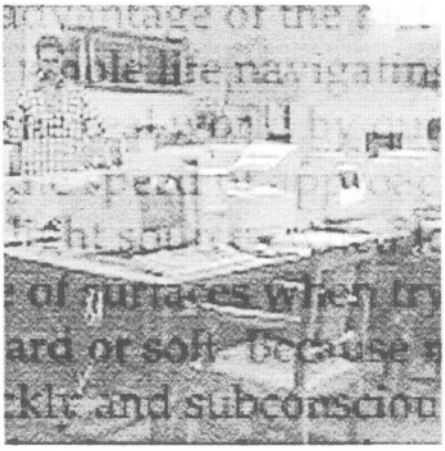


Figure 3.17. Reconstructed transparency image where the "lab" layer is focused but the "text" layer is defocused. (©Elsevier Science)

crude interpolation will suffice, thus saving in computation. A few situations were described where the generalized interpolation scheme is found to be suitable. This is corroborated with experimental results. Future work will involve further consolidating the theory by finding a class of decompositions which guarantees superiority of the proposed scheme.

Acknowledgments

Thanks to Jayashree Karlekar and Narasimha Kaulgud for help in capturing some of the images presented in this chapter.

References

[1] H. S. Hou and H. C. Andrews, "Cubic splines for image interpolation and digital filtering,," *IEEE Trans. on Accoust., Speech and Signal Processing*, vol. 26, no. 6, pp. 508–517, 1978.

[2] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. on Accoust., Speech and Signal Processing*, vol. 29, pp. 1153–1160, 1981.

[3] J. A. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolating methods for image resampling," *IEEE Trans. on Medical Imaging*, vol. 2, pp. 31–39, 1983.

[4] M. Unser, A. Aldroubi, and M. Eden, "Fast B-spline transforms for continuous image representation and interpolation," *IEEE Trans. on Image Processing*, vol. 13, no. 6, pp. 508–517, 1978.

[5] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in Pattern Recognition," *IEEE Trans. on Electronic Computers*, vol. 14, pp. 326–334, 1965.

[6] W. M. Brown, "Optimal prefiltering of sampled data," *IRE Trans. on Information Theory*, vol. IT-17, pp. 269–270, 1961.

[7] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, Jan. 1949.

[8] A. Aldroubi and M. Unser, "Sampling procedures in function spaces and asymptotic equivalence with Shannon's sampling theory," *Numer. Funct. Anal. Optimiz.*, vol. 15, pp. 1–21, Feb. 1994.

[9] Deepu Rajan and Subhasis Chaudhuri, "A generalized interpolation scheme for image scaling and super-resolution," in *Proc. of Erlangen Workshop '99 on Vision, Modelling and Visualization, University of Erlangen-Nuremberg*, Germany, Nov. 1999, pp. 301–308.

[10] G. Ramponi, "Warped distance for space-variant linear image interpolation," *IEEE Trans. on Image Processing*, vol. 8, no. 5, pp. 629–639, May 1999.

[11] S. Thurnhoffer and S. K. Mitra, "Edge-enhanced image zooming," *Optical Engineering*, vol. 35, pp. 1862–1869, July 1996.

[12] J. Duchon, "Spline minimizing rotation-invariant semi-norms in Sobolev space," in *Constructive Theory of Functions of Several Variables*, W. Schempp and K. Zeller, Eds. Springer-Verlag, Berlin, 1977.

[13] J. Meinguet, "Multivariate interpolation at arbitrary points made simple," *Journal of Applied Math. Phys.*, vol. 30, pp. 292–304, 1979.

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

[15] Graham F. Carey, *Computational Grids : Generation, Adaptation and Solution Strategies*, Taylor and Francis, 1997.

[16] M. J. D. Powell, "The uniform convergence of thin plate spline interpolation in two dimensions," *Numerishce Mathematik*, vol. 68, pp. 107–128, 1994.

[17] David G. Lowe, *Perceptual Organisation and Visual Recognition*, Kluwer Academic Press, 1985.

[18] S. Sarkar and K. L. Boyer, "Integration, inference, and management of spatial information using bayesian networks: Perceptual organization,," *IEEE Trans. on Pattern Anal. and Machine Intell.,*, vol. 15, no. 3, pp. 256–274, Mar. 1993.

[19] D. T. Clemens and D. W. Jacobs, "Space and time bounds on indexing 3D models from 2D images," *IEEE Trans. on Pattern Anal. and Machine Intell.,*, vol. 13, no. 10, pp. 1007–1017, Oct. 1991.

[20] S. V. Raman, S. Sarkar, and K. L. Boyer, "Hypothesizing structures in edge focused cerebral magnetic resonance images using graph theoretic cycle enumeration," *Computer Vision, Graphics and Image Understanding*, vol. 57, no. 1, pp. 81–98, Jan. 1993.

[21] Shih-Ping Liou, Arnold H. Chin, and Ramesh C. Jain, "A parallel technique for signal level perceptual organization," *IEEE Trans. on Pattern Anal. and Machine Intell.,*, vol. 13, pp. 317–325, Apr. 1991.

[22] T. F. Syeda-Mahmood, "Detecting perceptually salient texture regions in images," in *Proc. IEEE Workshop on Perceptual Organization and Computer Vision*, Santa Barbara, CA, USA, 1998.

[23] Y. Y. Schechner, N. Kiryati, and R. Basri, "Separating of transparent layers using focus," in *Proc. of Int. Conf. Computer Vision*, Bombay, India, January 1998, pp. 1061–1066.

[24] Deepu Rajan and Subhasis Chaudhuri, "A physics-based approach to generation of super-resolution images," in *Proc. Indian Conf. on Comp. Vis., Graphics and Image Proces.*, New Delhi, India, Dec. 1998, pp. 250–254.

[25] Shmuel Peleg and Gad Ron, "Nonlinear multiresolution: A shape-from-shading example," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1206–1210, Dec. 1990.

[26] B. K. P. Horn, *Robot Vision*, MIT Press, 1986.

[27] J. A. Fessler, "Nonparametric fixed interval smoothing with vector splines," *IEEE Trans. on Signal Processing*, vol. 399, no. 4, pp. 852–859, 1991.

[28] Mohammed Yeasin, *Visual analysis of human motion:Some applications in Biomedicine and Dextrous robot programming*, Ph.D. thesis, Elect. Engg. Department, Indian Institute of Technology, Bombay, 1998.

[29] Hassan Shekarforoush, Marc Berthod, Josiane Zerubia and Michael Werman, "Sub-pixel bayesian estimation of albedo and height," *International Journal of Computer Vision*, vol. 19, no. 3, pp. 289–300, 1996.

[30] H. Ur and D. Gross, "Improved resolution from sub-pixel shifted pictures," *CVGIP:Graphical Models and Image Processing*, vol. 54, pp. 181–186, March 1992.

[31] B.-T. Phong, "Illumination for computer generated pictures," *Communications of ACM*, vol. 18, no. 6, pp. 311–317, June 1975.

[32] N. Ohnishi, K. Kumaki, T. Yamamura, and T. Tanaka, "Separating real and virtual objects from their overlapping images," in *Proc. of ECCV*, Cambridge, UK, 1996, pp. 636–646.

[33] L. B. Wolf, "Using polarization to separate reflection components," in *Proc. of IEEE CVPR*, San Diego, 1989, pp. 363–369.

Chapter 4

RECONSTRUCTION OF A HIGH RESOLUTION IMAGE FROM MULTIPLE LOW RESOLUTION IMAGES

Brian C. Tom
Center for MR Research
Evanston Northwestern Healthcare
Evanston, IL 60201
briant@cmr.nunet.net

Nikolas P. Galatsanos
Illinois Institute of Technology
Dept. of Electrical and Computer Engineering
Chicago, IL 60616
npg@ece.iit.edu

Aggelos K. Katsaggelos
Northwestern University
Dept. of Electrical and Computer Engineering
Evanston, IL 60208
aggk@ece.nwu.edu

Abstract In this chapter the problem of reconstructing a high resolution image from multiple aliased and shifted by sub-pixel shifts low resolution images is considered. The low resolution images are possibly degraded by unknown blurs and their sub-pixel shifts are not known. This problem is described in the frequency and spatial domains. Algorithms for providing solutions to it are reviewed. In addition, two approaches are presented in detail for solving this low-to-high resolution problem. In the first of these two approaches registration and restoration is performed simultaneously using the expectation-maximization (EM) algorithm. The high resolution image is then reconstructed using regularized interpola-

tion which is performed as a separate step. For this reason this approach is abbreviated as **RR-I** which corresponds to registration/restoration-interpolation. In the second of these approaches registration, restoration and interpolation are perfomed simultaneously using the EM algorithm. Therefore this approach is abbreviated as **RRI** which corresponds to registration/restoration/interpolation. Numerical experiments are presented that demonstrate the effectiveness of the two approaches.

Keywords: High resolution images, image registration, multi-channel image restoration, regularized interpolation, expectation maximization algorithm.

1. Introduction

In applications such as remote sensing, military, and medical imaging, images with high-resolution are often required. Such images offer additional detail that may be critical in accurately analyzing the image. Currently, Charge-Couple-Devices (CCDs) are used to capture such high-resolution images digitally. Although this is adequate for most of today's applications, in the near future this will not be acceptable. This is because the technology of CCDs and high precision optics cannot keep up with the demand for images of higher and higher resolution. For example, in order to obtain images approaching (or surpassing) the quality of 35mm film, considered to be the quality criterion for non-electronic visual media, a resolution higher than High Definition Television (HDTV) is needed (greater than 2000×2000 pixels), and current CCD technology cannot achieve this very high resolution [1]. In addition, the presence of shot noise, which is unavoidable in any imaging system, prescribes an upper limit on the resolution of CCDs. This upper limit arises from the fact that while reducing the area of each CCD increases the resolution (more CCDs), the signal strength (number of photons hitting the CCD) is correspondingly decreased, while the noise strength remains roughly the same [40]. This limit on the size of each CCD is roughly 50 μm^2, and current CCD technology has almost reached this limit [1], With a lower CCD area, the Signal-to-Noise (SNR) ratio is too low for images to be useful.

Aside from the approaching limit posed by the presence of shot noise, cost is another concern in using high precision optics and CCDs. Launching a high resolution camera into space on board a satellite can be costly, and even risky. It is more cost-efficient to launch a cheaper camera with a lower resolution into orbit if higher resolution images can be obtained on the ground through image processing techniques.

Finally, another impediment to using high-resolution digital cameras is that often the imaging is done under less than ideal situations. In

military surveillance, for example, taking images of an enemy's troop movement is difficult at best because the enemy is taking steps to move at night, in fog, etc., to hamper surveillance. Weather provides another difficulty for remote sensing, where cloud cover may occlude the area of interest.

For these reasons, an alternative solution is necessary to obtain high-resolution images. Instead of obtaining the high-resolution image directly from a high-resolution digital camera, this approach uses signal processing techniques. First, several subsampled, and misregistered low-resolution images of the desired scene are obtained. These low-resolution images can be either obtained as a sequence taken over time, or taken at the same time with different sensors. A pictorial example of the overlay of three misregistered images is shown in Figure 4.1, where the sub-pixel shifts for each frame, δ_x, δ_y, are also shown. Figure 4.2 shows the relationship between the subsampled, multiple low-resolution images and the high-resolution image.

The main reason that a single high-resolution frame can be constructed from low-resolution frames is that the low-resolution images are subsampled (aliased) as well as misregistered with sub-pixel shifts. If the images are shifted by integer amounts, then each image contains the same information (intensity values at the same spatial location), and thus there is no new information that can be used. In this case, a simple interpolation scheme (bilinear, cubic spline, etc.) can be used to increase the resolution. However, if the images have sub-pixel shifts, and if aliasing is present, then each image cannot be obtained from the others, assuming each image has different shifts. New information is therefore contained in each low-resolution image, and can thus be exploited to obtain a high-resolution image.

In what follows, we review the relationships between high and low-resolution images first in the spatial and then in the frequency domain in Sections 1.1 and 1.2, respectively. In Section 2 we present a literature review of the solution approaches to this problem. In Section 3 we present the image model that we use. In Sections 4 and 5 we present the two approaches that we have proposed to solve this problem. Section 6 contains experimental results and Section 7 concludes this chapter. In the rest of this chapter the term interpolation will be used to describe the process of computing samples, of the high-resolution image from the low-resolution images, either in the spatial or the frequency domain.

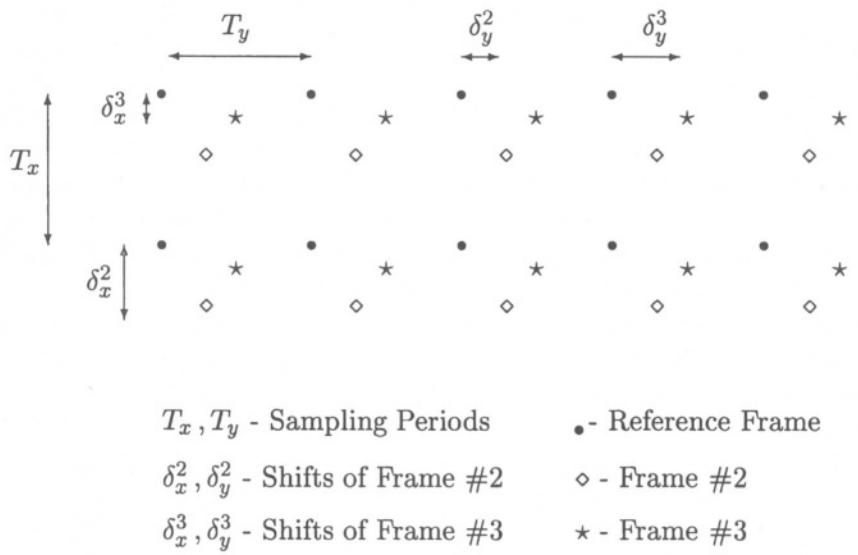


Figure 4.1. Overlay of three misregistered images $(\delta_x^1 = \delta_y^1 = 0)$

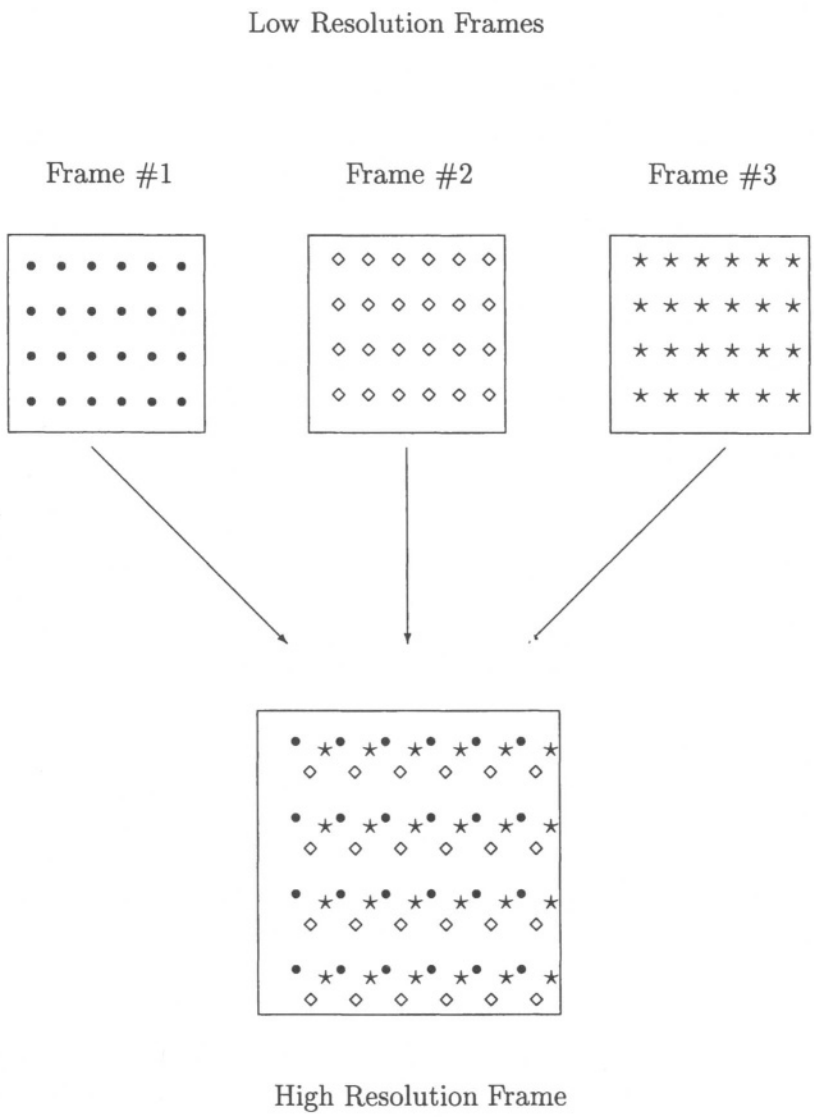


Figure 4.2. Relationship between low-resolution images and the high-resolution image

1.1. Spatial Domain Interpolation

Let us denote by $f(x,y)$ the continuous two-dimensional (2D) image, by $f_{HR}(m,n)$ the high-resolution discrete image of size $2L_xN \times 2L_yN$ and by $f^l_{LR}(m,n)$ the *l-th* low-resolution discrete image of size $N \times N$. They are related by

$$f_{HR}(m,n) = f(mT_x, nT_y)\ , \tag{4.1}$$

where T_x and T_y are the sampling periods in the x and y directions, and

$$f^l_{LR}(m,n) = f(mT'_x + \delta^l_x, nT'_y + \delta^l_y)\ , \quad \text{for } l = 1,2,\ldots P, \tag{4.2}$$

with the sampling periods T'_x and T'_y given by $T'_x = 2L_xT_x, T'_y = 2L_yT_y$ and δ^l_x and δ^l_y represent the shifts in the x and y direction of the lth low-resolution image with respect to a reference image. respectively. direction of the lth low-resolution image with respect to a reference image. Then the equation relating the high and the low-resolution images, also known as the interpolation equation, is given by [10]

$$f^l_{LR}(m,n) = \sum_{m_1=0}^{2L_xN-1} \sum_{n_1=0}^{2L_yN-1} f_{HR}(m_1,n_1) \frac{\sin\left[W_x\left(\delta^l_x + mT'_x - m_1T_x\right)\right]}{W_x\left(\delta^l_x + mT'_x - m_1T_x\right)} \cdot \frac{\sin\left[W_y\left(\delta^l_y + nT'_y - n_1T_y\right)\right]}{W_y\left(\delta^l_y + nT'_y - n_1T_y\right)}\ , \ l = 1,2,\ldots P, \tag{4.3}$$

where W_x and W_y are defined by

$$W_x = \frac{\pi}{T_x},\ \ W_y = \frac{\pi}{T_y}\ . \tag{4.4}$$

Ordering lexicographically indices m, n and l, equation (4.3) can be written in matrix-vector form as

$$\mathbf{f}_{LR} = \phi(\boldsymbol{\delta}) \cdot \mathbf{f}_{HR}, \tag{4.5}$$

where the vectors $\mathbf{f}_{LR}$ and $\mathbf{f}_{HR}$ are of dimensions $PN^2 \times 1$ and $4L_xL_yN^2 \times 1$, respectively, and $\phi(\boldsymbol{\delta})$ is the $PN^2 \times 4L_xL_yN^2, P \geq 4L_xL_y$ interpolation operator between the $P(N \times N)$ and the $2L_xN \times 2L_yN$ grids, and the shifts are given by

$$\begin{aligned} \boldsymbol{\delta} &= \left[\boldsymbol{\delta}^1, \boldsymbol{\delta}^2, \ldots, \boldsymbol{\delta}^P\right] \\ \boldsymbol{\delta}^j &= \left[\delta^j_x, \delta^j_y\right]\ , \ l = 1,2,\ldots P. \end{aligned}$$

The problem therefore at hand is, given the observation vector $\mathbf{f}_{HR}$ and the matrix $\phi(\boldsymbol{\delta})$, assuming the shifts are known, to solve Eq. (4.5) for $\mathbf{f}_{HR}$. In this case, it is in general possible to find the least-squares solution of Eq. (4.5), or the exact solution if $P = 4L_xL_y$ and $\phi(\boldsymbol{\delta})$ is invertible. This solution, however, is a computationally formidable task due to the sizes of the matrices involved. Additionally, in many applications the shifts are not known, and therefore need to be estimated from the available data. Even worse, in certain applications the low-resolution images are degraded due to deterministic blur and noise.

Equation (4.3) is pictorially represented in Fig. 1.3 for an one dimensional (1D) signal, for $L_x = 1$. Figure 1.3 (a) shows f^1_{LR} (bold dotted lines) obtained by multiplying the high-resolution signal by the interpolation kernel (dotted line). Since the shift is zero for this low-resolution image, interpolation corresponds to subsampling by a factor of 2. Figure 1.3 (b) shows f^2_{LR} obtained when the shift δ is not zero. The dotted line again in Fig. 1.3(b) depicts the shifted by δ sinc function which is the interpolation kernel that generates $f^2_{LR}(0)$ according to Eq. (4.3).

1.2. Frequency Domain Interpolation

The low to high-resolution problem is now described in the frequency domain. To the best of our knowledge, Tsai and Huang [49] first introduced the low to high-resolution problem in the frequency domain. Let $f^l(x,y) = f(x + \delta^l_x, y + \delta^l_y)$, $l = 1, 2, \cdots P$, be a set of spatially shifted version of the continuous image $f(x, y)$. Then in the Fourier domain, we have

$$F^l(u,v) = \exp\{2j\pi(\delta^l_x u + \delta^l_y v)\}F(u,v). \tag{4.6}$$

Image $f^l(x,y)$ is now sampled with the sampling periods T'_x and T'_y to generate $f^l_{LR}(m,n)$. Its $N \times N$ discrete Fourier transform (DFT) is given by

$$F^l_{LR}(i,k) = \sum_{m=0}^{N-1}\sum_{n=0}^{N-1} f^l_{LR}(m,n)\exp\left[-2j\frac{\pi}{N}(mi+nk)\right], i,k = 0,\ldots N-1. \tag{4.7}$$

Due to the sampling theorem, the continuous and discrete Fourier transforms are related by [10]

$$F^l_{LR}(i,k) = \frac{1}{T'_xT'_y}\sum_{m=-\infty}^{+\infty}\sum_{n=-\infty}^{+\infty} F^l\left[\frac{2\pi}{T'_x}(\frac{i}{N}+m), \frac{2\pi}{T'_y}(\frac{k}{N}+n)\right]. \tag{4.8}$$

If the Fourier transform of the original image, $F(u, v)$, is bandlimited such that it satisfies $|F(u,v)| = 0$ for $|u| \geq \frac{\pi}{T_x}$ and $|v| \geq \frac{\pi}{T_y}$ (i.e.,

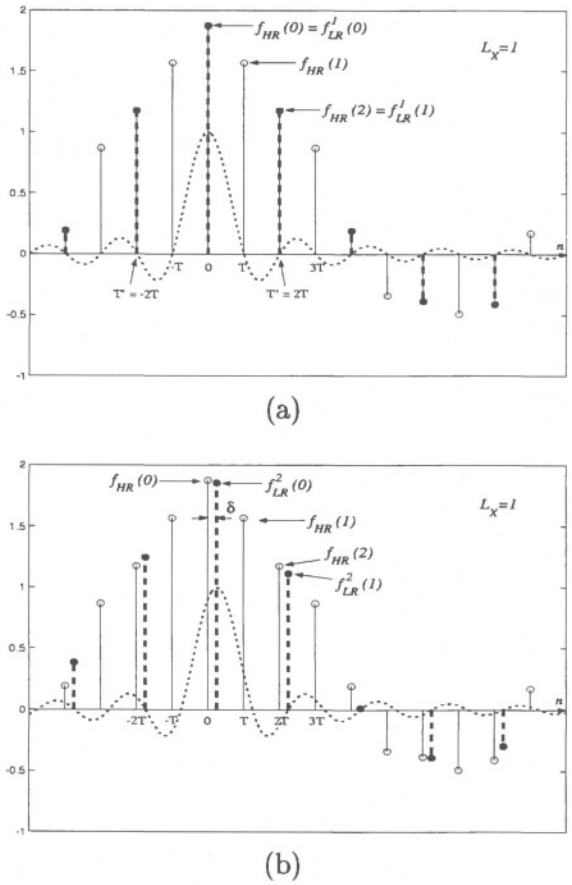


Figure 4.3. Graphical representation of the relationship between high and low-resolution signals (1D version of Eq. (4.3) with $L_x = 1$). Thin lines indicate the samples of the high-resolution signal, bold dotted lines the low-resolution signal, and dotted lines the interpolating kernel-sinc function. **(a)** $\delta = 0$; **(b)** $\delta \neq 0$.

$T_x = \frac{T_x'}{2L_x}$ and $T_y = \frac{T_y'}{2L_y}$ are the Nyquist sampling periods in the *x,y* directions), the DFT of the low-resolution signal in equation (4.8) is aliased. Then, each discrete $F_{LR}^l(i,k)$ is the sum of $2L_x \times 2L_y = 4L_xL_y$ discrete samples of the original $F(u, v)$ according to the equation

$$F_{LR}^l(i,k) = \frac{1}{T_x'T_y'} \sum_{m=-L_x}^{L_x-1} \sum_{n=-L_y}^{L_y-1} F^l \left[\frac{2\pi}{T_x'}(\frac{i}{N} + m), \frac{2\pi}{T_y'}(\frac{k}{N} + n) \right]. \quad (4.9)$$

Substituting $F^l(u,v)$ from Eq. (4.6) into Eq. (4.9) we obtain

$$F_{LR}^l(i,k) = \frac{1}{T_x'T_y'} \sum_{m=-L_x}^{L_x-1} \sum_{n=-L_y}^{L_y-1} \exp\left[j2\pi\{\frac{\delta_x^l}{T_x'}(\frac{i}{N} + m) + \frac{\delta_y^l}{T_y'}(\frac{k}{N} + n)\} \right] \cdot F\left[\frac{2\pi}{T_x'}(\frac{i}{N} + m), \frac{2\pi}{T_y'}(\frac{k}{N} + n) \right]. \quad (4.10)$$

Using lexicographic ordering for the indexes m, n in the right hand side and l in the left hand side of equation (4.10), we obtain in matrix vector form

$$\mathbf{F}_{LR}^{(i,k)} = \theta(\boldsymbol{\delta})^{(i,k)} \cdot \mathbf{F}^{(i,k)}, \quad (4.11)$$

where the dimensions of $\mathbf{F}_{LR}^{(i,k)}$, $\mathbf{F}^{(i,k)}$, and $\theta(\boldsymbol{\delta})^{(i,k)}$ are $P \times 1$, $4L_xL_y \times 1$ and $P \times 4L_xL_y$, respectively. The low to high-resolution problem again is to solve Eq. (4.11) for $\mathbf{F}_{LR}^{(i,k)}$ and the interpolation matrix $\theta(\boldsymbol{\delta})^{(i,k)}$, assuming the shifts are known. A major difference among the spatial domain formulations is that each discrete frequency of the high-resolution signal is recovered separately, which results in the solution (or inversion) of small $P \times P$ (if $P = 4L_xL_y$) matrices.

A pictorial representation of the relationship between the low and high resolution signals in the frequency domain is shown in Fig. 1.4, for an 1D signal, with $L_x = 1$ and $L_x = 2$. Figure 1.4(a) shows the spectrum of the continuous signal $F(u)$. Figures 1.4 (b) and (c) show one period of the continuous spectrum of the low-resolution signal for $L_x = 1$ and $L_x = 2$, respectively. In the first case ($L_x = 1$, i.e., subsampling by 2), two shifted versions of the spectrum $F(u)$ are used to form $F_{LR}^1(u)$, while in the second case ($L_x = 2$, i.e., subsampling by a factor of four) four shifted versions are used to form $F_{LR}^1(u)$. The sampled versions of the spectra shown in Figs. 1.4 (b), (c) follow from Eq. (4.10).

2. Literature Review

From the previous analysis it is clear that interpolation is possible both in the spatial and frequency domains. However, in either case

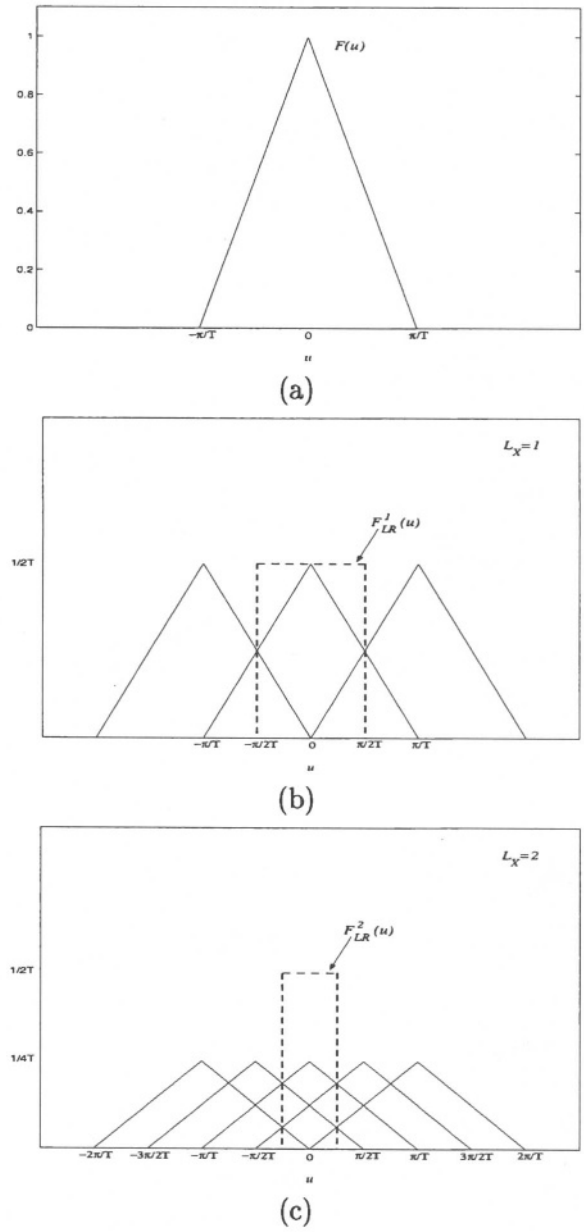


Figure 4.4. Graphical representation of the relationship between high and low-resolution signals in the frequency domain. (a) Spectrum of one dimensional high-resolution signal; spectra of the low-resolution signals for (b) $L_x = 1$, and (c) $L_x = 2$, respectively.

knowledge of the shifts of the low-resolution images is necessary. In most practical applications, in addition to the fact that these shifts are unkown, the low-resolution images might be degraded by blur and noise. Thus, before interpolation to reconstruct the high-resolution image it is necessary to restore the low-resolution images. Therefore, in practice reconstruction of high-resolution images involves the following three tasks: restoration of the low-resolution images, registration of the low-resolution images to find the shifts, and interpolation to reconstruct the high-resolution image.

The individual tasks associated with the problem of reconstructing high from low-resolution images (registration, restoration, interpolation) are on their own important problems that have been extensively researched. For the registration task, see for example [3, 5, 31, 33], for restoration in general, good reviews can be found in [23, 24, 4], while for interpolation, see for example [7, 9, 12, 28, 39, 50]. Since all low-resolution images are very similar, optimal restoration results are obtained if a multi-channel approach is used. According to this approach, the correlations between the low-resolution images (channels) are utilized in addition to within channel correlations. For a recent review on multi-channel image recovery approaches see [17]. In spite of this, significantly better results can be expected if instead of addressing each of these tasks independently a more comprehensive approach is taken. This is because all of the tasks associated with the problem of reconstructing a high-resolution image from low-resolution images are coupled. The general problem of reconstructing a single high-resolution image from multiple low-resolution images (all of which are taken of the same scene) has been investigated by several researchers, some of which are Frieden and Aumann [13], Stark and Oskoui [40], Aizawa *et. al* [1, 2], Tsai and Huang [49], Irani and Peleg [20], Tekalp *et. al* [41], Srinivas and Srinath [38], Kim *et. al* [30], Kim and Su [29], and Bose *et. al.* [6]. The differences among these works lie in the method used, assumptions made, and degree of degradations (if any) incurred by the sensor.

Of these approaches, the earliest comprehensive work was by Tsai and Huang [49]. They derived an interpolation equation that described the low-resolution image as a function of the high-resolution image and the shifts between the low resolution images in the Fourier domain. They also proposed a registration approach based on minimizing the energy of the high-resolution signal However, they did not consider the case when the images were noisy or blurred, and because of this, neglected to address the issue of inverting the interpolation matrix, which could prove difficult in the presence of noise. They also proposed an approach

to estimate the shifts based on minimizing the energy of the interpolated high-resolution signal outside the assumed bandwidth.

Kim, *et. al* [29, 30] extended this work to include noise and blur. In order to account for the blur in [30], Tikhonov regularization [42] was employed. Thus, they addressed both the restoration and interpolation sub-problems together. However, the issue of choosing the optimum regularization parameter was not addressed. Also, the cross-correlation terms were not used to restore the degraded images, which, according to [14, 47], is sub-optimal because of the additional information that they contain. In addition, it was assumed that the shifts were known.

Similarly, Srinivas and Srinath [38] combined the restoration and interpolation steps together, while assuming that the shifts were known exactly. A distinguishing feature from [6, 29, 30] is that they formulated the restoration sub-problem in a multi-channel framework, recognizing that the restoration could be improved if the cross-correlation terms were used. This agreed with results found in [14, 47].

In [41, 35] both Frieden's frequency domain method [13] and Stark's projection onto convex sets (POCS) method [40] were extended in order to account for both sensor noise and blur. A method was also proposed where the interpolation step was performed before the restoration step. The problem of reconstructing a band-limited signal from nonuniformly sampled points has been previously discussed in the literature [12, 28, 39, 50]. Following this interpolation step, the Wiener filter was used to restore the image (actually, any general restoration algorithm could have been used). However, this method still requires that the shifts be known.

A completely different approach to estimate the high-resolution image was developed by Irani and Peleg [20]. In this work, a method to estimate the displacements was presented, based on their earlier work (they also assume that the images could be slightly rotated with respect to the reference image). Instead of using a standard interpolating matrix to obtain the high-resolution image, they chose an iterative technique similar to the back-projection method commonly used in computed tomography. Experimental results of improved resolution was presented for both gray-scale and color images. They also showed that the high-resolution problem reduces to the standard restoration problem when the input is a single image, and no upsampling is necessary. An estimate of the point spread function (PSF) was obtained by evaluating control images (i.e., the literal definition of point spread function). However, additive noise is handled by simply averaging all of the contributions of the low-resolution pixels, and is not explicitly accounted for in the iterations. A similar low to high-resolution problem described in this

chapter so far can be formulated when multiple frames in a dynamic image sequence are considered (see for example, [37], [48]). The subpixel shifts in this case are due to the motion of objects which is represented by motion vectors of sub-pixel accuracy. This is a problem described in other chapters in this book, and is therefore not considered here. In [11] maximum *a posteriori* (MAP) estimation and POCS were applied to this problem. However, the imaging model assumed perfectly registered images and thus bypassed the difficulty of registration.

In summary, by reviewing the literature on this problem it is apparent that although the problem has a long history, very little work has been done in addressing this problem in a comprehensive manner. In other words, no comprehensive framework has been presented for combining all of the tasks that are involved in this problem, with the exception of possibly [19]. In [19] a MAP framework is proposed where simultaneous restoration, registration and interpolation was performed. A block-matching like algorithm was used to estimate the shifts between the low-resolution frames. However, to simplify computations a suboptimal optimization strategy was used according to which the high-resolution image and the shifts were estimated one at a time while keeping the other fixed.

In this chapter we propose two formulations to solve the high-resolution problem in a comprehensive manner. Both approaches combine the registration and restoration steps into a single step, thus, they are solved *simultaneously*. In particular, a multi-channel blur identification and restoration algorithm [47] is used to estimate the displacements while simultaneously restoring the image(s). This multi-channel framework improves the estimates of the shifts while simultaneously estimates the original, undegraded image.

The difference between these two formulations lies in the interpolation step. The first approach, called the **RR-I** (Registration Restoration - Interpolation) formulation, performs the interpolation step independently of the first two steps. This approach can be implemented easily by slightly modifying the multi-channel restoration and identification algorithm in [47]. The second approach, called **RRI** (Registration Restoration Interpolation), formulates all three sub-problems into a single equation, yielding an optimal solution at the expense of the increased computational cost. These two approaches are presented in Sections 4 and 5 of this chapter.

3. Imaging Model

Let P be the number of the available low-resolution images, and let each low-resolution image be of size $N \times N$. The discrete image observed at the ith sensor is given by

$$\mathbf{g}_{LR}^{i} = \sum_{j=1}^{P} H_{ij} \cdot \mathbf{f}_{LR}^{j} + \mathbf{v}^{i} \ , \quad i, j = 1, \ldots, P \ , \tag{4.12}$$

where $\mathbf{g}_{LR}^{i}$, $\mathbf{f}_{LR}^{i}$, and $\mathbf{v}^{i}$ represent the $N \times N$ observed, original and noise images, respectively, at the ith sensor (channel), and H_{ij} the discretized impulse response modeling *both* the within ($i = j$) and between ($i \neq j$) channel degradation mechanisms.

Equation (4.12) represents a multi-channel degradation model, that also allows cross-channel degradations. It can be written in matrix vector form as

$$\mathbf{g}_{LR} = \mathbf{H} \cdot \mathbf{f}_{LR} + \mathbf{v} \ , \tag{4.13}$$

where $\mathbf{g}_{LR}$ and $\mathbf{f}_{LR}$ are $PN^2 \times 1$ vectors given by

$$\mathbf{g}_{LR}^{T} = \left[\left(\mathbf{g}_{LR}^{1}\right)^{T}, \left(\mathbf{g}_{LR}^{2}\right)^{T}, \ldots, \left(\mathbf{g}_{LR}^{P}\right)^{T} \right] \tag{4.14}$$

$$\mathbf{f}_{LR}^{T} = \left[\left(\mathbf{f}_{LR}^{1}\right)^{T}, \left(\mathbf{f}_{LR}^{2}\right)^{T}, \ldots, \left(\mathbf{f}_{LR}^{P}\right)^{T} \right] \ , \tag{4.15}$$

and $\mathbf{g}_{LR}^{i}$ and $\mathbf{f}_{LR}^{i}$ are the lexicographic orders of $g_{LR}^{i}(m,n)$ and $f_{LR}^{i}(m,n)$, respectively. $\mathbf{H}$ is a $PN^2 \times PN^2$ matrix which has the form

$$\mathbf{H} = \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1P} \\ H_{21} & H_{22} & \cdots & H_{2P} \\ \vdots & \vdots & \vdots & \vdots \\ H_{P1} & H_{P2} & \cdots & H_{PP} \end{bmatrix} \ , \tag{4.16}$$

where each sub-matrix, H_{ij} is of size $N^2 \times N^2$, and represents the PSF between the ith and jth sensors. Since within each sensor the degradation operator represents a linear convolution, matrices H_{ij} are block-circulant. However, because the shifts are incorporated into the degradation operator, and are spatially varying, i.e., the shifts between grids i and j are not the same between grids $i + k$ and $j + k$, so that $H_{i+k,j+k} \neq H_{i,j}$, $\mathbf{H}$ is not block-circulant. Matrices of this form have been studied in the context of multi-channel problems and are named block semi-circulant (BSC) [14, 27, 17].

A different form of this matrix structure can be obtained if the channel index is used for ordering first, followed by the spatial indices. The

multi-channel vector is then given by

$$\mathbf{f}_{LR}^{T} = \left[f_{LR}^{1}(0)\ldots f_{LR}^{P}(0)f_{LR}^{1}(1)\ldots f_{LR}^{P}(N^2-1)\right]^{T} . \tag{4.17}$$

Using the definition

$$\mathbf{f}_{LR}(i) \triangleq \left[f_{LR}^{1}(i)f_{LR}^{2}(i)\ldots f_{LR}^{P}(i)\right]^{T} , \quad i = 0,\ldots,N^2-1 , \tag{4.18}$$

we can now write Eq. (4.17) as

$$\mathbf{f}_{LR}^{T} = \left[\mathbf{f}_{LR}^{T}(0)\ \mathbf{f}_{LR}^{T}(1)\ldots\mathbf{f}_{LR}^{T}(N^2-1)\right] . \tag{4.19}$$

When this lexicographic ordering is applied to the linear degradation model in Eq. (4.13) **H** the $PN^2 \times PN^2$ linear degradation operator has the form

$$\mathbf{H} = \begin{bmatrix} H(0) & H(1) & \cdot\cdot & H(N^2-1) \\ H(N^2-1) & H(0) & \cdot\cdot & H(N^2-2) \\ \vdots & \vdots & \cdot\cdot & \vdots \\ H(1) & H(2) & \cdot\cdot & H(0) \end{bmatrix} , \tag{4.20}$$

where the $P \times P$ sub-matrices (sub-blocks) have the form

$$H(m) = \begin{bmatrix} H_{11}(m) & H_{12}(m) & \cdot\cdot & H_{1P}(m) \\ H_{21}(m) & H_{22}(m) & \cdot\cdot & H_{2P}(m) \\ \vdots & \vdots & \cdot\cdot & \vdots \\ H_{P1}(m) & H_{P2}(m) & \cdot\cdot & H_{PP}(m) \end{bmatrix} , \quad 0 \leq m \leq N^2-1 . \tag{4.21}$$

Note that $H_{ii}(m)$ represents an intra-channel blurring operator, while $H_{ij}(m), i \neq j$ represents the inter-channel blur. Matrix **H** in Eq. (4.20) has a structure which is the dual to that of **H** in Eq. (4.16). In other words, it contains $P \times P$ non-circulant blocks that are arranged in a circulant fashion. Matrices of this form are called semi-block circulant (SBC) [27, 15, 17].

The covariance matrices of multi-channel signals where within-channel but not between-channel stationarity is assumed are also either BSC or SBC depending on the lexicographic ordering that was used to arrange the multi-channel vector $\mathbf{f}_{LR}$ [14, 17, 27].

Mathematical expressions containing SBC matrices and multi-channel vectors can be computed very efficiently in the DFT domain. $PN^2 \times N^2P$ SBC matrices are transformed in the DFT domain by the transformation

$$\mathcal{W}_{\mathcal{S}} = I_P \otimes W \tag{4.22}$$

where I_P is the $P \times P$ identity, W the $N^2 \times N^2$ 2-D DFT matrix and $\otimes$ the Kronecker product. Application of this transformation to $\mathbf{H}$ a $PN^2 \times N^2P$ SBC matrix and use of the diagonalization properties of the DFT for circulant matrices gives

$$\mathcal{W}_S^{-1}\mathbf{H}\mathcal{W}_S = \begin{bmatrix} S_0 & 0 & \cdots & 0 \\ 0 & S_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & S_{N^2-1} \end{bmatrix}, \tag{4.23}$$

where S_i a $P \times P$ matrix. In what follows the SBC representation will be used to convert equations in the DFT domain.

Substituting Eq. (4.5) into Eq. (4.13), the final equation relating all three sub-problems can be written as

$$\mathbf{g}_{LR} = \mathbf{H} \cdot \phi(\boldsymbol{\delta})\, \mathbf{f}_{HR} + \mathbf{v}\,. \tag{4.24}$$

Equation (4.13) is the starting equation for the **RR-I** formulation, where $\mathbf{f}_{LR}$ is to be solved for as an intermediate step. The next and final step is to solve Eq. (4.5) for the high resolution image, $\mathbf{f}_{HR}$, given $\mathbf{f}_{LR}$. In the **RRI** formulation, Eq. (4.24) is the governing equation, where $\mathbf{f}_{HR}$ is to be solved for directly from the observable, noisy images, $\mathbf{g}_{LR}$.

4. Simultaneous Registration and Restoration, RR-I Approach

In this section, the **RR-I** approach is presented [43, 44, 46]. This approach, shown in Fig. 4.5, solves the first two sub-problems simultaneously, and performs the interpolation step independently. More specifically, this approach first estimates $\mathbf{f}_{LR}$ and shifts $\boldsymbol{\delta}$ from Eq. (4.13) and then via interpolation reconstructs the high-resolution image, i.e., finds $\mathbf{f}_{HR}$ by inverting Eq. (4.5). In other words, the restoration and the registration step are combined. This is accomplished by selecting an image covariance model which is parameterized by the shifts. This model is estimated simultaneously with the restored low-resolution images in an iterative maximum likelihood framework that is based on the Expectation Maximization (EM) algorithm [8, 25].

4.1. Image Covariance Model

The incorporation of the sub-pixel shifts into the likelihood function is achieved by assuming a particular image covariance model for the low-resolution frames. A popular covariance model is the separable co-

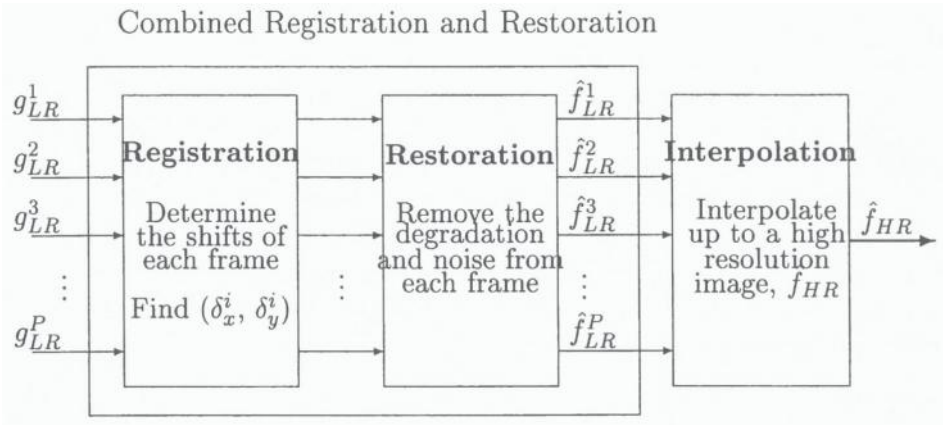


Figure 4.5. Block diagram of constructing a high-resolution frame from multiple low resolution frames, **RR-I** approach

variance model, given by [21]

$$r(m,n) = \sigma_f^2 \rho_1^{|m|} \rho_2^{|n|} \tag{4.25}$$

where $r(m,n)$ is the (m,n)th spatial location of the image covariance, σ_f^2 is the variance of the random field, and ρ_1, ρ_2 are the correlation coefficients in the x and y directions, respectively. This separable characteristic is highly desirable, and will prove to be very useful.

The covariance $\Lambda_{\mathbf{f}_{LR}}$ of the multi-channel low-resolution image is

$$\Lambda_{\mathbf{f}_{LR}} = \mathrm{E}\left[\mathbf{f}_{LR}\,\mathbf{f}_{LR}^H\right], \tag{4.26}$$

where E is the expectation operator. From Eq. (4.25), $\Lambda_{\mathbf{f}_{LR}}$ is block Toeplitz. Using Eq. (4.26) and (4.25), the entries of the sub-block of the image covariance are given by

$$[r\,(m,n)]_{i,j} = \sigma_f^2 \rho_1^{|i-j|} \rho_2^{|m-n|}\ , \tag{4.27}$$

where $[r(m,n)]_{i,j}$ is the (m,n)th element of the (i,j)th block. Using Eq. (4.27), when the lexicographic ordering in Eq. (4.15) is used for $\mathbf{f}_{LR}$, the covariance matrix of the low-resolution image can be shown to be

$$\Lambda_{\mathbf{f}_{LR}} = \sigma_f^2 S \otimes R_{1,f_{LR}} \otimes R_{2,f_{LR}}\ , \tag{4.28}$$

where $\otimes$ again denotes the Kronecker product of two matrices and the matrices $S, R_{1,f_{LR}}$, and $R_{2,f_{LR}}$ are respectively of sizes $P \times P$, $N \times N$, and $N \times N$. Their respective (i,j)th elements are equal to $\rho_1^{|\delta_x^i - \delta_x^j|}\rho_2^{|\delta_y^i - \delta_y^j|}$, ρ_1^{i-j} and ρ_2^{i-j}. Clearly the displacements are incorporated into the covariance matrix. The Kronecker product arises from the separability of the covariance model. The matrices $R_{1,f_{LR}}$ and $R_{2,f_{LR}}$ are $N \times N$ Toeplitz. However, as N grows and as we move away from the main diagonal $i = j$, ρ_1^{i-j} and ρ_2^{i-j} go to zero. Asymptotically then these Toeplitz can be approximated by circulant matrices [18]. The Kronecker product $R_{1,f_{LR}} \otimes R_{2,f_{LR}}$ is a $N^2 \times N^2$ block-circulant matrix. Thus, $\Lambda_{\mathbf{f}_{LR}}$ is a special case of a $PN^2 \times N^2P$ BSC matrix because of the Kronecker product based decomposition. In general BSC and SBC matrices do not have a Kronecker based decomposition.

4.2. Maximum Likelihood based Restoration-Registration

Multi-channel linear minimum mean squared error (LMMSE) restoration will be used to restore the low-resolution images $\mathbf{f}_{LR}$ from the observed images $\mathbf{g}_{LR}$. For this purpose knowledge of the covariance of the signal, $\Lambda_{\mathbf{f}_{LR}}$, and noise, $\Lambda_{\mathbf{v}}$, and the linear degradation $\mathbf{H}$ are necessary. Maximum likelihood (ML) estimation will be used for their estimation, $\mathbf{f}_{LR}$ and $\mathbf{v}$ are assumed uncorrelated Gaussian random processes, thus the observed image, $\mathbf{g}_{LR}$, is also Gaussian with zero mean, and pdf given by

$$f_G(\mathbf{g}_{LR}) = |2\pi(\mathbf{H}\Lambda_{\mathbf{f}_{LR}}\mathbf{H}^H + \Lambda_{\mathbf{v}})|^{-1/2} \cdot \exp\left\{-\frac{1}{2}\mathbf{g}_{LR}^H(\mathbf{H}\Lambda_{\mathbf{f}_{LR}}\mathbf{H}^H + \Lambda_{\mathbf{v}})^{-1}\mathbf{g}_{LR}\right\} . \quad (4.29)$$

To emphasize that $f_G(\mathbf{g}_{LR})$ is parameterized, we rewrite it as $f_G(\mathbf{g}_{LR}; \phi)$, where $\phi = \{\Lambda_{\mathbf{f}_{LR}}, \Lambda_{\mathbf{v}}, \mathbf{H}\}$ represents the quantities of interest. The ML estimation of this parameter set is that set ϕ_{ML} which maximizes the likelihood function $f_G(\mathbf{g}_{LR}; \phi)$, or its logarithm, that is

$$\phi_{ML} = \arg\left\{\max_{\phi} f_G(\mathbf{g}_{LR}; \phi)\right\} . \quad (4.30)$$

Taking the logarithm of Eq. (4.29) and disregarding constant multiplicative and additive terms, the maximization of the log-likelihood function becomes the minimization of the function $L(\phi)$, given by

$$L(\phi) = \log|\mathbf{H}\Lambda_{\mathbf{f}_{LR}}\mathbf{H}^H + \Lambda_{\mathbf{v}}| + \mathbf{g}_{LR}^H(\mathbf{H}\Lambda_{\mathbf{f}_{LR}}\mathbf{H}^H + \Lambda_{\mathbf{v}})^{-1}\mathbf{g}_{LR} . \quad (4.31)$$

However, minimizing $L(\phi)$ explicitly as written in Eq. (4.31) with respect to $\mathbf{H}$, $\Lambda_{\mathbf{f}_{LR}}$, and $\Lambda_{\mathbf{v}}$ is a difficult problem due to the size of the matrices involved as well as its high degree of nonlinearity. The alternative and more suitable approach is to transform Eq. (4.31) into the frequency domain, and use an iterative technique, such as the EM algorithm, to minimize this objective function.

The likelihood function, after discarding constant terms, can be written as

$$L(\vartheta) = \log \left| \mathbf{H}\Lambda_{\mathbf{f}_{LR}}\mathbf{H}^T + \Lambda_{\mathbf{v}} \right| + \mathbf{g}_{LR}^T \left(\mathbf{H}\Lambda_{\mathbf{f}_{LR}}\mathbf{H}^T + \Lambda_{\mathbf{v}} \right)^{-1} \mathbf{g}_{LR} , \quad (4.32)$$

where ϑ represents the unknown parameter set, given by $\vartheta = \{\sigma_i^2, \rho_1, \rho_2, \sigma_f^2, \delta\}$.

For linear Gaussian problems the application of the EM algorithm is well studied. In [25, 32] the EM algorithm was applied to the single-channel linear image restoration blur identification problem. In [46] the EM algorithm was applied to the multi-channel image restoration and blur identification problem.

Without going through the analysis steps, we present here the results derived in [46, 47]. If as complete data we select $(\mathbf{f}_{LR}^T, \mathbf{g}_{LR}^T)$, the function that has to be minimized iteratively, instead of the likelihood in Eq. (4.32), for obtaining the estimates of ϑ at the $k+1$ iteration is

$$\begin{aligned} L(\vartheta; \vartheta^{(k)}) = \log \left| \Lambda_{\mathbf{f}_{LR}} \right| + \operatorname{tr} \left\{ \left(\Lambda_{\mathbf{f}_{LR}}^{-1} + \mathbf{H}^T\Lambda_{\mathbf{v}}^{-1}\mathbf{H} \right) \Lambda_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}^{(k)} \right\} \\ + \mu_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}^{(k)T} \left(\Lambda_{\mathbf{f}_{LR}}^{-1} + \mathbf{H}^T\Lambda_{\mathbf{v}}^{-1}\mathbf{H} \right) \mu_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}^{(k)} \\ - \left(\mathbf{g}_{LR}^T\Lambda_{\mathbf{v}}^{-1}\mathbf{H}\mu_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}^{(k)} + \mu_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}\mathbf{H}^T\Lambda_{\mathbf{v}}^{-1}\mathbf{g}_{LR} \right) + \mathbf{g}_{LR}^T\Lambda_{\mathbf{v}}^{-1}\mathbf{g}_{LR} , \end{aligned} \quad (4.33)$$

where $\Lambda_{\mathbf{f}_{LR}}^{(k)}$ and $\Lambda_{\mathbf{v}}^{(k)}$ are the estimates of $\Lambda_{\mathbf{f}_{LR}}$ and $\Lambda_{\mathbf{v}}$ at the k iteration and are parameterized by $\vartheta^{(k)}$. The conditional mean and covariance are given by

$$\mu_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}^{(k)} = \Lambda_{\mathbf{f}_{LR}}^{(k)}\mathbf{H}^T \left(\mathbf{H}\Lambda_{\mathbf{f}_{LR}}^{(k)}\mathbf{H}^T + \Lambda_{\mathbf{v}}^{(k)} \right)^{-1} \mathbf{g}_{LR} \quad (4.34)$$

$$\Lambda_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}^{(k)} = \Lambda_{\mathbf{f}_{LR}}^{(k)} - \Lambda_{\mathbf{f}_{LR}}^{(k)}\mathbf{H} \left(\mathbf{H}\Lambda_{\mathbf{f}_{LR}}^{(k)}\mathbf{H}^T + \Lambda_{\mathbf{v}}^{(k)} \right)^{-1} \mathbf{H}\Lambda_{\mathbf{f}_{LR}}^{(k)} \quad (4.35)$$

The EM algorithm is an iterative algorithm and consists of the E and M steps. During the E-step of the k^{th} iteration the estimates $\vartheta^{(k)}$ (the shifts and the parameters, $\rho_1, \rho_2, \sigma_f^2$, which define the covariance of $\mathbf{f}_{LR}$ as well as the noise variance, σ_i^2) are used to compute the conditional

mean $\mu^{(k)}_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}$ and conditional covariance $\Lambda^{(k)}_{\mathbf{f}_{LR}|\mathbf{g}_{LR}}$. During the M-step of the $(k+1)^{th}$ iteration the conditional mean and covariances from the k^{th} iteration are used in the objective function, $L(\vartheta;\vartheta^{(k)})$ which is then minimized with respect to ϑ. These values of ϑ form then $\vartheta^{(k+1)}$, the new parameter estimates. Because of the Gaussian assumption and the linearity of the observation model the conditional mean is the Linear Minimum Mean Square Error (LMMSE) estimate of $\mathbf{f}_{LR}$. Thus the restored image and the unknown parameters are updated in every iteration.

4.3. Formulation of L in the frequency domain

One important issue that remains to be addressed is the implementation of the EM algorithm for this problem. Direct computation of Eqs. (4.33), (4.34) and (4.35) is not possible due to the very large size of the matrices involved. To solve this problem, it is first trasformed in the DFT domain based on the lexicographic ordering in Eq. (4.17), which yields SBC $PN^2 \times N^2P$ matrices. These matrices when transformed in the DFT domain give SBD matrices, as explained earlier.

The covariance of the multi-channel low-resolution vector, because of the Kronecker decomposition gives a very simple SBD DFT domain representation. This representation using the Kronecker product properties is given by

$$\mathcal{W}_{\mathcal{S}}{}^{-1}\Lambda_{\mathbf{f}_{LR}}\mathcal{W}_{\mathcal{S}} = \sigma_f^2\mathcal{W}_{\mathcal{S}}{}^{-1}R\otimes S\mathcal{W}_{\mathcal{S}} = \sigma_f^2 W^{-1}RW\otimes S\,, \qquad (4.36)$$

where $R = R_1 \otimes R_2$, since R_1 and R_2 are $N \times N$ circulant matrices, R is $N^2 \times N^2$ block-circulant [21], and W the 2D-DFT matrix, with $W = W_1 \otimes W_1$,where W_1 is the $N \times N$ 1D-DFT matrix. Thus,

$$W^{-1}RW = W_1^{-1}R_1W_1 \otimes W_1^{-1}R_2W_1 = \Theta_1 \otimes \Theta_2, \qquad (4.37)$$

where $\Theta_1 = W_1^{-1}R_1W_1 - 1$ is the diagonal matrix of R_1, whose entries are given by the DFT of the first row of R_1, or

$$\Theta_1(m) \;= \sum_{n=0}^{N-1}\rho_1^n w_N^m \;= \frac{1-\rho_1^N}{1-\rho_1 w_N^m}\,,$$

where $w_N = \exp(-\frac{j2\pi}{N})$ and $w_N^N = 1.0$. Similarly, Θ_2 is diagonal with entries

$$\Theta_2(m) \;= \sum_{n=0}^{N-1}\rho_2^n w_N^m \;= \frac{1-\rho_2^N}{1-\rho_2 w_N^m}\,.$$

Thus, $\Theta_F = diag\{\Theta(0,0), \cdots \Theta(N-1, N-1)\}$ with $\Theta_F(m,n) = \Theta_1(m) \cdot \Theta_2(n)S$ a $P \times P$ matrix.

All SBC matrices in Eq. (4.33), (4.34) and (4.35) can be transformed in the DFT domain. Using the properties of the Kronecker product we can then write Eq. (4.33) as

$$\begin{aligned} L(\vartheta; \vartheta^{(k)}) &= PN^2 \log(\sigma_f^2) + PN \log|\Theta_1| + PN \log|\Theta_2| \\ &+ N^2 \log|S| + \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} J(m,n) , \end{aligned} \tag{4.38}$$

where

$$\begin{aligned} J(m,n) &= \log|\Theta_v(m,n)| + \mathrm{tr}\left\{\left(\left(\sigma_f^2 \Theta_1(m)\Theta_2(n)S\right)^{-1}\right.\right. \\ &+ \left.\left.\Theta_H^H(m,n)\Theta_v^{-1}(m,n)\Theta_H(m,n)\right) B(m,n)\right\} \\ &- \frac{1}{N^2}\mathrm{tr}\left\{\Theta_v^{-1}(m,n)\left(\Psi(m,n)\Theta_H^H(m,n)\right.\right. \\ &+ \left.\left.\Theta_H(m,n)\Psi^H(m,n)\right) + G(m,n)G^H(m,n)\right)\right\} , \end{aligned} \tag{4.39}$$

and

$$\begin{aligned} B(m,n) &= \Theta_{F|g}(m,n) + \frac{1}{N^2} M_{F|g}(m,n)\, M_{F|g}^H(m,n) \\ \Psi(m,n) &= G(m,n) M_{F|g}^H(m,n) . \end{aligned} \tag{4.40}$$

In the previous equations $G(m,n) = [G^1(m,n), G^2(m,n) \cdots G^p(m,n)]^T$, a $P \times 1$ vector of the (m, n) frequency of the DFT of $\mathbf{g}_{LR}$, the observed image; $\Theta_{F|g}(m,n)$ and $M_{F|g}(m,n)$ are $P \times P$ matrices and $P \times 1$ vectors, respectively, containing the DFT representations of the SBC conditional covariance and the multi-channel conditional mean. They are given by

$$\begin{aligned} \Theta_{F|g} &= \Theta_F - \Theta_F \Theta_H^H \left(\Theta_H \Theta_F \Theta_H^H + \Theta_V\right)^{-1} \Theta_H \Theta_F \\ M_{F|g} &= \Theta_F \Theta_H^H \left(\Theta_H \Theta_F \Theta_H^H + \Theta_V\right) G , \end{aligned} \tag{4.41}$$

where the frequency notation (m, n) has been dropped for brevity, and $\Theta_H(m,n)$, and $\Theta_V(m,n)$ are $P \times P$ matrices resulting from the DFT representation of the SBC $PN^2 \times N^2P$ $\mathbf{H}$ and $\Lambda_\mathbf{V}$ matrices, respectively.

Thus, in the M-step of the algorithm by minimizing $L(\vartheta; \vartheta^{(k)})$ derived in Eqs. (4.37)-(4.41) with respect to ϑ, we obtain the estimates of $\{\sigma_i, \rho_1, \rho_2, \sigma_f^2, \delta\}$. In the E-step the conditional statistics in Eq. (4.41) are computed. Due to space constraints, we do not present furhter details of the implementation in this chapter (for more details see [46, 47]).

4.4. Iterative Interpolation for the RR-I Formulation

Applying the previously described EM algorithm yields estimates of the displacements and the restored low-resolution images (see [46] for details). The final step is to interpolate the intensity values of the desired high-resolution image. Mathematically, the interpolation equation, Eq. (4.5) needs to be solved. Before discussing the inversion, however, a good model for the interpolation matrix, ϕ, needs to be chosen. The form of this matrix can be given as

$$\phi(\delta) = \begin{bmatrix} \phi_{0,0} & \phi_{0,1} & \cdots & \phi_{0,N} & \cdots & \phi_{0,N^2-1} \\ \phi_{1,0} & \phi_{1,1} & \cdots & \phi_{1,N} & \cdots & \phi_{1,N^2-1} \\ \vdots & & \ddots & & & \vdots \\ \phi_{N^2-1,0} & \phi_{N^2-1,1} & \cdots & \phi_{N^2-1,N} & \cdots & \phi_{N^2-1,N^2-1} \end{bmatrix}, \tag{4.42}$$

where each $4L_xL_y \times P$ sub-matrix $\phi_{i,j}$ is given by

$$\phi_{i,j} = \begin{bmatrix} \phi^1_{i,j}(0) & \phi^1_{i,j}(1) & \cdots & \phi^1_{i,j}(4L_xL_y-2) & \phi^1_{i,j}(4L_xL_y-1) \\ \vdots & & \ddots & & \vdots \\ \phi^P_{i,j}(0) & \phi^P_{i,j}(1) & \cdots & \phi^P_{i,j}(4L_xL_y-2) & \phi^P_{i,j}(4L_xL_y-1) \end{bmatrix}, \tag{4.43}$$

$$\phi^\ell_{ij}(k) = \operatorname{sinc}\left\{\pi\left(\delta^\ell_x + \left\lfloor\frac{i}{N}\right\rfloor 2L_x - \left\lfloor\frac{j}{N}\right\rfloor 2L_x - \left\lfloor\frac{k}{2L_y}\right\rfloor\right)\right\} \cdot$$
$$\operatorname{sinc}\left\{\pi\left(\delta^\ell_y + \bmod(i,N)\,2L_y - \bmod(j,N)\,2L_y - \bmod(k,2L_y)\right)\right\} \tag{4.44}$$

where L_x, L_y are integers specifying the degree of interpolation in the x and y directions, respectively. For $L_x = L_y = 1$ the image size doubles in both dimensions. In the following discussion, $P = 4L_xL_y$, in order to make ϕ a square matrix. In this case the total number of pixels in both the high-resolution image and the four low-resolution images is the same, since the number of pixels in each low-resolution image is one quarter the number of pixels of the high-resolution image. In Eq. (4.44), $\lfloor x \rfloor$ denotes the greatest integer of x, and $\bmod(x,y) \equiv x$ modulo y. As a reminder, the relationships between the sampling period of the low-resolution frame and that of the high-resolution frame were taken to be $\frac{T'_x}{T_x} = 2L_x$, $\frac{T'_y}{T_y} = 2L_y$.

While the lexicographic ordering of the low-resolution images has already been discussed, the lexicographic ordering for the high-resolution image needs to be also specified. Due to the interlacing among the low-resolution images, the ordering for the high-resolution image should keep

the same spatial coordinate system. In this case $\phi(\delta)$ becomes block-Toeplitz at the matrix level, and can therefore be approximated by a semi-block circulant matrix (SBC), which can then be diagonalized by the same transform matrix as in Sec. 3 [46]. Once Eq. (4.5) is written in the frequency domain, it follows that Eq. (4.24) can also be expressed in the frequency domain.

4.5. Regularization

Direct inversion of the interpolation matrix, $\phi(\delta)$, is not a viable approach due to the large dimensions of the matrix and the ill-posedness of the problem. That is, small errors (noise) in the estimates of the shifts will be enhanced in the estimates of the high-resolution image. Regularization techniques modify the inverse problem appropriately so that this noise amplification is controlled, while providing a meaningful solution to the original problem. Regularization has also been used in conjunction with iterative techniques for the restoration of noisy and degraded images [22, 23, 26]. Therefore, using an iterative regularized approach, Eq. (4.5) can be rewritten as

$$\mathbf{f}_{HR}^{(k+1)} = \mathbf{f}_{HR}^{(k)} + b\left(\phi^T(\delta)\mathbf{f}_{LR} - \left(\phi^T(\delta)\phi(\delta) + \lambda Q^T Q\right)\right)\mathbf{f}_{HR}^{(k)}, \quad (4.45)$$

where b is a parameter controlling the convergence and the speed of convergence, Q is the regularizing operator and λ the regularizing parameter [26]. The regularizing operator Q is typically chosen to be a Laplacian operator, usually a five point Laplacian. The choice of λ is a very important issue, and has been studied extensively [16]. A good initial choice is

$$\lambda = \frac{\text{tr}\left\{\phi^T\phi\right\}}{\text{tr}\left\{Q^T Q\right\}}, \quad (4.46)$$

where this choice weights equally both the high pass energy constraint (Q) and the noise power constraint (fidelity to the data) [36].

5. Simultaneous Restoration, Registration and Interpolation: RRI Approach

In this section, the **RRI** formulation is outlined. A major difference between the two formulations is that in **RR-I**, the output are the low-resolution, restored images, $\mathbf{f}_{LR}$, which need to be interpolated up to $\mathbf{f}_{HR}$ iteratively, while for the **RRI** formulation, the output is the final, restored, high-resolution image, $\mathbf{f}_{HR}$, obtained directly from the observed noisy images, $\mathbf{g}_{LR}$. Another major difference is that in the **RR-I** case the

shifts axe expressed as part of the covariance matrix of the low-resolution images only while in the **RRI** formulation, they are part of the interpolation operator also. The inclusion of the interpolation operator into the degradation equation still enables us to enter the frequency domain as in the **RR-I** formulation, since the interpolation matrix is semi-block circulant [46]. Due to this inclusion of the interpolation matrix into the degradation operator, the equations for the noise variances and power spectra are unchanged in form from Sec. 4, where the only difference is that the degradation matrices in Sec. 4 are replaced by two matrices - the degradation matrix and the interpolation matrix.

Again, using the same assumption of a Gauss-Markov ergodic model on Eq. (4.24), the pdf of the low-resolution observations can be written as

$$f_G(\mathbf{g}_{LR}) = \left|2\pi\left(\mathbf{H}\phi^{(k)}(\boldsymbol{\delta})\Lambda_{\boldsymbol{f}_{HR}}\phi^{T}(\boldsymbol{\delta})\boldsymbol{H}^{T} + \Lambda_{\boldsymbol{v}}\right)\right|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{g}_{LR}^{T}\left(\boldsymbol{H}\phi^{(k)}(\boldsymbol{\delta})\Lambda_{\boldsymbol{f}_{HR}}\phi^{(k)}(\boldsymbol{\delta})^{T}\boldsymbol{H}^{T} + \Lambda_{\boldsymbol{v}}\right)^{-1}\boldsymbol{g}_{LR}\right\} \tag{4.47}$$

Taking the logarithm of Eq. (4.47) and ignoring constant additive and multiplicative terms, the maximization of the log-likelihood function becomes the minimization of

$$\begin{aligned} L(\vartheta) &= \log\left|\mathbf{H}\phi^{(k)}(\boldsymbol{\delta})\Lambda_{\mathbf{f}_{HR}}\phi^{(k)^{T}}(\boldsymbol{\delta})\mathbf{H}^{T} + \Lambda_{\mathbf{v}}\right| \\ &+ \mathbf{g}_{LR}^{T}\left(\mathbf{H}\phi^{(k)}(\boldsymbol{\delta})\Lambda_{\mathbf{f}_{HR}}\phi^{(k)^{T}}(\boldsymbol{\delta})\mathbf{H}^{T} + \Lambda_{\mathbf{v}}\right)^{-1}\mathbf{g}_{LR} \end{aligned} \tag{4.48}$$

Comparing Eq. (4.48) (representing the **RRI** formulation) with Eq. (4.32) (representing the **RR-I** formulation) it is clear that the only difference is the inclusion of the interpolation operator, $\phi(\boldsymbol{\delta})$, in other words, we obtain the former equation from the latter by replacing $\mathbf{H}$ by $\mathbf{H}\phi(\boldsymbol{\delta})$. We have followed similar steps in minimizing $L(\vartheta)$ in these two equations (Eqs. (4.48) and (4.32)). That is, the EM algorithm is used in minimizing Eq. (4.48) with the same choice of complete data as in Sec. (4). Furthemore since $\phi(\boldsymbol{\delta})$ is an SBC matrix, as was described in Sec. (4.4), Eq. (4.48) is written in the discrete frequency domain, where the E-step of the EM algorithm is formulated. The minimization step, however, is now considerably more complicated than in the **RR-I** case. This stems from the fact mentioned earlier that the unknown shifts $\boldsymbol{\delta}$ appear now both in the interpolation sinc functions, the entries of $\phi(\boldsymbol{\delta})$, and the image covariance function. Differentiating therefore $L(\vartheta)$ in Eq. (4.48) with respect to $(\boldsymbol{\delta})$ is considerably more complicated. All the details of

RRI algorithm can be found in [46, 45], but are not presented here due to lack of space.

6. Experimental Results

The performance of the **RR-I** approach described in Sec. 4 is demonstrated with the following experiment. Four 128×128 low-resolution images, shown in Fig. 4.6, were generated from the 256×256 image in Fig. 4.7 (upper left) using Eq. (1.3) and the shifts shown in the second column of Table 4.1. White Gaussian noise was added to all of the frames, resulting in SNR of 30 dB. The estimated shifts by applying the EM algorithm of Eqs. (1.38-1.41) are shown in the last column of Table 4.1, along with the initial values of the shifts. These values are then used in interpolation via Eq. (1.45) to generate the image shown in Fig. 4.7 (lower left). In the same figure are also shown the bilinearly interpolated image (upper right) from the (undegraded) zero shift low-resolution image, as well as, the interpolated image using the estimated (lower left) and true values of the shifts (lower right). It is clear that the **RR-I** approach performs considerably better than bilinear interpolation. In particular, note the increased legibility of the "F-16" on the tail of the airplane, as well as the words "U.S. Air Force" on the body when comparing the upper right image to the lower left image. Further improvement in in the legibility of the words "U.S. Air Force" can be seen when comparing the lower left image to the lower right image. The peak signal to noise ratio (PSNR), defined by

$$PSNR = 10 \log \left\{ \frac{255^2}{\frac{1}{N^2}\|x - \hat{x}\|^2} \right\} , \tag{4.49}$$

where $x, \hat{x}$ denote the original and estimated images, respectively, and $N \times N$ the size of the images, was used to quantify the results. The PSNR values for the bilinearly interpolated case (of the undegraded low resolution zero shift low-resolution image) and the **RR-I** algorithm using the estimated and true shifts in the interpolation step are listed in Table 4.2. The differences in PSNRs between the bilinearly interpolated case and the **RR-I** approaches are significant, and increase with the accuracy of the estimated shifts.

It is also interesting to note that according to this experiment the difference between the high-resolution images generated by the **RR-I** approach with the estimated and true values of the shifts (lower left and lower right images, respectively) is not as large as might have been suggested by the error in the shift estimation.

In order to test the **RRI** approach, the 256×256 cameraman image was used in another experiment, and four 128×128 low-resolution images were generated again using Eq. (1.3). These low-resolution images were then blurred by a 3×3 Gaussian linear space-invariant filter with variance of 1.0. Additive noise was added resulting in a blurred SNR of 30 dB for each frame. The degraded low-resolution frames are shown in Fig. 4.8. In order to compare the **RRI** results with those from a bilinear interpolation, the zero shifted degraded low-resolution frame was restored

frame	True	Initial	Estimated
2	(0.1400,0.2100)	(0.2000,0.3000)	(0.189222,0.271433)
3	(0.3400,0.4100)	(0.5000,0.6000)	(0.357290,0.473951)
4	(0.6700,0.7900)	(0.6000,0.7000)	(0.614208,0.712440)

Table 4.1. True, initial, and estimated shifts by the RR approach.

Bilinearly Interpolated	RR-I (est shifts)	RR-I (true shifts)
28.78 dB	30.28 dB	31.17 dB

Table 4.2. PSNRs from the bilinearly interpolated and RR-I approaches.

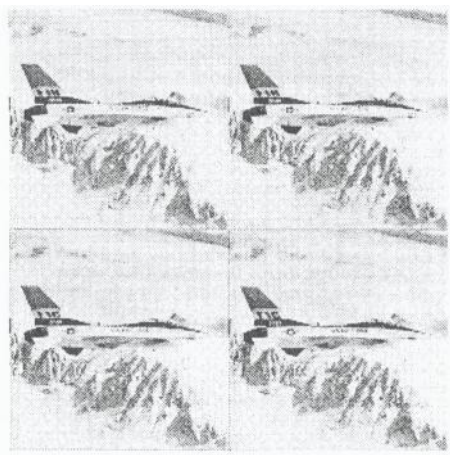


Figure 4.6. Four 128 × 128 low-resolution images

using an iterative Wiener filter with the blur PSF known. This restored low-resolution image was then bilinearly interpolated to 256 × 256. The image on the left of Fig. 4.9 shows this result, while the image on the right is from the **RRI** approach, where the blur PSF was also known, but the sub-pixel shifts and the noise variances were unknown. Note the additional detail afforded by the **RRI** approach along the legs of the tripod and the man's face. Furthermore, the background of the bilinearly interpolated image shows significant artifacts from the restoration process, while that from the **RRI** approach is much smoother. The PSNRs for these two images are 21.77 and 22.52 dB, respectively. The true, ini-

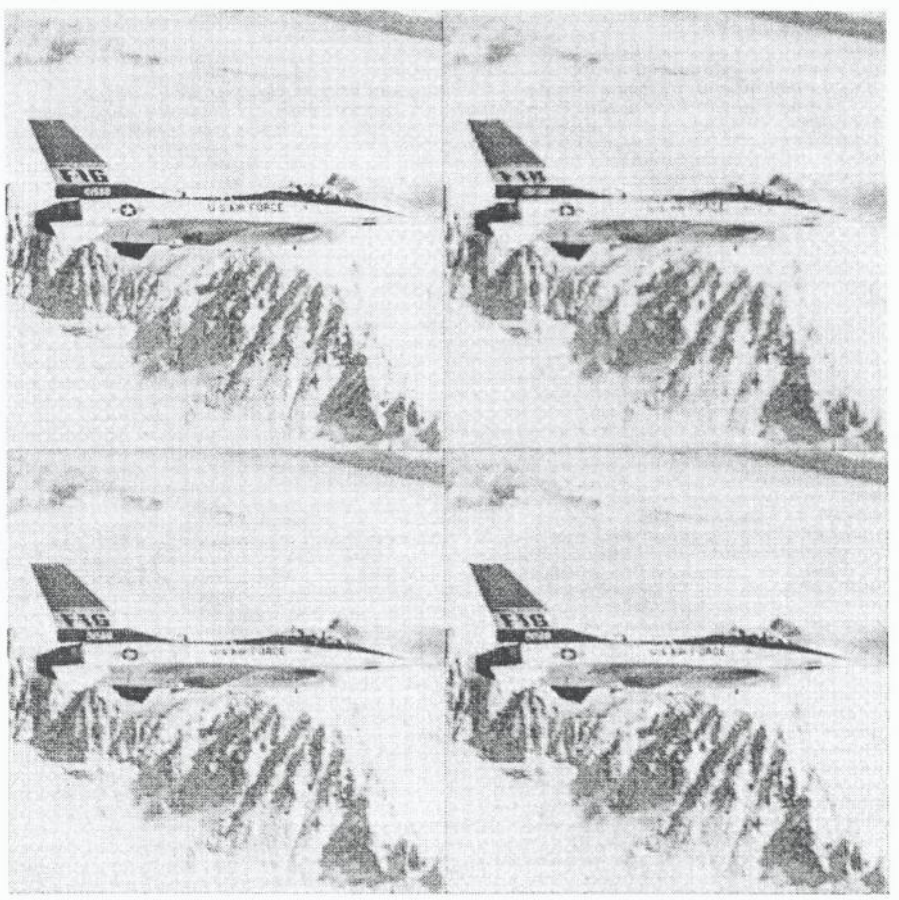


Figure 4.7. (upper left): original high-resolution image; (upper right): bilinearly interpolated image from the zero shift low-resolution image; results of the RR-I algorithm using: (lower left) estimated shifts; (lower right) true shifts

frame	True	Initial	Estimated
2	(0.1500,0.6500)	(0.3000,0.7000)	(0.2497,0.4531)
3	(0.6000,0.4000)	(0.7000,0.4500)	(0.5218,0.4500)
4	(0.2500,0.7500)	(0.4000,0.7000)	(0.2727,0.5092)

Table 4.3. True, initial, and estimated shifts by the **RRI** approach.

tial, and estimated shifts are shown in Table 4.3. Most of the estimated shifts were reasonably close to the true values, with the exception of δ_y^2 and δ_y^4.

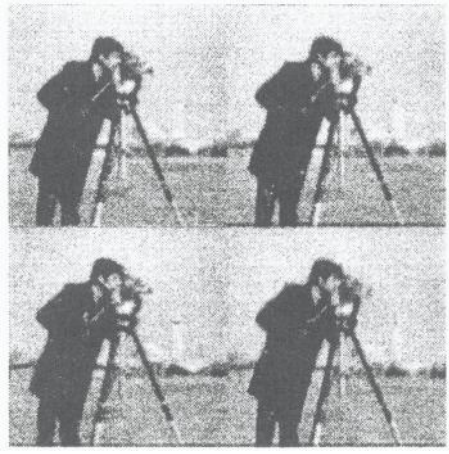


Figure 4.8. Degraded low-resolution images.

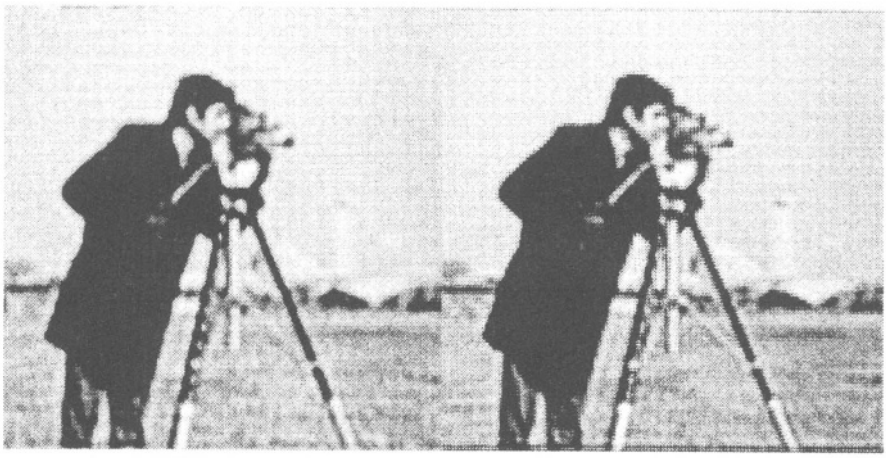


Figure 4.9. Bilinearly interpolated image from restored low-resolution image and reconstructed image from **RRI** approach using estimated shifts.

7. Conclusions and Future Work

In this chapter the problem of reconstructing a high-resolution image from multiple degraded low-resolution images is addressed. We first formulated the problem in the spatial and frequency domains and reviewed the relevant literature. We then presented a unified framework based on maximum likelihood estimation and the EM algorithm that simultaneously deals with all the subtasks involved; namely that of registration, restoration, and interpolation. Although the framework is presented in the discrete spatial domain, the problem is transformed and solved in the discrete frequency domain by appropriately exploiting the structures of the matrices appearing in the likelihood function. This framework is elegant and quite straightforward to understand. However, the resulting M-step of the EM algorithm involves a complex and non convex minimization. This makes the resulting estimates of the subpixel shifts dependent on the initial conditions used for the optimization. Current work involves the exploration of the frequency domain relationship between the high and the low-resolution images and the formulation of the problem in a way which is better suited for non convex problems, such as Graduated Non Convexity [34].

Acknowledgments

The authors would like to acknowledge the assistance of Carlos Luna and Passant Karunaratne, at Northwestern University, during the preparation of the manuscript.

References

[1] K. Aizawa, T. Komatsu, and T. Saito, "A Scheme for Acquiring Very High Resolution Images Using Multiple Cameras," *IEEE Proc. ICASSP-92*, San Francisco, CA, vol. III, pp. 289-292, 1992.

[2] K. Aizawa, T. Komatsu, T. Saito, and M. Hatori, "Subpixel Registration for a High Resolution Imaging System Using Multiple Imagers," *IEEE Proc. ICASSP-93*, Minneapolis, MN, vol. V, pp. 133-136, 1993.

[3] P. E. Anuta, "Spatial Registration of Multispectral and Multitemporal Digital Imagery Using Fast Fourier Transform Techniques," *IEEE Trans. Geoscience Electronics*, vol. GE-8, no. 4, pp. 353-368, Oct. 1970.

[4] M. R. Banham and A. K. Katsaggelos, "Digital Image Restoration," *IEEE Signal Processing Magazine*, vol. 14, no. 2, pp. 24-41, March 1997.

[5] D. I. Barnea and H. F. Silverman, "A Class of Algorithms for Fast Digital Image Registration," *IEEE Trans. Computers*, vol. C-21, no. 2, pp. 179-186, 1972.

[6] N. K. Bose, H. C. Kim, H. M. Valenzuela, "Recursive Implementation of Total Least Squares Algorithm for Image Reconstruction from Noisy, Undersampled Multiframes," *IEEE Proc. ICASSP-93*, Minneapolis, MN, vol. V, pp. 269-272, April 1993.

[7] C. Cenker, H. G. Feichtinger and H. Steir, "Fast Iterative and Non-Iterative Reconstruction of Band-Limited Functions from Irregular Sampling Values," *IEEE Proc. ICASSP-91*, Toronto, pp. 1773-1776, 1991.

[8] A. D. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. B39, pp. 1-37, 1977.

[9] F. DeNatale, G. Desoli, D. Giusto, and G. Vernazza, "A Spline-Like Scheme for Least-Squares Bilinear Interpolation of Images," *IEEE Proc. ICASSP-93*, Minneapolis, MN, vol. V, pp. 141-144, 1993.

[10] D. Dudgeon, and R. Mersereau, *Multidimensional Digital Signal Processing*, Prentice Hall 1984.

[11] M. Elad and A. Feuer, "Restoration of a Single Superresolution Image from Several Blurred, Noisy, and Undersampled Measured Images," *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1647-1657, December 1997.

[12] H. G. Feichtinger and K. Gröchenig, "Iterative Reconstruction of Multivariate Band-Limited Functions from Irregular Sampling Values," *SIAM J. Math. Anal.*, vol. 23, no. 1, pp. 244-261, Jan. 1992.

[13] B. R. Frieden and H. H. G. Aumann, "Image Reconstruction from Multiple 1-D Scans Using Filtered Localized Projections," *Appl. Opt.*, vol. 26, pp. 3615-3621, 1987.

[14] N. P. Galatsanos and R. T. Chin, "Digital Restoration of Multichannel Images," *IEEE Trans. Acoust., Speech, Signal* Proc., vol. 37, no. 3, pp. 415-421, March 1989.

[15] N.P. Galatsanos, A.K. Katsaggelos, R.T. Chin, A. Hillery, "Least Squares Restoration of Multi-Channel Images," *IEEE Trans. Signal Processing*, vol. 39, no. 10, pp. 2222-2236, Oct. 1991.

[16] N.P. Galatsanos and A.K. Katsaggelos, "Methods for Choosing the Regularization Parameter and Estimating the Noise Variance in Image Restoration and their Relation," *IEEE Trans. Image Processing*, vol. 1, pp. 322-336, July 1992.

[17] N. P. Galatsanos, M. Wernick, and A. K. Katsaggelos, "Multichannel Image Recovery", in *Handbook of Image and Video Processing*, A. Bovik, editor, ch. 3.7, pp. 161-174, Academic Press, 2000.

[18] R. M. Gray, "On the Asymptotic Eigenevalue Distribution of Toeplitz Matrices", *IEEE Trans. on Information Theory*, vol. IT-18, pp. 725-730, November 1972.

[19] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint MAP Registration and High-Resolution Image Estimation Using a Sequence of Undersampled Images", *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1621-1633, December 1997.

[20] M. Irani and S. Peleg, "Improving Resolution by Image Registration," *CVGIP: Graphical Models and Image Proc.*, vol. 53, pp. 231-239, May 1991.

[21] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, 1988.

[22] M. G. Kang and A. K. Katsaggelos, "General Choice of the Regularization Functional in Regularized Image Restoration," *IEEE Trans. Image Proc.*, vol. 4, no. 5, pp. 594-602, May 1995.

[23] A. K. Katsaggelos, "Iterative Image Restoration Algorithms," *Optical Engineering*, vol. 28, no. 7, pp. 735-748, July 1989.

[24] A. K. Katsaggelos, ed., *Digital Image Restoration*, New York: Springer-Verlag, 1991.

[25] A. K. Katsaggelos and K. T. Lay, "Identification and Restoration of Images Using the Expectation Maximization Algorithm," in *Digital Image Restoration*, A.K. Katsaggelos, editor, Springer-Verlag, 1991.

[26] A. K. Katsaggelos, J. Biemond, R. M. Mersereau, and R. W. Schafer, "A Regularized Iterative Image Restoration Algorithm," *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp. 914-929, April 1991.

[27] A. K. Katsaggelos, K. T. Lay, and N. P. Galatsanos, "A General Framework for Frequency Domain Multi-Channel Signal Processing," *IEEE Trans. Image Proc.*, vol. 2, no. 3, pp. 417-420, July 1993.

[28] S. P. Kim and N. K. Bose, "Reconstruction of 2-D Bandlimited Discrete Signals from Nonuniform Samples," *IEE Proc.*, vol. 137, pt. F, no. 3, pp. 197-204, June 1990.

[29] S. P. Kim, N. K. Bose and H. M. Valenzuela, "Recursive Reconstruction of High Resolution Image From Noisy Undersampled Multiframes," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 38, no. 6, pp. 1013-1027, June 1990.

[30] S. P. Kim, W. Y. Su, "Recursive High-Resolution Reconstruction of Blurred Multiframe Images," *IEEE Proc. ICASSP-91*, Toronto, pp. 2977-2980, 1991.

[31] S. P. Kim, W. Su, "Subpixel Accuracy Image Registration By Spectrum Cancellation," *IEEE Proc. ICASSP-93*, Minneapolis, MN, vol. V, pp. 153-156, 1993.

[32] K. T. Lay and A. K. Katsaggelos, "Image Identification and Restoration Based on the Expectation-Maximization Algorithm," *Optical Engineering*, vol. 29, pp. 436-445, May 1990.

[33] M. S. Mort and M. D. Srinath, "Maximum Likelihood Image Registration With Subpixel Accuracy," *Proc. SPIE*, vol. 974, pp. 38-44, 1988.

[34] M. Nikolova, J. Idier, and A. Mohammad-Djafari, "Inversion of Large-support Ill-posed Linear Operators Using a Piecewise Gaussian MRF," IEEE Transactions on Image Processing, vol. 7, no. 4, pp. 571-585, 1998.

[35] A. Patti, M. Sezan and A. Tekalp, "Superresolution Video Reconstruction with Arbitrary Sampling Lattices and Non-zero Apperture Time,", *IEEE Trans. Image Processing*, vol. 6, pp. 1064-1076, August 1997.

[36] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd. ed., Cambridge University Press, 1992.

[37] R. R. Schultz, R. L. Stevenson, "Extraction of High-Resolution Stills from Video Sequences", *IEEE Trans. in Image Processing*, vol. 5, no. 6, pp. 996-1011, June 1996.

[38] C. Srinivas and M. D. Srinath, "A Stochastic Model-Based Approach for Simultaneous Restoration of Multiple Misregistered Images," *Proc. SPIE*, vol. 1360, pp. 1416-1427, 1990.

[39] K. D. Sauer and J. P. Allebach, "Iterative Reconstruction of Band-Limited Images from Nonuniformly Spaced Samples," *IEEE Trans. Circuits and Systems*, vol. 34, no. 10, pp. 1497-1505, Oct. 1987.

[40] H. Stark and P. Oskoui, "High-resolution Image Recovery from Image-Plane Arrays, Using Convex Projections," *J. Opt. Soc. Amer. A*, vol. 6, no. 11, pp. 1715-1726, Nov. 1989.

[41] A. M. Tekalp, M. K. Ozkan, and M. I. Sezan, "High-Resolution Image Reconstruction from Lower-Resolution Image Sequences and Space-Varying Image Restoration," *IEEE Proc. ICASSP 92*, San Francisco, vol. III, pp. 169-172, 1992.

[42] A. Tikhonov and V. Arsenin, *Solution of Ill-Posed Problems*, John Wiley and Sons, 1977.

[43] B. C. Tom and A. K. Katsaggelos, "Reconstruction of a High Resolution Image from Multiple Degraded Mis-Registered Low Resolution Images," *Proc. SPIE, Visual Communications and Image Processing*, Chicago, IL, vol. 2308, pt. 2, pp. 971-981, Sept. 1994.

[44] B. C. Tom, A. K. Katsaggelos, and N. P. Galatsanos, "Reconstruction of a High Resolution from Registration and Restoration of Low Resolution Images," *IEEE Proc. International Conference on Image Processing*, Austin, TX, vol. 3, pp. 553-557, Nov. 1994.

[45] B. C. Tom and A. K. Katsaggelos, "Reconstruction of a High-Resolution Image by Simultaneous Registration, Restoration, and Interpolation of Low-Resolution Images," *IEEE Proc. International Conference on Image Processing*, Washington D.C., vol. 2, pp. 539-542, Oct. 1995.

[46] B. C. Tom, "Reconstruction of a High Resolution Image from Multiple Degraded Mis-registered Low Resolution Images," *Ph.D. Thesis*, Northwestern University, Department of ECE, December 1995.

[47] B. C. Tom, K. T. Lay and A. K. Katsaggelos, "Multi-Channel Image Identification and Restoration Using the Expectation-Maximization Algorithm," *Optical Engineering*, "Special Issue on Visual Communications and Image Processing", vol. 35, no. 1, pp. 241-254 Jan. 1996.

[48] B. C. Tom and A. K. Katsaggelos, "Resolution Enhancement of Monochrome and Color Video Using Motion Compensation," *Trans Image Proc.*, vol. 10, no. 2, pp. 278-287, Feb. 2001.

[49] R. Y. Tsai and T. S. Huang, "Multiframe Image Restoration and Registration," *Advances in Computer Vision and Image Processing*, vol. 1, T. S. Huang, ed., Greenich, CT: Jai Press, ch. 7, pp. 317-339, 1984.

[50] S. Yeh and H. Stark, "Iterative and One-Step Reconstruction from Nonuniform Samples by Convex Projections," *J. Opt. Soc. Amer. A*, vol. 7, no. 3, pp. 491-499, 1990.

This page intentionally left blank

Chapter 5

SUPER-RESOLUTION IMAGING USING BLUR AS A CUE

Deepu Rajan*
School of Biomedial Engineering
Indian Institute of Technology-Bombay
Powai, Mumbai-400 076. India.
dr@doe.cusat.edu

Subhasis Chaudhuri
Department of Electrical Engineering
Indian Institute of Technology-Bombay
Powai, Mumbai-400 076. India.
sc@ee.iitb.ac.in

Abstract In this chapter, we present a parametric method for generating a super-resolution image from a sequence consisting of blurred and noisy observations. The high resolution image is modeled as a Markov random field (MRF) and a maximum a posteriori (MAP) estimation technique is used for super-resolution restoration. Unlike other super-resolution imaging methods, the proposed technique does not require sub-pixel registration of given observations. A simple gradient descent method is used to optimize the cost. The discontinuities in the intensity process can be preserved by introducing suitable line processes. Superiority of this technique to standard methods of image interpolation is illustrated. The motivation for using blur as a cue is also explained.

Keywords: Image restoration, MRF, parametric super-resolution.

*On leave from Dept. of Electronics, Cochin University of Science and Technology, Cochin - 682 022. India.

1. Introduction

Super-resolution deals with obtaining still images and video at a resolution higher than that of the sensor used in recording the image. The objective is to undo the effects of aliasing due to undersampling, loss of high frequency detail due to point spread function (PSF) averaging and due to motion or out-of-focus optical blurring. As indicated in Chapter 1, most of the techniques for super-resolution restoration reported so far involve multiple sub-pixel shifted images of a scene which implies the availability of more number of samples for image reconstruction. But the attendant pre-requisite for such an approach is that of registration of the observed frames or the estimation of the sub-pixel shifts. In some cases these shifts are assumed to be known, e.g. [1]. In this chapter, we avoid the task of registration by considering decimated, blurred and noisy versions of an ideal high resolution image which are used to generate a super-resolved image. There are no spatial shifts but the images are captured with different camera blurs.

The phenomenon of blurring is inherent during the formation of an image due to the low resolution of the point spread function of the capturing device. Blurring can also arise due to the relative motion between the camera and the scene. In the case of real aperture imaging, we know that the blur at a point is a function of the depth of the scene at that point. Hence, we see the blur as a *natural* cue in a low resolution image and hence, it should be exploited. Thus, the motivation behind using blur as a cue is the fact that it is already present in the low resolution observations. When different images of a scene are captured using different camera parameter (focus, aperture, etc.) settings, the relative blur between the observations are known. Here, in this work, we assume that the blurs are known. Of course, in the practical case where the blurs are unknown, there are techniques by which they can be estimated [2]. Assuming the blurs to be known and that the high resolution image can be represented by a Markov random field, we perform a super-resolved restoration of the observations to obtain the high resolution image.

Contextual constraints are essential in the interpretation of visual information. The theory of Markov random fields (MRF) provides a convenient and consistent way of modeling context dependent entities. The practical use of MRF models is largely ascribed to the equivalence between MRFs and Gibbs distribution established by Hammersely and Clifford and further developed by Besag [3]. This equivalence enables the modeling of vision problems by a mathematically sound and tractable means for image analysis in the Bayesian framework. From the com-

putational perspective, the local property of MRFs leads to algorithms which can be implemented in a local and massively parallel manner.

In this chapter we model the super-resolved image to be estimated as an MRF. Together with statistical decision and estimation theories, the MRF model enables the formulation of objective functions in terms of established optimality criteria. Maximum *a posteriori* probability (MAP) is one of the most popular statistical criteria for optimality. In the MAP-MRF framework, the objective is the joint posterior probability of the MRF labels. The prior model for the super-resolved image is chosen in such a way that the resultant cost function is convex. Consequently a fast optimization technique such as gradient descent minimization suffices. One can obtain an improved result for super-resolution if appropriate line fields are included in the cost function to preserve the discontinuities in the intensity process. [4]. However, in this case the computational requirement goes up. In situations where the computational cost is not an issue, simulated annealing can be used for optimization. A relatively faster optimization technique would be the graduated non-convexity algorithm[5].

In the following section, we briefly review the theory of MRFs. Section 3 describes the low resolution image formation in terms of the unkown high resolution image. In Section 4, the cost function using the MAP estimator is derived. Section 5 presents experimental results and conclusions are presented in Section 6.

2. Theory of MRF

Since the theory of MRF has been used for super-resolution restoration of observations, a quick review of the MRF is given in this section for completeness of discussion.

A lattice S is a square array of pixels $\{(i,j)|1 \leq i,j \leq N\}$. A random field is the triplet $< \Omega, \Psi, P >$ where Ω is the sample space, Ψ is the class of Borel sets on Ω and P is a probability measure with domain Ψ. A random field model is a distribution for an M-dimensional random vector $\mathbf{X}$, which contains a random variable X_t for the 'label' at site t. Label could be gray values, pattern classes, etc.

The sites in S are related to each other through a neighborhood system defined as

$$N = \{N_i, \quad \forall i \in S\}$$

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

1 a site is not neighboring itself : $i \notin N_i$

2 the neighboring relationship is mutual : $i \in N_{i'} \Leftrightarrow i' \in N_i$

The pair (S, N) constitutes a graph where S contains the nodes and N determines the arcs between the nodes. A *clique* c for (S, N) is defined as a subset of sites in S in which all pairs of sites are mutual neighbors. Cliques can occur as singletons, doublets, triplets and so on. The first

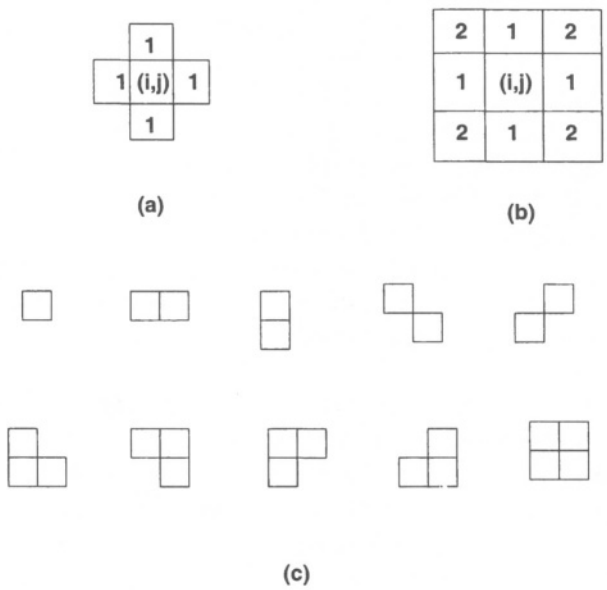


Figure 5.1. Illustration of (a) first and (b) second order neighborhood and (c) associated cliques

and second order neighborhoods of a site and their corresponding cliques are shown in Figure 5.1.

A discrete Gibbs random field (GRF) provides a global model [6] for an image by specifying the following probability mass function :

$$P(\mathbf{X} = \mathbf{x}) = e^{-U(\mathbf{x})}/Z \tag{5.1}$$

where $U(\mathbf{x})$ is called the *energy* function, $\mathbf{x}$ is a vector of 'labels' and Z is a normalizing constant called the *partition* function, given by $Z = \sum_{\mathbf{x}} e^{-U(\mathbf{x})}$. The joint distribution indicates that smaller the energy of a particular realization $\mathbf{x}$, the more likely it is to occur. A potential function $V_c(\mathbf{x})$ is associated with each clique and the energy function can be expressed as

$$U(\mathbf{x}) = \sum_{c \in \mathcal{C}} V_c(\mathbf{x})$$

where $\mathcal{C}$ is the set of all cliques in a neighborhood system. A GRF is parameterised by the choice of clique functions, e.g., [7]

$$\begin{aligned} V_c(\mathbf{x}) &= \zeta_c, \text{ if all sites in clique } c \text{ have the same label} \\ &= -\zeta_c, \text{ otherwise.} \end{aligned} \tag{5.2}$$

While a GRF describes the global properties of an image in terms of the joint distribution of labels for all pixels, an MRF is defined in terms of local properties. A random field, with respect to a neighborhood system, is a discrete MRF if its probability mass function satisfies the following properties [8, 9]: (Notation : $S\backslash t$ is the set of all sites in S excluding site t. ∂t is the set of all sites in the neighborhood of site t, excluding site t itself.)

1 (Positivity) $P(\mathbf{X} = \mathbf{x}) > 0, \forall \mathbf{x} \in \Omega$

2 (Markov Property) $P(X_t = x_t | \mathbf{X}_{S\backslash t} = \mathbf{x}_{S\backslash t}) = P(X_t = x_t | \mathbf{X}_{\partial t} = \mathbf{x}_{\partial t})$

3 (Homogeneity) $P(X_t = x_t | \mathbf{X}_{\partial t} = \mathbf{x}_{\partial t})$ is the same for all sites t.

While an MRF is characterized by its local property, viz., the Markovianity, a GRF is characterized by its global property. The utility of MRFs arises from the Hammersley-Clifford theorem which states that $\mathbf{X}$ is *an MRF on S with respect to N if and only if* $\mathbf{X}$ *is a GRF on S with respect to N* [3, 6]. The theorem provides a simple way of expressing the joint probability $P(\mathbf{X} = \mathbf{x})$ by specifying the clique potential functions $V_c(\mathbf{x})$. This enables the a *priori* knowledge to be encoded into the estimation process.

A variety of physical phenomena is characterized by *smoothness*. It is one of the most common assumptions in computer vision models, especially those formulated as MRFs [4, 10]. The line process model, introduced by Geman and Geman [4] assumes piecewise smoothness whereby the smoothness constraint is "switched off" at points where the magnitude of the signal derivative exceeds certain threshold. It is defined on a dual lattice that has two sites corresponding to the vertical and horizontal line fields whose elements are $v(i, j)$ and $l(i, j)$, respectively, that take on binary values from {0,1} resulting in corresponding binary line fields $\mathbf{V}$ and $\mathbf{L}$. The *a priori* Gibbs distribution in (5.1) can be modified to

$$P(\mathbf{X} = \mathbf{x}, \mathbf{L} = \mathbf{l}, \mathbf{V} = \mathbf{v}) = e^{-U(\mathbf{x},\mathbf{l},\mathbf{v})}/Z$$

where the partition function $Z = \sum_{\mathbf{x},\mathbf{l},\mathbf{v}} e^{-U(\mathbf{x},\mathbf{l},\mathbf{v})}$. The on-state of the line-process variable indicates that a discontinuity, in the form of a high gradient, is detected between neighboring points, e.g., $l(i, j) =$

1 if $|x(i,j) - x(i-1,j)|$ >Threshold, else $l(i,j) = 0$. Each turn-on of a line-process variable is penalized by a quantity γ so as to prevent spurious discontinuities. Thus, for the so-called *weak membrane* model [5], the energy function is written as

$$\begin{aligned} U(\mathbf{x},\mathbf{l},\mathbf{v}) &= \sum_{c\in\mathcal{C}} V_c(\mathbf{x},\mathbf{l},\mathbf{v}) \\ &= \sum_{i,j}[(x(i,j) - x(i,j-1))^2(1 - v(i,j)) \\ &\quad +(x(i,j+1) - x(i,j))^2(1 - v(i,j+1)) \\ &\quad +(x(i,j) - x(i-1,j))^2(1 - l(i,j)) \\ &\quad +(x(i+1,j) - x(i,j))^2(1 - l(i+1,j))] \\ &\quad +\gamma[l(i,j) + l(i+1,j) + v(i,j) + v(i,j+1)] \quad (5.3) \end{aligned}$$

It may be noted that the energy function defined as above is not differentiable and hence minimization of the energy function is computationally demanding. One may, however, define a smoothly varying line process $l(i,j)$ and $v(i,j)$ such that the energy function is differentiable.

3. Modeling the Low Resolution Observations

The low resolution image sensor plane can be perceived as a collection of $M_1 \times M_2$ square sensor elements. The low resolution intensity values are denoted as $\mathcal{Y} = \{y(i,j)\}$, $i = 0,\ldots,M_1-1$ and $j = 0,\ldots,M_2-1$. If the downsampling parameters are q_1 and q_2 in the horizontal and vertical directions, respectively, then the high resolution image will be of size $q_1M_1 \times q_2M_2$. Without loss of generality, we can assume $q_1 = q_2 = q$, and therefore the desired high-resolution image $\mathcal{Z}$ will have intensity values $\{z(k,l)\}$, $k = 0,\ldots,qM_1 - 1$ and $l = 0,\ldots,qM_2 - 1$. Given $\{z(k,l)\}$, the process of obtaining $\{y(i,j)\}$ is written as

$$y(i,j) = \frac{1}{q^2}\sum_{k=qi}^{(q+1)i-1}\sum_{l=qj}^{(q+1)j-1} z(k,l) \quad (5.4)$$

i.e., the low resolution intensity is the average of the high resolution intensities over a neighborhood of q^2 pixels. This decimation model simulates the integration of light intensity that falls on the high-resolution detector.

Each of the decimated images is blurred by a different, but known linear space invariant blurring kernel. Motivation for having such a blur is already given in Section 1. Elad and Feuer [11, 12] have shown that in this case super-resolution restoration is possible even if there is no

relative motion between the input images. They derive the following necessary condition for super-resolution to be possible for images that are not represented parametrically :

$$q^2 \leq \min\{[2m+1]^2 - 2, p\} \tag{5.5}$$

where $(2m + 1) \times (2m + 1)$ is the size of the blurring kernel and p is the number of input images. In the current study, since the high resolution image has been modeled by an MRF and since the model parameters (in terms of clique potential) are assumed to be known, one can recover the high resolution image with much fewer observations. Hence, although more number of blurred observations of a scene do not provide any additional information in the same sense as sub-pixel shifts of the camera or changing illuminant directions do, it is, nevertheless, possible to achieve super-resolution with these blurred samples, provided equation (5.5) is satisfied. Even if only the relations among the blurring functions are known, as is the case in, say, depth from defocus problems [13, 2], it is tantamount to knowing all the blurs provided any one of them is known. Finally, i.i.d. zero mean Gaussian noise is added to the decimated and blurred images. Noise is assumed to be uncorrelated in different low resolution images.

Next, we formally state the problem by casting it in a restoration framework. There are p observed images $\{\mathcal{Y}_m\}_{m=1}^{p}$ each of size $M_1 \times M_2$ which are the decimated, blurred and noisy versions of a single high resolution image $\mathcal{Z}$ of size $N_1 \times N_2$, where $N_1 = qM_1$ and $N_2 = qM_2$. If $\underline{\mathbf{y}}_m$ is the $M_1M_2 \times 1$ lexicographically ordered vector containing pixels from the low resolution image $\mathcal{Y}_m$, then a vector $\underline{\mathbf{z}}$ of size $q^2M_1M_2 \times 1$ containing pixels of the high resolution image can be formed by placing each of the $q \times q$ pixel neighborhoods sequentially so as to maintain the relationship between a low resolution pixel and its corresponding high resolution pixel. After incorporating the blur matrix and the noise vector, the image formation model is written as

$$\underline{\mathbf{y}}_m = H_m D \underline{\mathbf{z}} + \underline{\mathbf{n}}_m, \quad m = 1, \ldots, p \tag{5.6}$$

where D is the decimation matrix of size $M_1M_2 \times q^2M_1M_2$, H is the blurring matrix (PSF) of size $M_1M_2 \times M_1M_2$, $\underline{\mathbf{n}}_m$ is the $M_1M_2 \times 1$ noise vector and p is the number of low resolution observations. In the current study, we assume the blur kernel to be shift invariant so that the matrix H is block-Toeplitz. In a practical case, where the cue comes from the natural blur due to defocus [2], the blur will be shift varying and hence the matrix H will not be block-Toeplitz. The decimation

matrix D consists of q^2 values of $\frac{1}{q^2}$ in each row and has the form [14]

$$D = \frac{1}{q^2}\begin{bmatrix} 1\,1\ldots1 & & & 0 \\ & 1\,1\ldots1 & & \\ & & \ddots & \\ 0 & & & 1\,1\ldots1 \end{bmatrix} \tag{5.7}$$

It may be noted that $\underline{\mathbf{z}}$, representing the high resolution intensity process is not lexicographically ordered unlike $\underline{\mathbf{y}}_m$. For a lexicographically ordered data $\underline{\mathbf{z}}$, the matrix D will have a different structure.

Thus, the model indicates a collection of low resolution images, each of which differs from the others in the blur matrix, which is akin to changing the focus of a stationary camera looking at a stationary scene. However, as noted earlier, here we assume that the blurs are known. Since we have assumed noise to be zero mean i.i.d, the multivariate probability density function of $\underline{\mathbf{n}}_m$ is given by

$$P(\underline{\mathbf{n}}_m) = \frac{1}{(2\pi)^{\frac{M_1 M_2}{2}} \sigma_\eta^{M_1 M_2}} \exp\left\{-\frac{1}{2\sigma_\eta^2} \underline{\mathbf{n}}_m{}^T \underline{\mathbf{n}}_m\right\}, \tag{5.8}$$

where σ_η^2 denotes the variance of the noise process. Our problem now reduces to estimating $\underline{\mathbf{z}}$ given $\underline{\mathbf{y}}_m$'s, which is clearly an ill-posed, inverse problem.

Although the process of decimation followed by blurring is the reverse of the more intuitive process of decimation of a blurred image, we have observed from simulations that the results are very similar in both the cases. Moreover, the computational overhead in the model of equation (5.6) will increase, if the positions of H_m and D are swapped. Mathematically, both the models (i.e., positions of H_m and D swapped) are equivalent as H_m and D are both known.

4. MAP Estimation of the Super-resolution Image

The maximum *a posteriori* (MAP) estimation technique is used to obtain the high resolution image $\underline{\mathbf{z}}$ given the ensemble of low resolution images, i.e.,

$$\hat{\underline{\mathbf{z}}} = \arg\max_{\underline{\mathbf{z}}} P(\underline{\mathbf{z}}|\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_2, \ldots, \underline{\mathbf{y}}_p) \tag{5.9}$$

From Bayes' rule, this can be written as

$$\hat{\underline{\mathbf{z}}} = \arg\max_{\underline{\mathbf{z}}} \frac{P(\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_2, \ldots, \underline{\mathbf{y}}_p|\underline{\mathbf{z}}) P(\underline{\mathbf{z}})}{P(\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_2, \ldots, \underline{\mathbf{y}}_p)}. \tag{5.10}$$

Since the denominator is not a function of $\hat{\underline{\mathbf{z}}}$, equation (5.10) can be written as

$$\hat{\underline{\mathbf{z}}} = \arg\max_{\underline{\mathbf{z}}} P(\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_2, \ldots, \underline{\mathbf{y}}_p|\underline{\mathbf{z}})P(\underline{\mathbf{z}}). \tag{5.11}$$

Taking the log of posterior probability,

$$\hat{\underline{\mathbf{z}}} = \arg\max_{\underline{\mathbf{z}}}[\log P(\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_2, \ldots, \underline{\mathbf{y}}_p|\underline{\mathbf{z}}) + \log P(\underline{\mathbf{z}})]. \tag{5.12}$$

Hence, we need to specify the prior image density $\mathrm{P}(\underline{\mathbf{z}})$ and the conditional density $P(\underline{\mathbf{y}}_1, \underline{\mathbf{y}}_2, \ldots, \underline{\mathbf{y}}_p|\underline{\mathbf{z}})$.

4.1. Prior Model for the Super-resolution Image

MRF models have been widely used to solve computer vision problems because of their ability to model context dependency, since interpretation of visual information necessitates an efficient description of contextual constraints. As mentioned earlier, the utility of MRF models arises from the Hammersley-Clifford theorem which describes the equivalence of the local property that characterizes an MRF and the global property which characterizes a Gibbs random field (GRF). The lexicographically ordered high resolution image $\underline{\mathbf{z}}$ satisfying the Gibbs density function is now written as

$$P(\underline{\mathbf{z}}) = \frac{1}{Z}\exp\{-\sum_{c\in\mathcal{C}} V_c(\underline{\mathbf{z}})\} \tag{5.13}$$

where Z is a normalizing constant known as the partition function, $V_c(.)$ is the clique potential and $\mathcal{C}$ is the set of all cliques in the image. In order to employ a simple and fast minimization technique like gradient descent, it is desirable to have a convex energy function. More importantly, the minimization procedure should not get trapped in a local minima. To this end, we consider pair wise cliques on a first order neighborhood and impose a quadratic cost which is a function of finite difference approximations of the first order derivative at each pixel location, i.e.,

$$V_c(\underline{\mathbf{z}}) = \frac{1}{\lambda}\sum_{k=1}^{N_1}\sum_{l=1}^{N_2}[(z(k,l) - z(k,l-1))^2 + (z(k,l) - z(k-1,l))^2] \tag{5.14}$$

where λ can be viewed as a "tuning" parameter. It can be interpreted as the penalty for departure from smoothness in $\underline{\mathbf{z}}$. In this study, we make no attempt at finding out the parameters that constitute the MRF model for a given scene. Indeed, if one has access to the correct model and the corresponding parameters, one is likely to perform much better restoration. Here, we demonstrate that the method performs well even while using a simple model to describe the intensity process.

It is well known that in images, points having significant change in the image irradiance carry important information. In order to incorporate provisions for detecting such discontinuities, Geman and Geman [4] introduced the concept of line fields located on a dual lattice. We describe a prior using horizontal and vertical line fields in Section 4.4 and use graduated non-convexity (GNC) algorithm to optimize the corresponding cost function. As mentioned earlier, where computational issues do not arise, one could go in for simulated annealing (SA) in which case, we observe a significant improvement in the performance of the restoration process.

4.2. MAP Solution

From equation (5.12), since the noise process $\underline{\mathbf{n}}_m$'s are independent,

$$\begin{aligned}\hat{\underline{\mathbf{z}}} &= \arg\max_{\underline{\mathbf{z}}} \left[\log \prod_{m=1}^{p} P(\underline{\mathbf{y}}_m|\underline{\mathbf{z}}) + \log P(\underline{\mathbf{z}}) \right] \\ &= \arg\max_{\underline{\mathbf{z}}} \left[\sum_{m=1}^{p} \log P(\underline{\mathbf{y}}_m|\underline{\mathbf{z}}) + \log P(\underline{\mathbf{z}}) \right]. \end{aligned} \tag{5.15}$$

Since noise is assumed to be i.i.d Gaussian, from equations (5.6) and (5.8) we obtain

$$\begin{aligned} P(\underline{\mathbf{y}}_m|\underline{\mathbf{z}}) &= \left[\sum_{m=1}^{p} \log \frac{1}{(2\pi\sigma_\eta^2)^{\frac{M_1 M_2}{2}}} \exp\left\{ -\frac{\| \underline{\mathbf{y}}_m - H_m D\underline{\mathbf{z}} \|^2}{2\sigma_\eta^2} \right\} \right] \\ &= -\sum_{m=1}^{p} \frac{\| \underline{\mathbf{y}}_m - H_m D\underline{\mathbf{z}} \|^2}{2\sigma_\eta^2} - \frac{M_1 M_2}{2} \log(2\pi\sigma_\eta^2), \end{aligned} \tag{5.16}$$

where σ_η is the noise variance. Substituting in (5.15) and using (5.13) we get,

$$\begin{aligned}\hat{\underline{\mathbf{z}}} &= \arg\max_{\underline{\mathbf{z}}} \left[\sum_{m=1}^{p} -\frac{\| \underline{\mathbf{y}}_m - H_m D\underline{\mathbf{z}} \|^2}{2\sigma_\eta^2} - \sum_{c\in C} V_c(\underline{\mathbf{z}}) \right] \\ &= \arg\min_{\underline{\mathbf{z}}} \left[\sum_{m=1}^{p} \frac{\| \underline{\mathbf{y}}_m - H_m D\underline{\mathbf{z}} \|^2}{2\sigma_\eta^2} + \sum_{c\in C} V_c(\underline{\mathbf{z}}) \right] \end{aligned} \tag{5.17}$$

Substituting equation (5.14) into equation (5.17), the final cost function is obtained as

$$\hat{\underline{\mathbf{z}}} = \arg\min_{\underline{\mathbf{z}}} \left[\sum_{m=1}^{p} \frac{\| \underline{\mathbf{y}}_i - H_m D\underline{\mathbf{z}} \|^2}{2\sigma_\eta^2} + \frac{1}{\lambda} \sum_{k=1}^{N_1} \sum_{l=1}^{N_2} [(z(k,l) - z(k,l-1))^2 \right.$$

$$+ \quad (z(k,l) - z(k-1,l))^2] \right] \quad (5.18)$$

The above cost function is convex in terms of the unknown image $\underline{\mathbf{z}}$ and hence a simple gradient descent optimization can be used to minimize it. It may be mentioned here that although the super-resolved image $\underline{\mathbf{z}}$ has been assumed to be an MRF, the low resolution observations $\underline{\mathbf{y}}_m$ do not constitute separate MRFs, and hence a multi-resolution MRF model based super-resolution scheme will not work [15].

4.3. Gradient Descent Optimization

The cost function of equation (5.18) consists of two parts - the first part is the error between the observation model and the observed data and the second part is the regularization term which is minimized when $\underline{\mathbf{z}}$ is smooth. It is not sufficient to minimize the error term alone since this will lead to excessive noise amplification due to the ill-posed nature of the inverse problem. The contribution of the two terms are controlled by the noise variance σ_η^2 and the regularization parameter λ. The gradient of (5.18), at the n^{th} iteration is given by

$$g^{(n)} = \frac{1}{\sigma_\eta^2} \sum_{m=1}^{p} D^T H_i^T (H_m D \underline{\mathbf{z}}^{(n)} - \underline{\mathbf{y}}_m) + \frac{G^{(n)}}{\lambda} \quad (5.19)$$

where $G^{(n)}$ at location (k,l) in the super-resolution lattice is given by

$$\begin{aligned} G^{(n)}(k,l) \quad = \quad & 2[4z^{(n)}(k,l) - z^{(n)}(k,l-1) - z^{(n)}(k,l+1) \\ & -z^{(n)}(k-1,l) - z^{(n)}(k+1,l)]. \end{aligned}$$

The estimate at $(n+1)^{th}$ iteration,

$$\underline{\mathbf{z}}^{(n+1)} = \underline{\mathbf{z}}^{(n)} - \alpha g^{(n)},$$

where α is the step size, is computed iteratively until $\| \underline{\mathbf{z}}^{(n+1)} - \underline{\mathbf{z}}^{(n)} \| <$ Threshold. The initial estimate $\underline{\mathbf{z}}^{(0)}$ is chosen as the bilinear interpolation of the available least blurred, low resolution image. It should be noted here again that the necessary condition for obtaining the super-resolution image given in equation (5.5) is not applicable here as the super-resolved image $\underline{\mathbf{z}}$ is modeled by an MRF unlike in [11] where $\underline{\mathbf{z}}$ is not represented parametrically. It is the parameteric representation of the super-resolved image $\underline{\mathbf{z}}$ (in terms of the MRF model) along with the blur cue that helps us in obtaining the super-resolution.

4.4. Preservation of Discontinuities

Presence or absence of discontinuities conveys important information such as change in surface orientation, depth, texture, etc. The concept of

line fields on a dual lattice, consisting of sites corresponding to vertical and horizontal line fields, was introduced in [4]. The horizontal line field element $l(i,j)$ connecting site (i,j) to $(i,j-1)$ aids in detecting a horizontal edge, while the vertical line field element $v(i,j)$ connecting site (i,j) to $(i-1,j)$ helps in detecting a vertical edge. Note that we have chosen $l(i,j)$ and $v(i,j)$ to be binary variables in this study. However, one can use continuous variables as well without much changing the problem formulation [9]. The advantage of using continuous variable line fields lies in having a differentiable cost function when a gradient-based optimization method can still be used. The log of the prior distribution in equation (5.13), observing the normalizing term in other parameters, becomes

$$\sum_{c\in C} V_c(\underline{\mathbf{z}}) = \sum_{k,l}[\mu e_{zs} + \gamma e_{zp}] = V(\underline{\mathbf{z}}), \quad (say), \tag{5.20}$$

where

$$\begin{aligned} e_{zs} &= (z(k,l) - z(k,l-1))^2(1 - v(k,l)) \\ &+ (z(k,l+1) - z(k,l))^2(1 - v(k,l+1)) \\ &+ (z(k,l) - z(k-1,l))^2(1 - l(k,l)) \\ &+ (z(k+1,l) - z(k,l))^2(1 - l(k+1,l)), \\ \text{and } e_{zp} &= l(k,l) + l(k+1,l) + v(k,l) + v(k,l+1) \end{aligned}$$

Here e_{zs} is the same smoothness term but punctuated through the incorporation of the appropriate line fields, and the term e_{zp} quantifies the amount of punctuation in the otherwise smooth field. Given a preset threshold, if the gradient at a particular location is above that threshold, the corresponding line field is activated to indicate the presence of a discontinuity. The term multiplying γ provides a penalty for every discontinuity so created. Putting the above expression into equation (5.17), we arrive at the modified cost function

$$\hat{\mathbf{z}} = \arg\min_{\underline{\mathbf{z}}} \left[\sum_{m=1}^{p} \frac{\| \underline{\mathbf{y}}_m - H_m D \underline{\mathbf{z}} \|^2}{2\sigma_\eta^2} + V(\underline{\mathbf{z}}) \right]. \tag{5.21}$$

When the energy function is non-convex, there is a possibility of the steepest descent type of algorithms getting trapped in a local minima. As shown earlier, our cost function was chosen to be convex. This saved us from the requirement of using a computationally intensive minimization technique like simulated annealing (SA). However, on inclusion of line field terms in the cost function to account for discontinuities in the image, the gradient descent technique is liable to get trapped in local minima. We see the similarity between the above cost function and the

energy function of the weak membrane formulation [5]. Hence, the GNC algorithm is apt for carrying out the minimization. Although the results indicate an improvement over the gradient descent approach, still better estimates of the super-resolved image $\underline{\mathbf{z}}$ are observed using SA since it guarantees the attainment of the global minima as opposed to the GNC algorithm which is a sub-optimal one.

5. Experimental Results

In this section, we present results of simulation of the technique on various images. Figure 5.2 shows two of the five low resolution noisy images of Lena, CT and Pentagon, each of size 64 × 64, obtained by decimating the respective original images and blurring the decimated images with Gaussian blurs, repeated here that the reverse process of blurring the original image and decimating it, does not produce significant difference from the results reported here. Although the Gaussian blur has an infinite extent, for purpose of computation we chose the kernel size according to an extent of $\pm 3\sigma$, where σ is the blur parameter. Each low resolution observation contains zero mean Gaussian noise with variance 5.0.

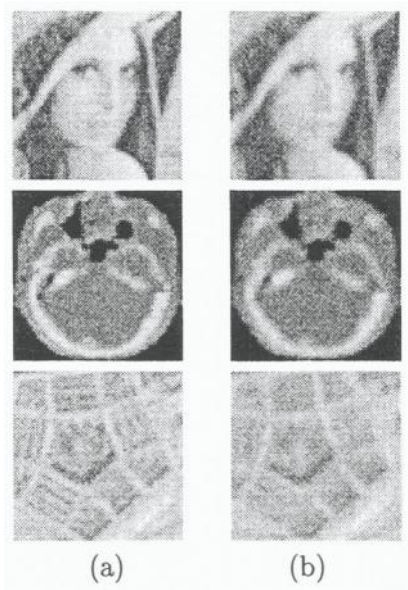


Figure 5.2. Low resolution, noisy images of Lena, CT and Pentagon with blurs (a) $\sigma = 0.7$ and (b) $\sigma = 1.1$, respectively.

First, we present the results of super-resolution using the gradient descent method. As mentioned in Section 4.3, the initial estimate of the

high resolution image is the bilinear interpolation of the least blurred observation. The smoothness parameter λ was chosen as 16.75 for the Lena and CT images and 20.0 for the Pentagon image. These parameters were chosen mostly on an adhoc basis and no effort has been made in this study to arrive at their optimal values. The step size was initially chosen as 0.1 and was reduced by a factor of 0.99 after each iteration. For large values of λ, the data consistency term in equation (5.18) dominates, producing excessive blockiness in the expanded image. On the other hand, a small value of λ causes over-smoothing. The super-resolved Lena image using the gradient descent optimization scheme is shown in Figure 5.3. Results of zero order hold expansion and cubic spline

Figure 5.3. Super-resolved Lena image obtained using gradient-descent optimization.

interpolation of the least blurred Lena image are shown in Figures 5.4(a) and 5.4(b), respectively. The blockiness in the zero-order hold expanded

Figure 5.4. Lena image (a) zero order hold expanded and (b) cubic spline interpolated, respectively.

image is clearly discernible, while the spline interpolated image is not

only significantly noisy, but, as is expected of splines, it is also blurred due to over-smoothing. On the other hand, the proposed algorithm also does deblurring in addition to removing noise and generating a super-resolved image. The super-resolved CT and Pentagon images with 5 low resolution observations using the proposed method are shown in Figure 5.5. The gradient descent method was used for the optimization purpose. The cynuses in the bone near the right middle edge of the CT image which are not visible in the low resolution observations show up clearly in the super-resolved image. The super-resolved Pentagon image contains more details of the circular central part than any of the low-resolution images.

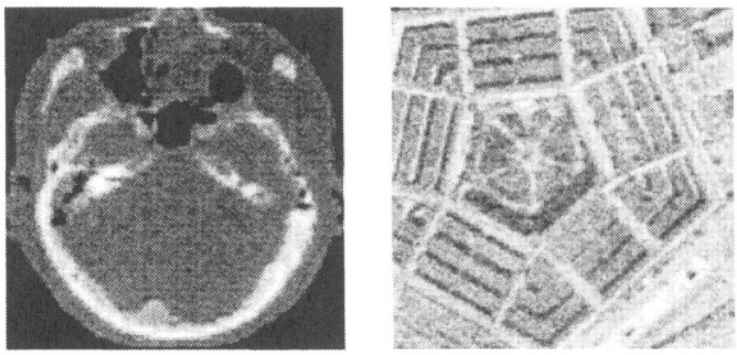


Figure 5.5. Super-resolved images of (a) CT and (b) Pentagon with 5 low resolution observations, using the gradient-descent method.

The mean squared error between the original image and generated super resolved image is defined as

$$MSE = \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}(\hat{z}_{i,j} - z_{i,j})^2}{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}(z_{i,j})^2}. \tag{5.22}$$

Table 5.1 shows the comparison of the MSE for the proposed method with standard methods like zero-order hold and cubic spline interpolation. Notice the significant drop in the MSEs for CT and Pentagon images in going from cubic spline interpolation to the proposed technique using gradient descent optimization technique.

In another experiment, only two low resolution observations, each of Lena, CT and Pentagon were constructed, out of which one was not blurred and the other was blurred with $\sigma = 0.5$. The mean squared errors of the super-resolved Lena, CT and Pentagon images shown in Figure 5.6 were 0.003586, 0.021446 and 0.009416, respectively. Compare

Table 5.1. Comparison of MSEs for different interpolation schemes.

Method	Lena	CT	Pentagon
ZOH	0.012061	0.899187	0.054042
Cubic spline	0.011870	0.452966	0.052309
Gradient Descent	0.003531	0.021216	0.010676
GNC	0.002255	0.021085	0.006962
SA	0.002143	0.007615	0.004186

this to the results given in the third row in table 5.1. This is not very different from the results we obtained when all the input images were defocused. Hence, there is no appreciable gain in having focused images in the low resolution ensemble. The proposed technique is, therefore, suitable for low-resolution images that are blurred, since the algorithm inherently performs a deblurring operation. It was observed that there was a marked progressive reduction in the mean square errors till four input observations; however, for five or more images, the errors did not decrease significantly, although more number of images helps in smoothing out noise.

Next, we present results of minimization of the modified cost function when line processes are used to preserve discontinuity. As before, we consider 5 low resolution observations. The super-resolved images using GNC as the optimization technique are shown in Figure 5.7. The MSE for this method is indicated in Table 5.1. Visually, there is a significant reduction in noise of the super-resolved images generated using the discontinuity preserving method. In yet another experiment, we carried out the optimization of the modified cost function using SA, but with the output of the GNC algorithm as the initial estimate of the super-resolved image. The super-resolved Lena, CT and Pentagon images obtained using this method are shown in Figure 5.8. Notice that the estimates of the GNC algorithm have undergone further deblurring resulting in a sharper image, e.g. around the eyes of Lena and on the Pentagon image as a whole. We noted earlier that in order to avoid the computational expense of simulated annealing, we opted for a convex cost function by choosing a suitable expression for the clique potentials. However, with incorporation of line fields and optimization using the GNC, which is proven to have a faster convergence than SA, we obtained a better estimate of the super-resolved image. When computational complexity is not an issue, we could go further and use the SA to obtain still better estimates.

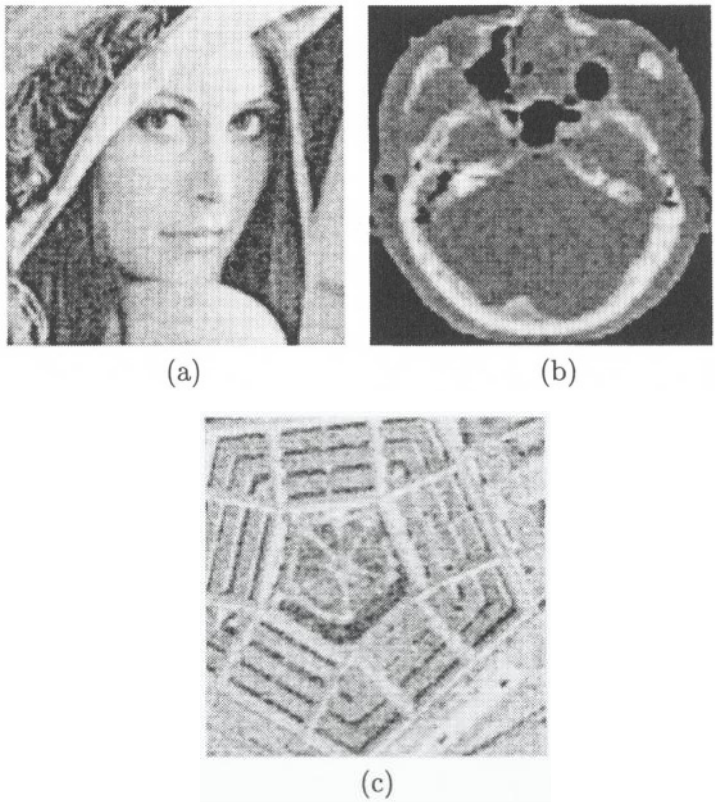
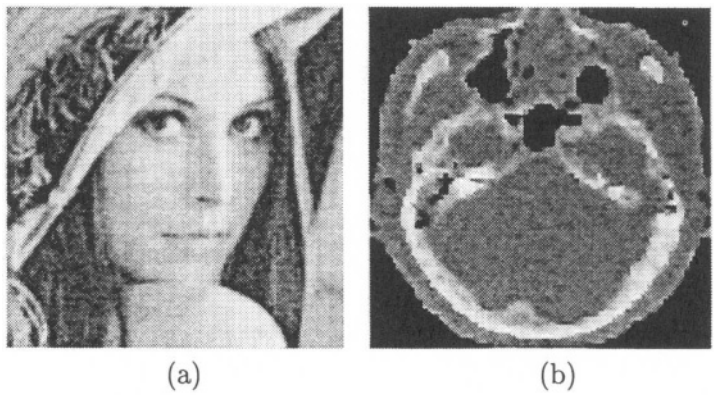


Figure 5.6. Super-resolved (a) Lena, (b) CT and (c) Pentagon images using only two low resolution observations with $\sigma = 0$ and $\sigma = 0.5$.

Simulations were also carried out to investigate the effect of the number of observations on the quality of the super-resolved image. As shown in Figure 5.9, the mean square errors decrease as the number of low resolution observations increases. The plots also illustrate the superiority of the discontinuity preserving method to the gradient descent approach. As noted earlier, the flat nature of the plots for the gradient descent approach implies that the errors do not reduce significantly, although more number of images do contribute to smoothening out noise. On the other hand, an increase in the number of images does bring about a substantial reduction in errors when line fields are included in the cost function.

Finally, we present some results from our on-going study on recovering a super-resolved image from a sequence of space varying blurred



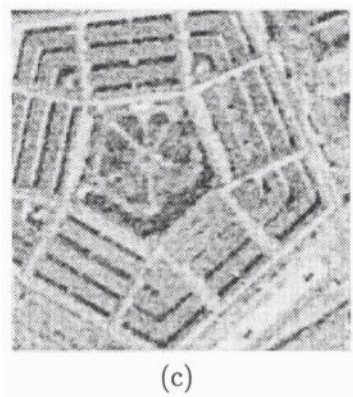


Figure 5.7. Super-resolved (a) Lena, (b) CT and (c) Pentagon images using the GNC optimization scheme.

observations. The details are available in [16]. We consider the case where the blurring is constant over a certain contiguous region of the image and then the amount of blurring varies linearly over a second region and finally is constant again over the remaining part of the image. Such a variation in the blur kernel occurs when the images are captured with a finite aperture lens and when the depth in the scene has similar variations. Two such blurred images of the sail image are shown in Figure 5.10(a) and (b) and the super-resolved image is shown in Figure 5.10(c). Unlike in previous experiments, the blur kernel was not assumed to be known; rather it was estimated locally at each pixel using the depth from defocus technique. The super-resolved image recovery technique has performed quite well. The numerals on the sail as well as the thin lines on the sail are discernible.

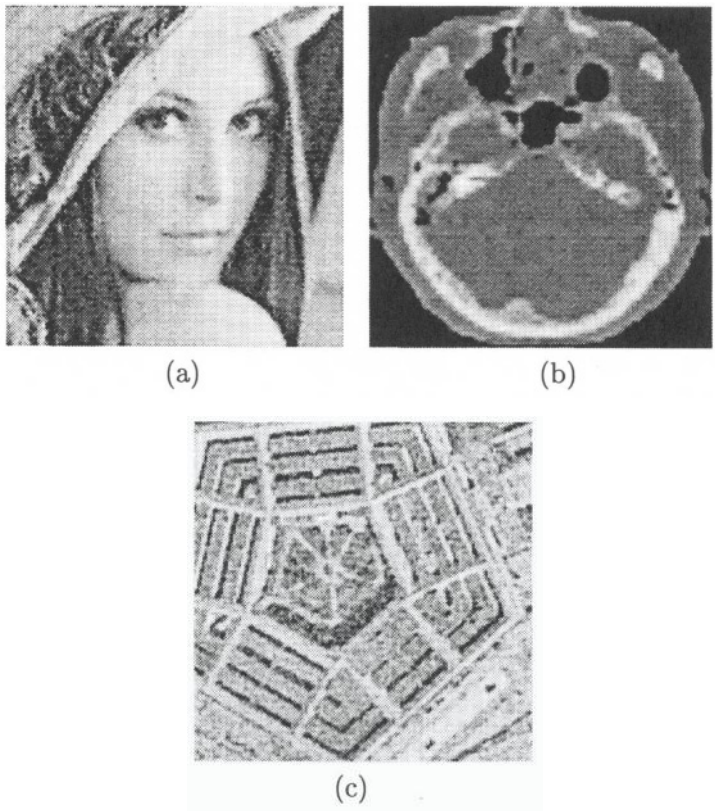


Figure 5.8. Super-resolved (a) Lena, (b) CT and (c) Pentagon images using simulated annealing (SA).

A second example of super-resolution from space varying blurred images is illustrated through low resolution observations of the Text image. Each observation of size 46 × 212 is blurred by a space varying blur having a step profile. Two of the five low resolution images are shown in Figure 5.11(a) and (b). Due to the step-like variation in the blur profile, we notice the text getting progressively blurred from the left to the right end of the input images. The estimated super-resolved text image is shown in Figure 5.11(c) in which the text is readable throughout. The above two examples illustrate that depth related defocus blur could be used as a natural cue for generating super-resolution images.

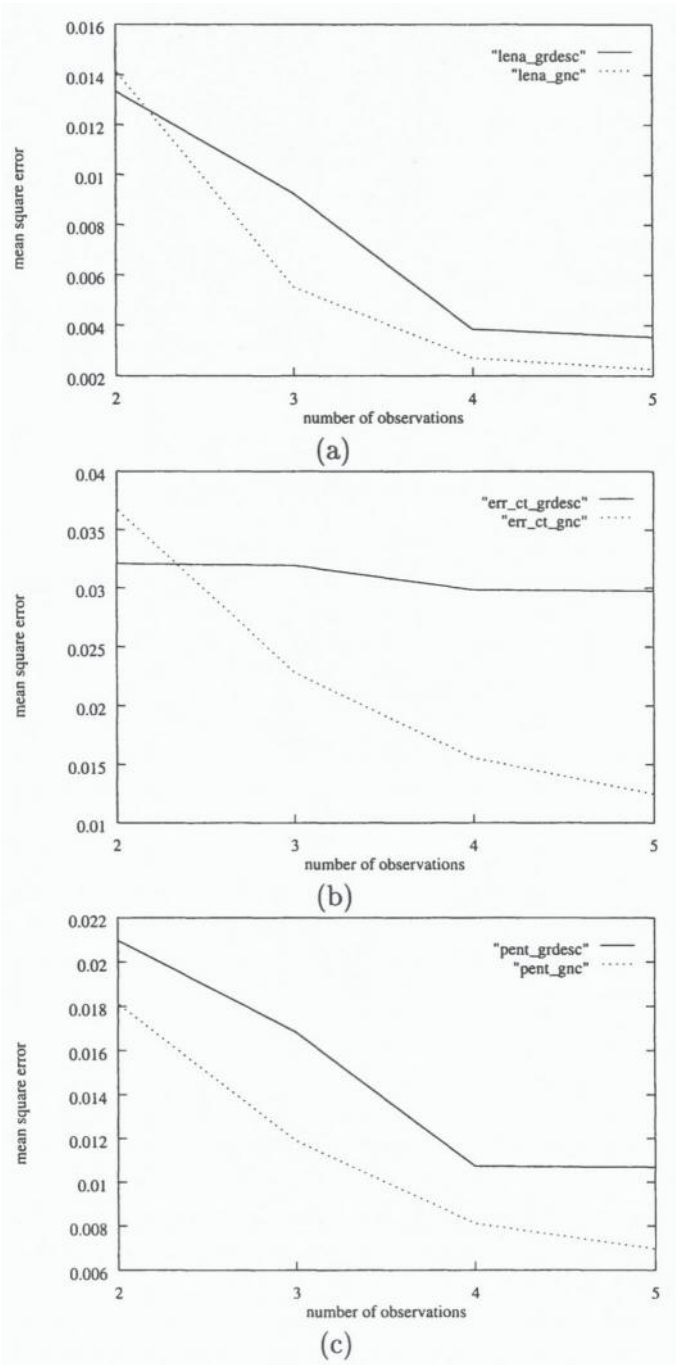


Figure 5.9. Comparison of mean square errors between the gradient-descent and the discontinuity preserving (GNC) approaches for (a) Lena, (b) CT and (c) Pentagon images, as the number of observations increases.

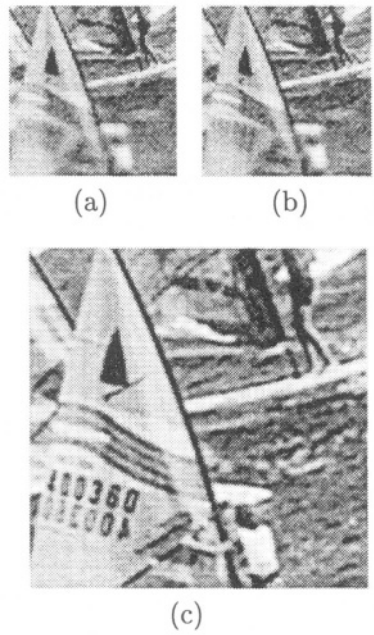


Figure 5.10. (a) and (b) Two of the low resolution Sail images and (c) the super-resolved Sail image.

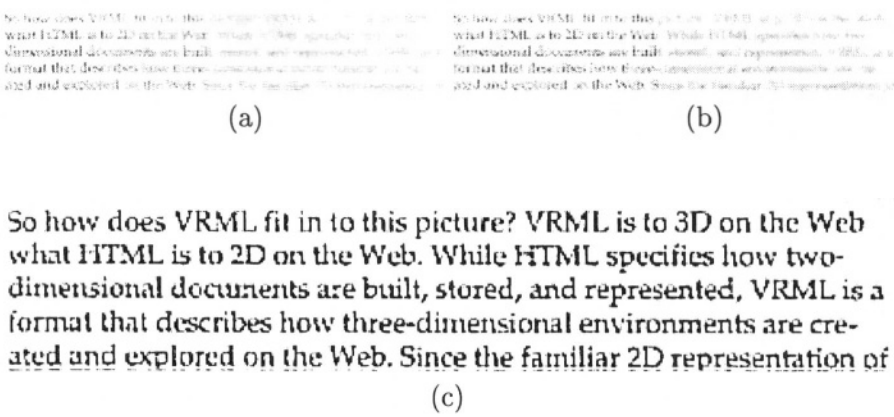


Figure 5.11. (a) and (b) Two of the low resolution Text images and (c) the super-resolved Text image.

6. Conclusions

This chapter addressed the problem of generating a super-resolution image from a sequence of blurred, decimated and noisy observations of

an ideal image. A MAP-MRF approach was used to minimize the function. Initially, the energy function was chosen to be convex by selecting the finite difference approximation of the first order derivative of the intensity at each pixel location. This enabled the use of steepest-descent type of algorithms to be used for minimization. The errors are seen to level off after about 35 iterations for all the images considered in this paper. Comparison with zero order hold and spline interpolation techniques shows that the proposed method is superior. Since there is no relative motion between the observed images, as is the case in most of the previous work in super-resolution, the difficult tasks of image registration and motion estimation are avoided. For the same reason, the performance of the proposed scheme cannot be compared with those obtained using motion-based super-resolution techniques. Next, the cost function was modified to include line fields to preserve discontinuitites, which is then minimized using the GNC algorithm. Since GNC is a sub-optimal optimization technique, we also used the more computationally intensive simulated annealing algorithm. In addition to significant noise reduction, the sharpness in the image was also observed to be enhanced.

In this chapter we assumed that the blurring kernels are known. However, in most practical situations, this may not be the case. Hence, the next natural step is to look at the problem of super-resolved restoration with unknown blurs. This translates to a joint blur identification and super-resolution restoration problem. Investigations are currently underway [16] to achieve the above.

References

[1] H. Ur and D. Gross, "Improved resolution from sub-pixel shifted pictures," *CVGIP:Graphical Models and Image Processing*, vol. 54, pp. 181–186, March 1992.

[2] S. Chaudhuri and A. N. Rajagopalan, *Depth from defocused images : A real aperture imaging approach*, Springer-Verlag, New York, 1999.

[3] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of Royal Statistical Society, Series B*, vol. 36, pp. 192–236, 1974.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and the Bayesian restoration of image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[5] Andrew Blake and Andrew Zisserman, *Visual Reconstruction*, MIT Press, 1987.

[6] R. Kindermann and J. L. Snell, *Markov Random Fields and their applications*, American Mathematical Society, Providence, RI, 1980.

[7] H. Derin and H. Elliot, "Modeling and segmentation of noisy and textured images using gibbs random fields," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, pp. 3–55, 1987.

[8] Richard C. Dubes and Anil K. Jain, "Random field models in image analysis," *Journal of Applied Statistics*, vol. 16, no. 2, pp. 131–164, 1989.

[9] S. Z. Li, *Markov Random Field Modelling in Computer Vision*, Springer-Verlag, Tokyo, 1995.

[10] J. L. Marroquin, *Probabilistic solution of inverse problems*, Ph.D. thesis, MIT AI Lab, 1985.

[11] Michael Elad and Arie Feuer, "Restoration of a single super-resolution image from several blurred, noisy and undersampled measured images," *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1646–1658, December 1997.

[12] M. Elad and A. Feuer, "Restoration of a single super-resolution image from several blurred, noisy and undersampled measured images," Tech. Rep. EE Pub No. 967, Dept. of Electrical Engg, Technion, Israel Instt. of Technology, May 1995.

[13] A. N. Rajagopalan and S. Chaudhuri, "Space-variant approaches to recovery of depth from defocused images," *Computer Vision and Image Understanding*, vol. 68, no. 3, pp. 309–329, Dec. 1997.

[14] R. R. Schultz and R. L. Stevenson, "A Bayesian approach to image expansion for improved definition," *IEEE Trans. on Image Processing*, vol. 3, no. 3, pp. 233–242, May 1994.

[15] S. Krishnamachari and Rama Chellappa, "Multiresolution gauss-markov random field models for texture segmentation," *IEEE Trans. on Image Processing*, vol. 6, no. 2, pp. 251–266, February 1997.

[16] Deepu Rajan and Subhasis Chaudhuri, "imultaneous estimation of super-resolved intensity and depth maps from low resolution defocused observations of a scene," in *Proc. of International Conf. on Compuner Vision*, Vancouver, Canada, 2001.

[17] Deepu Rajan, *Some new approaches to generation of super-resolution images*, Ph.D. thesis, School of Biomedical Engineering, Indian Institute of Technology, Bombay, 2001.

This page intentionally left blank.

Chapter 6

SUPER-RESOLUTION VIA IMAGE WARPING

Theory, Implementation and Evaluation

Terrance E. Boult, Ming-Chao Chiang and Ross J. Micheals
Computer Science and Engineering Deparartment
Lehigh University, Bethlehem, PA 18015, USA
tboult@lehigh.edu

Abstract

This chapter focuses on three issues: supporting image warping algorithms for super-resolution, examples of how image warping algorithms impact super-resolution image quality, and the development of quantitative techniques for super-resolution algorithm evaluation.

The warping approach proposed in this chapter is based on the integrating resampler [Chiang and Boult, 1996] which warps the image while both enforcing the underlying image reconstruction and satisfying the *imaging consistent constraint* [Boult and Wolberg, 1993]. The imaging consistent constraint requires that the image reconstruction yields a function which, when convolved with the imaging system's point-spread function (PSF), is consistent with the input image. Many popular reconstruction techniques, including bilinear and natural cubic splines, do not satisfy the imaging consistent constraint. In this chapter, we review imaging consistent warping algorithms, how they form the core of the integrating resampler, and their implementation.

Although imaging consistent warping techniques can be used in other super-resolution implementations, such as those discussed in Chapter 8, we present its use in a simpler direct approach: warping followed by a straightfoward fusion. Examples are provided on grayscale images of simple patterns, text, and human faces. The use of priors in the fusion, such as those used in Chapter 10 could further enhance the results, but we analysize the simpler approach to isolate the impact of the warping algorithm.

The chapter then discusses the important problem of quantitative evaluation and presents a summary of two different quantitative experiments: using OCR and face recognition as metrics. These experiments

clearly show the importance of high-quality reconstruction and warping to super-resolution. Perhaps more importantly, these experiments show that even when images are qualitatively similar, quantitative differences appear in machine processing. As the super-resolution field is pushed towards its boundardies, the ability to measure progress, even if it is small, becomes increasingly important.

Keywords: Super-Resolution, Imaging-Consistent Restoration/Reconstruction, Integrating Resampling, Integrating Resampler, Quantitative Evaluation, OCR, Bi-linear Resampling, Image Reconstruction, Image Restoration, Image Warping, Balanced Repeated Replicates, Replicate Statistics, Face Recognition.

1. Background and Introduction

The fundamental purpose of image warping is to allow the reshaping of image geometry for a variety of applications. Inherent in any super-resolution algorithm that uses multiple images is the alignment, or "matching," of data among the images—the computation of a mapping from each pixel in the low resolution image to a pixel in the super-resolution image. Except in specalized devices that intentionally cause precise sub-pixel shifts, alignment is almost always to a regular grid, and hence can be viewed as a general warp of the input image. General image warping, as is needed for super-resolution, requires the underlying image to be resampled at non-integer, and generally spatially-varying locations. Hence, super-resolution requires sub-pixel image reconstruction, but is not necessarily amenable to efficient image reconstruction via convolution. When the goal of warping is to produce output for human viewing, only moderately accurate image intensities are needed. In these cases, techniques using bilinear interpolation have been found sufficient. However, as a step for applications such as super-resolution, the precision of the warped intensity values is often important. As we shall show in this chapter, super-resolution based on bilinear image reconstruction may not be sufficient.

One of the first explicit uses of image warping for super-resolution was in [Peleg et al., 1987, Keren et al., 1988]. Peleg and Keren estimated an initial guess of the high-resolution image, and simulated the imaging process via warping so that the difference between the observed and simulated low-resolution images was minimized. Irani and Peleg [Irani and Peleg, 1991, Irani and Peleg, 1993] used a back-projection method similar to that used in tomography to minimize the same difference. Bascle et al. [Bascle et al., 1996] extended this back-projection method

to include a simple motion blur model. We note, however, that all previous work has ignored the impact of image warping techniques.

Not all prior image-based work has used image warping. Algebraic approaches do have some significant advantages; analysis of the underlying linear systems may constrain the blur kernel enough to permit the computation of new information. Also, algebraic approaches are more naturally extended to allow for Bayesian estimation and the use of priors. Note, however, that algebraic constraints still require sub-pixel evaluation of the input for each pixel in the super-resolution image, which is tantamount to warping. One can view warping as a pre-processing that takes the spatial alignment and matching information and generates reconstructed images that would make solution of the algebraic equations more efficient. The lack of high-quality reconstruction for warping may be the unstated reason that algebraic techniques have not embraced warping.

This chapter is structured as follows. In Section 2, the image formation process and the relationships between restoration, reconstruction, and super-resolution are briefly reviewed. The integrating resampler - an efficient method for warping using imaging-consistent reconstruction & restoration algorithms - is given in Section 3. In Section 4, we introduce the super-resolution algorithms considered in our analysis. Quantitative measurement of super-resolution imaging using three different applications is shown in Section 5.

2. Image Formation, Image Restoration and Super-Resolution

To address the problem of super-resolution, we need to first understand the process of image formation, reconstruction, and restoration. Although previous chapters provide most of the necessary background, to better describe our warping and super-resolution techniques, we briefly review the image formation process and sensor model as proposed in [Boult and Wolberg, 1993].

Generally, image formation can be described as a cascade of filtering operations. There is an overall blur applied at each pixel, $h(x,y)$, that can be decomposed as the sequence of operations as shown figure Fig. 6.1. Let $f(x,y)$ be the intensity distribution of a scene in front of a lens aperture. That distribution is acted upon by the blurring component of the lens, $h_1(x_y)$, yielding $f_1(x,y)$. The application of a geometric distortion function, $h_2(x,y)$, produces image $f_2(u,v)$. At this point, $f_2(u,v)$ strikes the image sensor where it undergoes additional blurring

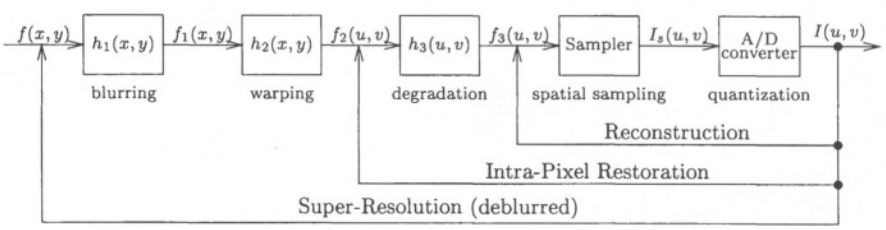


Figure 6.1. The image formation process and the relationship between restoration, reconstruction, and super-resolution.

by a point spread function, $h_3(u,v)$, which generates image $f_3(u,v)$. This blurring reflects the limitation of the sensor to accurately resolve each point without the influence of neighboring points. We choose to use a simple model wherein this blurring takes place within one pixel because for CCD and CID cameras the physical boundaries between photosites generally allow only insignificant charge transfer between pixels. Image $f_3(u,v)$ undergoes spatial sampling as it hits the discrete CCD or CID photosites. The combination of convolution with the photosite blur h_3 and sampling is known as area sampling and reflects the finite size of a discrete photosite. If h_3 was assumed to be an impulse, then we have point sampling. While point sampling is often assumed for theoretical considerations, it is not true in practice. In either case, intensities in the sampled image I_s are now defined only for integer values of u and v. The digital image $I(u, v)$ is obtained via an analog-to-digital converter that quantizes the samples of I_s. Note that parts of this decomposition are more conceptual than physical since, for instance, the geometric and blurring components occur simultaneously.

Reconstruction and restoration start with I (and models for one or more blur kernels $h_i(x,y)$), and seek to solve for one or more $f_j(x,y)$. Recovering an approximation of $f(x, y)$ is known as image *restoration* and is of considerable interest in image processing. The most common formulation of that problem, however, is actually recovering a discretized, rather than continuous, form of f. Recovering $f_2(x,y)$ might be called intra-pixel restoration, though it is is not commonly discussed in the literature.

Given this image formation model we might define super-resolution as the use of multiple images and/or prior model information to recover an approximation to $f(x, y)$ better than what would be obtained by image reconstruction followed by debluring using knowledge of $h_j(x,y)$. This definition includes approximating the image at a larger size with reason-

able approximations for frequencies higher than those representable at the original size. While it may seem non-traditional it also includes improving the SNR while keeping the image size fixed. Given such an SNR improved image, one could simply perform a finer resampling and deblur to obtain a super-resolution with increased spatial resolution. Note that since debluring amplifies noise, the increased SNR can have a more significant result on the super-resolution image than might be initially expected. In practice, however, one would want to improve the SNR at the higher spatial resolution to reduce the impact of reconstruction artificats when increasing the resolution.

Because of the multiple, and different degradations in this imaging model, we will define two different types of super-resolution that will be considered in this chapter. Recovering a discrete approximation, with resolution higher than I_s, to f_1 is called (plain) super-resolution and appoximation to f is called super-resolution with deblurring. Note that super-resolution with deblurring requires knowledge of the primary bluring kernel — a reasonable assumption for simple lens blur but tenuous for atomospheric or depth of field effects. Because super-resolution increases the signal-to-noise ratio in the approximation to f_1, it significantly ameliorates the ill-conditioned nature of deblurring.

3. Imaging-Consistency and The Integrating Resampler

Image reconstruction plays a key role in all super-resolution algorithms. Given the finite set of samples, there is an uncountably infinite number of functions that satisfy the data, and hence, image interpolation involves adding regularization constraints to allow a unique function to be defined, given the image data. Often there is a need to balance computational complexity against the sophisticated nature of the assumptions and constraints. The many constraints developed in the design of image reconstruction filters have been extensively discussed: in books [Andrews and Hunt, 1977, Pratt, 1978, Gonzalez and Wintz, 1987, Pavlidis, 1982, Wolberg, 1990, Pratt, 1990], articles [Simon, 1975, Andrews and Patterson, 1977, Hou and Andrews, 1987, Park and Schowengerdt, 1982, Reichenbach and Park, 1989, Jain, 1989, Oakley and Cunningham, 1990], and comparison papers [Parker and and D.E. Troxel, 1983, Mitchell and Netravali, 1988, Maeland, 1988]. Many of these constraints are related to how well the underlying filter approximates the ideal sinc filter. Even the "ideal" sinc interpolation is based on the assumption that the image is an infinite signal sampled at or above its Nyquist rate. While it is true that optics limit image bandwidth, it

need not result in images that are Nyquist sampled. If the underlying function f was Nyquist sampled, then except for noise removal, there is no need for super-resolution.

In [Boult and Wolberg, 1993], a new constraint was added to the mix of potential assumptions for image reconstruction: requiring the algorithm to be *imaging-consistent.* An algorithm is called imaging-consistent if it is the exact solution for some input function, which, according to the imaging model, would have generated the measured input. This constraint is particularly important for super-resolution because it means each resampling would, when subjected to the imaging model, actually be consistent with the measured image.

For image reconstruction, we can achieve a imaging-consistent reconstruction by first restoring the image to yield an approximation to f_2, then performing an additional blur by the pixel's PSF. Although restoration is ill-posed, blurring produces an image reconstruction that is totally consistent with the input data, regardless of the resampling rate. The use of image restoration technique permits the work presented in this chapter to achieve image reconstruction in a fundamentally different way than traditional approaches. Our approach is in the spirit of the work of [Huck et al., 1991], where it is argued that sampling and image formation should be considered together. Imaging-consistent algorithms directly combine knowledge of image formation and sampling into the reconstruction & restoration process. The way that knowledge is used, however, is quite different from [Huck et al., 1991].

Imaging-consistent algorithms follow quite naturally from a general approach to algorithm development known as information-based complexity (IBC) (see [Traub et al., 1988]). From IBC, it can be shown that the imaging-consistent algorithms enjoy very good error properties for many definitions of error. In particular, imaging-consistent algorithms have, within the presribed space of functions, an error at most twice that of any algorithm for *any* error measure defined as a weighted norm on the space of solutions (e.g., L^2, or even a weighted least-squares measure). Note that most image-quality measures yielding a scalar are error measures of this type — e.g., the QSF measure of [Drago and Granger, 1985, Granger, 1974], QSF extensions that include contrast effects, any weighted integral of the modulation transfer function (MTF), and the measure of [Park and Schowengerdt, 1982] when weighted and integrated over frequency v. For the algorithms discussed here we presume the space of functions are continuous and piecewise analytic with a bounded first derivative in each piece. More discussion of these error properties, and alternative spaces of functions, can be found in [Chiang, 1998].

Of course, an algorithm that performed a full restoration followed by blurring could be computationally expensive. Fortunately, with some effort, the imaging consistency constraint can be applied in a functional way, and incorporated into a very efficient algoirthm. In this chapter, only an overview is provided. For more details, and the derivation of four other imaging consistent algoirthms, see [Boult and Wolberg, 1993, Chiang, 1998]. One-dimensional image models are presented herein, but since higher dimensions may be treated separably, the process is easily extended.

The simplest imaging-consistent method to consider is based on a piecewise quadratic model for the image. If we assume each photosite PSF (h_3) is a Rect filter (1.0 inside the pixel, zero otherwise), an imaging consistent algorithm is easy to derive. To ensure that the function is continuous and local, we define the value of the reconstruction at the pixel boundaries k_i and k_{i+1} to be equal to E_i and E_{i+1}. Any method of approximation could be used to compute E_i, though our examples will only include cubic convolution or linear interpolation. See [Chiang, 1998, Section 2.3] for a more detailed derivation and more examples.

Given the values E_i at the pixel edges, an imaging consistent constraint is that the integral across the pixel must equal V_i. This results in exactly three constraints:

$$g_i(-1/2) = E_i; \quad g_i(1/2) = E_{i+1}; \quad \int_{-1/2}^{1/2} g_i(x)\,dx = V_i. \tag{6.1}$$

From Eq. (6.1), one can derive the following quadratic polynomial

$$\begin{aligned} g_i(x) &= 3(E_i + E_{i+1} - 2V_i)x^2 + (E_i - E_{i+1})x \\ &\quad -(E_i + E_{i+1} - 6V_i)/4, \end{aligned} \tag{6.2}$$

where $-1/2 \le x \le 1/2$. Using cubic convolution with parameter a to derive E_i and E_{i+1} yields

$$\begin{aligned} E_i &= \frac{1}{8}(aV_{i-2} + (4-a)V_{i-1} + (4-a)V_i + aV_{i+1}), \\ E_{i+1} &= \frac{1}{8}(aV_{i-1} + (4-a)V_i + (4-a)V_{i+1} + aV_{i+2}). \end{aligned} \tag{6.3}$$

So that the cubic convolution kernel resembles the sinc funtion, the parameter a is generally in the range [-3, 0], with the values –0.5, –0.75, and –1.0 having special significance (see [Simon, 1975, Keys, 1981, Park and Schowengerdt, 1983]). Note that with $a = 0$, we have

$$E_i = (V_{i-1} + V_i)/2 \quad \text{and} \quad E_{i+1} = (V_i + V_{i+1})/2. \tag{6.4}$$

In other words, for $a = 0$, cubic convolution interpolation of the edge values (i.e., midpoints between pixels) is equal to the value given by bilinear interpolation.

When applied over the entire image, Eq. (6.2) yields f_2, an intra-pixel restoration. If an imaging-consistent reconstruction is desired, it may be obtained from the intra-pixel restoration via covolution with the pixel PSF. Assuming a Rect PSF for the pixel, one can integrate Eq. (6.2) to derive a functional form for reconstruction. The result is the per-pixel cubic polynomial

$$\begin{aligned} G_i(x) &= \int_{x-1/2}^{1/2} g_i(z)\,dz + \int_{-1/2}^{x-1/2} g_{i+1}(z)\,dz \qquad (6.5) \\ &= (E_{i+2} - E_i - 2(V_{i+1} - V_i))x^3 \\ &\quad + (2E_i - E_{i+1} - E_{i+2} + 3(V_{i+1} - V_i))x^2 + (E_{i+1} - E_i)x + V_i, \end{aligned}$$

where $0 \le x \le 1$ spans from the center of one input pixel to the next.

It is interesting to note that if $a = 0$ as in Eq. (6.4) (i.e. linear interpolation) is used to determine E_i, the resulting imaging-consistent reconstruction (Eq. (6.5)) is tantamount to cubic convolution with the "optimal" value of $a = -0.5$ — proof can be found in [Chiang, 1998, Section 2.4]. No other value of a yields a reconstruction that satisfies the imaging-consistent constraint with a simple PSF. That is, if we use cubic convolution with $a \neq 0$ to estimate E_i, the resulting imaging consistent polynomial is not equivalent to any cubic convolution. We have found that using cubic convolution with $a = -0.5$ to estimate E_i is one of the best imaging consistent algorithms and it is the value used for most of the examples in this chapter.

This section presented a model that is globally continuous and analytic except on the pixel boundaries, which results in a per pixel model which is quadratic after restoration (cubic after reconstruction). In [Boult and Wolberg, 1993, Chiang, 1998], we also present/analyze alternatives that are globally differential or smoother, and also models that have multiple polynomials per pixel.

3.1. Imaging Consistent Warping: The integrating resampler

To define an imaging-consistent warping, we generalize the idea of the imaging-consistent reconstruction/restoration. Whereas imaging-consistent reconstruction assumes that the degradation models are identical for both input and output, imaging-consistent warping allows both the input and output to have their own degradation models, and also allows for the degradation models to vary its size for each output pixel.

The imaging-consistent algorithms described above and in [Boult and Wolberg, 1993] are linear filters. We designed them for use in what we call the *integrating resampling* approach. For the super-resolution results described herein, we consider only the integrating resampler assuming a Rect PSF filter as described in [Chiang and Boult, 1996], which we refer to as QRW.

As described before, our model of image formation requires the image to be spatially sampled with a finite area sampler. This is tantamount to a weighted integral being computed on the input function. Because we have a functional form for the restoration, we can simply integrate this function with the PSF for the output area sampler. In this section, although we assume that the output sampler has a Rect PSF, it should be noted that there is no limitation on other potential degradation models. However, Rect is used not only to simplify the algorithms, but because it is a good model for super-resolution where each photosite is represented with a pixel.

When resampling the image and warping its geometry, this new approach allows for efficient pre-filtering and post-filtering. Additionally, because a functional form of the input has already determined, no spatially-varying filtering is needed, unlike a case using a direct inverse mapping.

Computing the exact value of the imaging-consistent warped value (the integrated restored function weighted by the PSF) can represented in functional form if the mapping function has a functional inverse and the PSF is simple. In general, however, super-resolution algorithms may have complex maps requiring numerical integration, since such maps cannot be represented in closed form. To reduce the computational complexity, we propose a scheme where for within each input pixel, we use a linear approximation to the spatial warp, but use the full non-linear warp to determine the location of pixel boundaries. This integrating resampler, first used in [Boult and Wolberg, 1992] and formally described in [Chiang and Boult, 1996], also handles anti-aliasing of partial pixels in a straightforward manner.

Assume n input pixels are being mapped into k output pixels according to the mapping function $m(t)$. Let m_i be the mapped location of pixel i, for $i=0,\ldots,n$. Compute δ_j, $j=0,\ldots,k$, as the linear approximation to the location of $m^{-1}(j)$, as shown in Fig. 6.2. To avoid fold-over problems, we assume that the mapping function is strictly increasing. For an approach to modeling fold-over, see [Wolberg and Boult, 1989].

For efficient computation of the integral as well as the ability to perform proper antialiasing, the integrating resampler, given as pseudo code in Fig. 6.3, "runs" along the input and output determining in which im-

```
for (i = j = 0; j ≤ k; j++) {
  while (i < n − 1 && m_{i+1} < j) i++;
  δ_j = i + (j − m_i)/(m_{i+1} − m_i); }
```

Figure 6.2. Linear approximation to the location of $m^{-1}(j)$.

i	V_i	E_i	g_i	R
0	0	0	0	0
1	0	0	0	0
2	0	0	$-47.8x^2 - 15.9x + 3.9$	$-47.8t^3 - 15.9t^2 + 3.9t$
3	0	-15.94	$334.6x^2 - 143.4x - 27.8$	$334.6t^3 - 143.4t^2 - 27.8t - 91.6$
4	255	127.50	$-334.6x^2 + 143.4x + 282.8$	$334.6t^3 - 143.4t^2 - 27.8t - 219.1$
5	255	270.94	$47.8x^2 - 15.9x + 251.0$	$47.8t^3 - 15.9t^2 + 251.0t - 135.4$
6	255	255.0	255	255t -127.5
7	255	255.0	255	255t -127.5

Table 6.1. A simple step edge, and the resulting values and polynomials that serve as input to the integrating resampler. In computing the edge values, we presume pixel replication outside the image.

age the next pixel boundary will be crossed. To do this, there are two variables: *inseg* $\in [0, 1]$, which represents the fraction of the current input pixel left to be consumed, and *outseg*, which specifies the amount of input pixels required to fill the current output pixel. In the integrating resampler, the function $R(t; g)$ is obtained from the definite integral of an imaging-consistent restoration $g(x)$ as

$$R(t; g) = \int_{-0.5}^{t} \mathrm{PSF}(x)\, g(x)\, dx \tag{6.6}$$

which, natrually, changes according to each pixel. An example showing image values V_i, edge values E_i, imaging consistent intra-pixel restoration g_i, and imaging consistent reconstuction R is presented in table 6.1. Note this is *not* the same cubic polynomial as Eq. (6.5) - an integral over a full pixel size by our prevous definitions, implies a combination of two different quadratics. The table shows values only within individual pixels.

Assuming proper update to the algorithm's state, whenever *inseg* < *outseg*, we know that the input pixel will finish first, so it may be consumed. If, on the other hand, it happens that $inseg \geq outseg$, the output pixel will finish first, so an output is produced. Thus, in each iteration of the loop we either consume one input pixel or produce one output pixel. Therefore, the algorithm requires at most $k + n$ iterations.

Pad the input; compute k_l, k_r, and i_l, the indices to the leftmost and rightmost output pixels and the index to the leftmost input pixel that contributes to the output; and compute the linear approximation to the location of $\delta_j = m^{-1}(j)$, for $j = k_l, \ldots, k_r + 1$.

```
nfactor = δ_{k_l+1} − δ_{k_l};          //  set up for normalization
δ_{k_l} = MAX(δ_{k_l}, 0);              //  ensure that δ_{k_l} is nonnegative
inseg = 1.0 − FRACTION(δ_{k_l});        //  fraction of input pixel left to be
                                            consumed
outseg = δ_{k_l+1} − δ_{k_l};           //  #input pixels mapped onto one
                                            output pixel
acc = 0.0;                              //  reset accumulator for next output
                                            pixel
for(j = 0; j < δ_{k_l}; j++)out[j++] = 0;   //  zero out the garbage at left end
for (i = i_l, j = k_l; j <= k_r; ) {        //  while there is output to produce
  Use the current pixel (in[i]) and
  neighbors to update R(), the integral
  of the restoration g().
  left = 1.0 − inseg;                   //  get left endpoint for integration
  if (inseg < outseg) {                 //  if we will consume input pixel first
    acc += R(1) − R(left);              //  add integral to end of output pixel
    i++;                                //  index into next input pixel
    if (i == n) {                       //  check end condition
      if (normalize) acc /= nfactor;    //  normalize the output, if appropri-
                                            ate
      out[j] = acc;                     //  init output
      break;                            //  exit from the loop
    }
    outseg −= inseg;                    //  inseg portion has been filled
    inseg = 1.0;                        //  new input pixel will be available
  }
  else {                                //  Else we will produce output pixel
                                            first
    acc += R(left + outseg) − R(left);  //  add integral to end of output pixel
    if (normalize) acc /= nfactor;      //  normalize the output, if appropri-
                                            ate
    out[j] = acc;                       //  init output
    j++;                                //  index into next output pixel
    acc = 0.0;                          //  reset accumulator for next output
                                            pixel
    inseg −= outseg;                    //  outseg portion of input has been
                                            used
    outseg = δ_{j+1} − δ_j;             //  new output size
    nfactor = outseg;                   //  need for normalization
  }
}
for(j=k_r+1;j<k;j++)out[j++]=0;         //  zero out the garbage at right end
```

Figure 6.3. The integrating resampler assuming a output model of an Rect PSF filter. See text for discussion.

The underlying idea of this integrating resampler can be found in the work of Fant [Fant, 1986] who proposed an efficient bilinear warping algorithm. With some effort, one can see that by setting $R(t) = v_i t - 0.5$, the integrating resampler implements a bilinear warp.

In summary, the contribution discussed in this section is twofold:

1 the generalization of Fant's orginal algorithm into the integrating resampler which supports the use of advanced imaging-consistent reconstruction algorithms, and

2 the provision for modeling real lens effects by using real warps that affect the image radiance. In Fant's original work, the goal was to warp images for graphic effects, and hence to affect geometry without disturbing the intensities. To do this, the algorithm maintains knowledge of the input size and normalizes the integral to account for this size, giving a normalized intensity. Thus, if a constant image was stretched to twice its normal width, it would change shape but retain the same intensities. If a lens was placed into an imaging system so as to double the width of the image on the sensor plane, then the value measured would be halved. The algorithm is flexible enough to support both "graphics" and "lens" modeling. If the super-resolution is over images that vary because of, say, atmospheric variations, or if we are correcting for lens distortions, an unnormalized warp should used.

These contributions are at the foundations of our fusion of image consistent warping and super-resolution.

4. Warping-based Super-Resolution

We turn now to the use of warping for super-resolution. As described in earlier chapters, super-resolution refers to the process of constructing high-resolution images from low-resolution image sequences. Given the image sequence X_L, our warping-based super-resolution algorithm is formulated, as follows:

Define Reference and Mapping Choose one of the images, say X_p, as the reference image, and compute the motion field between all the images and the reference image.

Warp Scale up the reference image using QRW, and then use QRW to warp all the images to the reference image based on the motion field and scale computed in the previous step.

Fusion Obtain a super-resolution image by fusing all the images together.

Deblur (If desired) Deblur the resulting super-resolution image using $h_1(x, y)$.

This method presumes that lens blur, h_1, is approximately the same for all images. Otherwise, the deblurring step must be performed before the fusion stage. However, because of the noise amplification caused by deblurring in each image, deblurring before fusion is not as effective and should be used only when necessary.

We presume that a dense motion-field is computed between each image in the sequence - a straightforward calculation for motion that locally approximates a rigid transform. The motion field computations for the examples presented in this chapter are based on a sum-of-square difference matching algorithm with 11x11 and 7x7 template windows for the first and second experiments, respectively. In each case, the matching is a dense disparity surface. Sub-pixel estimates are obtained by fitting a quadratic to each point at which the match is unique and to its two neighbors. When off-the-shelf lenses and cameras are used, pre-warping can be used to remove the distortions. In [Chiang and Boult, 1996], we showed that the pre-warping with integrating resampler can improve the match quality given significant lens distortior.

In the face-based experiments, we did not have access to a face image database with a large number of views of the same subject. (We used the FERET database; more on this later). Therefore, we used a set of synthetic downsamplings from a single image to generate a sequence of perspective warps. Since the experiment called for warping so many images we directly used the matching information defined by the synthetic mappings. The mappings for each face were randomly generated, but the same set was used for both the bilinear and the QRW super-resolution warpings.

The fusion step is not the focus of this chapter, nor of our past work. We have tried several different approaches to fuse the images together, including a simple averaging or a median filter. Our experiments show that the median filter is better, though often not much better than the averaging filter. Median filtering is used for the OCR experiments and simple averaging for the others. More advanced techniques using priors, (see Chapter 10), could probably produce better results but would still be expected to benefit from the increased quality in fusion input. Again, the experiments in this chapter sought to isolate the effects of warping.

The test data shown in this section was taken using two different Sony cameras, models XC-77 and XC-999, captured by a Datacube MV200 System. Fig. 6.4 show our an experimental result with Fig. 6.5 showing the same results except that all the images are normalized so that the

dynamic ranges are identical. All the resulting super-resolution images are 256 × 256, and were scaled-up by a factor of approximately four (4). We note that the previous works [Gross, 1986, Peleg et al., 1987, Keren et al., 1988, Irani and Peleg, 1991, Irani and Peleg, 1993, Bascle et al., 1996] report results only scaling by a factor of two (2).

Fig. 6.4 shows the super-resolution results of our first example. Fig. 6.4 (a) shows an input image blown up by a factor of 4 using pixel replication so that the value of each pixel can easily be seen. Fig. 6.4 (b) shows super-resolution by our implementation of the back-projection method described in [Irani and Peleg, 1991] (not the ordinal authors, see [Chiang and Boult, 2000, Chiang, 1998] for details); Fig. 6.4 (c) shows super-resolution using bilinear resampling followed by deblurring; Fig. 6.4 (d) shows super-resolution using QRW followed by deblurring. Also, for the purpose of comparison, we assume that Figs. 6.4 (c) and 6.4 (d) have undergone the same degradation before sampling. Fig. 6.4 (e) and (f) show the super-resolution results without deblurring.

Fig. 6.5 shows the results after all the images are normalized so that the dynamic ranges are identical, as follows:

$$I_n = \frac{255}{\max_u - \min_u}(I_u - \min_u)$$

where I_n and I_u are, respectively, the normalized image and the image to be normalized, $\max_u$ and $\min_u$ are, respectively, the minimum and maximum intensity values of the image to be normalized.

Fig. 6.6 shows an example captured with a Sony XC999, which is a one-chip color camera. Note the target is similar to (yet different) from that in the first example. Fig. 6.6a shows one of the original images blown up by a factor of 4; it can be easily seen that inter-frame motion is involved in this case. Fig. 6.6b shows super-resolution using QRW followed by deblurring. Obviously, our super-resolution method removes most of the interframe motion and significantly improves the sharpnes of the image.

We implemented the back-projection method proposed in [Irani and Peleg, 1991] and found it somewhat difficult to work with since it is sensitive to the choice of its parameters called normalizing factors. For the comparisons, we tried many normalizing factors and chose one that resulted in the back-projected images (Fig. 6.4d) with minmal sum-of-square difference (SSD) between the observed and simulated images. It is worth pointing out that in this particular case, SSD is not necessarily a good error measure because it is not robust. Furthermore, the same normalizing factor does not always give the best result in terms of the

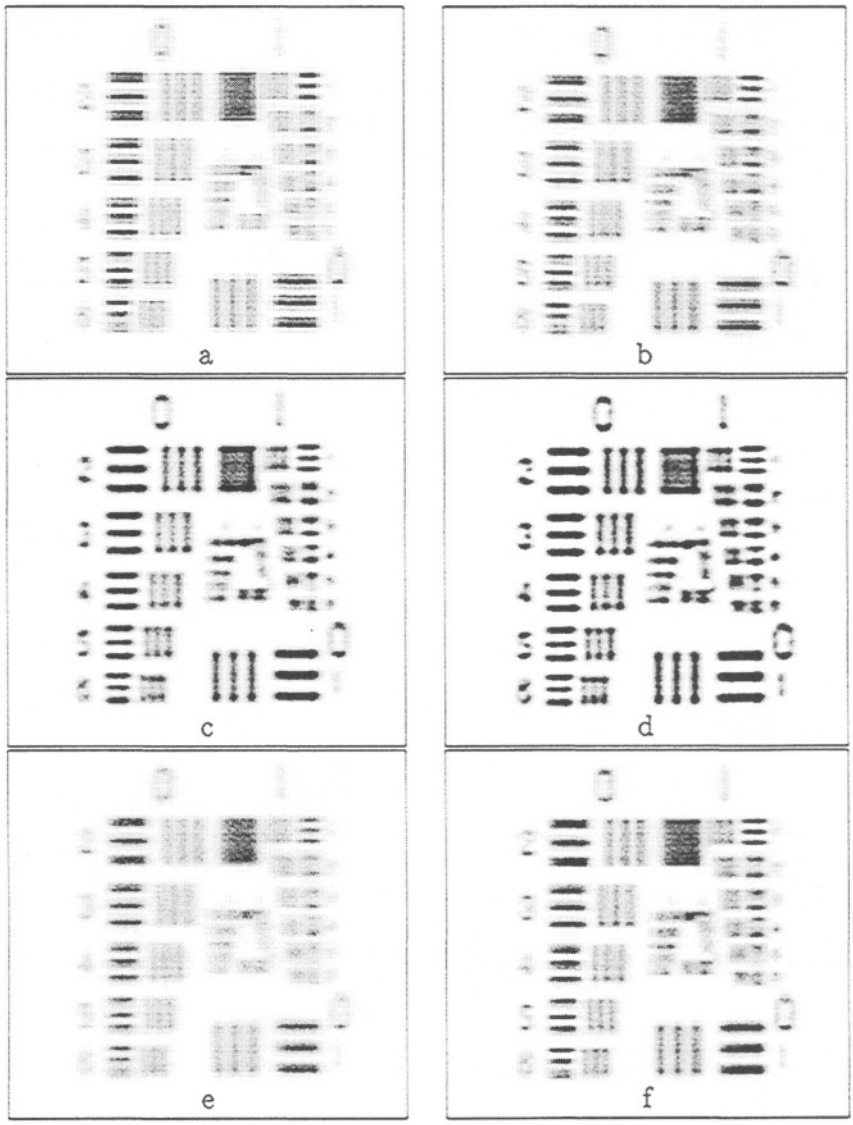


Figure 6.4. Final results from an 8 image sequence (64x64) taken by XC-77. (a) An orginal image blown up by a factor of 4 using pixel replication; (b) super-resolution by back-projection using bilinear resampling to simulate the image formation process and (e) as the initial guess; (c) super-resolution using bilinear warping followed by deblurring; (d) super-resolution using QRW followed by deblurring. Image (e) shows (c) (bilinear warping) without deblurring and (f) shows (d) (QRW) without deblurring.

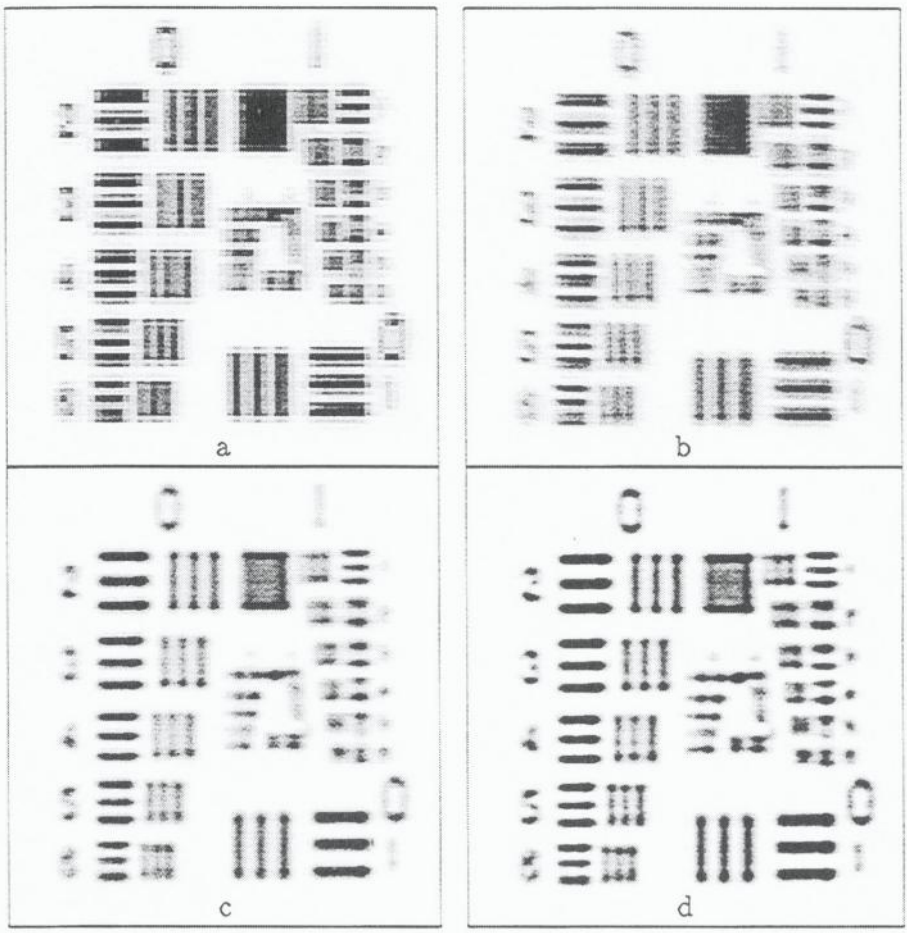


Figure 6.5. Results of Fig. 6.4 after being normalized to have the same dynamic range. (a) An orginal image blown up by a factor of 4 using pixel replication; (b) super-resolution by back-projection (c) super-resolution using bilinear warping followed by deblurring; (d) super-resolution using QRW followed by deblurring.

error measure when different resampling algorithms are used or when the input set is changed.

Results from our experiments show that the direct method we propose herein is not only computationally cheaper, but it also gives results comparable to or better than those using back-projection. Moreover, it is easily seen from Fig. 6.4 that the integrating resampler outperforms traditional bilinear resampling. Not surprisingly, our experiments show that most of the additional information carried by each image is concentrated on the high frequency part of the image. This observation also explains why the integrating resampler outperforms bilinear resampling.

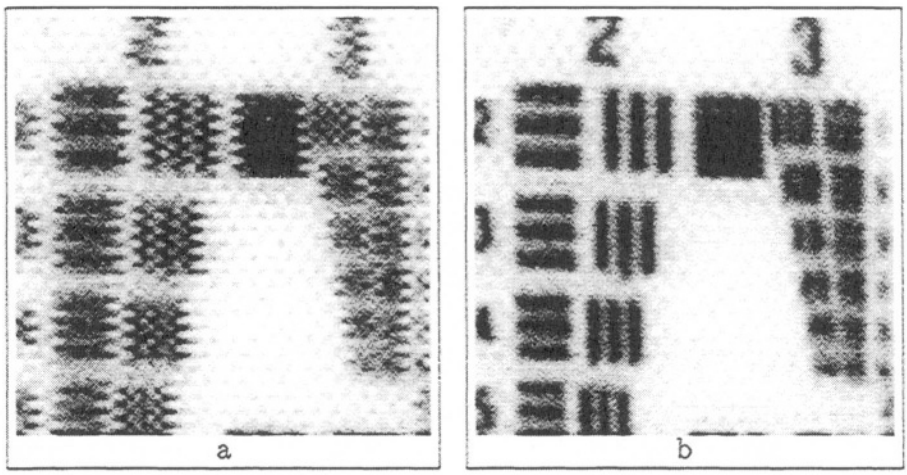


Figure 6.6. Super-resolution results from a very noisy image sequence of 32 images (64x64) taken by XC-999. (a) one of the original images blown up by a factor of 4; (b) super-resolution with QRW followed by deblurring.

As was shown in [Boult and Wolberg, 1993], when viewed as a reconstruction filter, bilinear causes more blurring than the imaging-consistent reconstruction of Eq. (6.5).

	QRW		Back-Projection	
Time in seconds	SPARC	Pentium	SPARC	Pentium
Warping	1.94	3.52	NA	NA
Fusion	0.04	0.25	NA	NA
Deblurring	0.81	1.31	NA	NA
Total	2.79	5.08	12.33	24.80

Table 6.2. Running times for our first examples (8 images) assuming the same degradation model. See text for details.

Table 6.2 gives the running times of our first example, as measured in 1996, using a Sun 143MHz Ultra SPARC running Solaris 2.5 and a 120MHz Pentium running Linux 2.0.12. Note that for both the maintenance of precision and ease of implementation, all operations were performed in double-precision floating point, and neither algorithm was explicitly optimized. Also, note that the motion field computation, required by both methods, is included in the timings. This required the warping and fusion of 8 images with the final size 256×256.

As shown in Table 6.2, for this example, our method is more than four times faster than our implementation of Irani's back-projection method. In general, Irani's back-projection method takes an amount

of time roughly proportional to both the number of iterations and the degradation model. Our experiments show that although each iteration of Irani's back-projection method takes approximately 65% of the running time of our method, the algorithm performs a minimum of two iterations. Thus, even in its best case, Irani's back-projection method is about 30% slower than our method. Our experiments also suggest that more than three iterations are often required to minimize the sum-of-square difference implying that the direct warping approach is often more than 100% to 200% faster than Irani's back-projection method.

5. Quantitative Evaluation

In this section, we turn our discussion to the quantitative measurement of super-resolution. Historically, as new super-resolution techniques have been developed, the standard practice has been to present a few examples of the technique to allow the reader to reach their own qualitative conclusions. It is difficult to justify comparisons between super-resolution algorithms that make different fundamental assumptions — for instance regular sub-pixel shifts vs. aperture changes vs. object motion. However, in order to make progress, we need to be able to quantatively measure the improvements of algorithms.

Some seemingly natural measures would be a blind measure of image quality; for instance, some measure of difference between a high-resolution image and the recovered super-resolution image or some type of spectral analysis seeing how well high-frequencies are recovered. Blind image-quality metrics are, however, fraught with problems as an overall measure of super-resolution algorithms because they are inherently task independent and disregard the underlying signal. Image differences from ground truth have been used in many areas as a measure of quality for comparison, but it remains difficult to decide how the differences should be weighted. In spectral analysis, we can look at how well the resulting super-resolution spectrum compares with the original. In section 5.1 we briefly review the quantitative and qualitative spectral analysis from [Chiang, 1998, Chapter 4].

While we have explored image differences and spectral measures, it is difficult to reduce them to a simple quantitative measure to allow comparison of two algorithms. The major difficulty with either difference or spectral analysis is how to combine the different spatial or spectral differences to a comparison metric. Simple approaches such as RMS of the difference is not very meaningful, just as RMS error is not a very good measure of image quality for compression. Also note that for super-resolution magnifications of more than double, the original

images contain frequencies so far above the Nyquist rate of the small images that the reconstruction techniques have no hope of recovering them. The intensity variations measured in these regions are a mixture of blurring and unmodeled aliasing. Neither super-resolution nor any resampling/reconstruction algorithm can recover all the lost information. While we are aware of these differences (their existence cannot be avoided), their significance is unknown. Rather than attempting to define what is important in some generic image sense, we believe that task oriented measures are more appropriate.

We present two realistic problems for which super-resolution is a natural augmenting tool. Since the metrics are computed using commercial products, their implementation and ease by which the evaluation may be reproduced are straightforward. The first problem, optical character recognition, or OCR, is considered in Section 5.2. The second, face-based human identification, is presented in Section 6. In both domains, image detail is critical to the systems' final performance.

In our OCR experiment, input is obtained from a hand-held camera, as opposed to the more traditionally used flat-bed scanner. Limited by NTSC resolution, we will show not only that super-resolution can significantly increase the recognition rate, but also the importance of warp quality. The experiment, described in section 5.2 and [Chiang and Boult, 2000], however, has drawbacks. The quantitative analysis used only a small number of samples. In addition, binary nature of the input may allow over-enhancements to cause increased recognition rates. Because of these limitations, we sought an additional domain.

Another approach, based on appearance matching and pose estimation, can be found in [Chiang, 1998, Chapter 7]. That analysis used grayscale images captured at two different resolutions and compared the results of running SLAM [Nene et al., 1994], an appearance-based recognition & pose algorithm, over various super-resolution images. The results were, in general, consistent with the OCR problem, except that there were instances where blurring the image actually increased accuracy of the recognition & pose estimation. In these unexpected cases, super-resolution did not help. Again, the sample size was small, and the targets with significant numbers of high-contrast edges may have dominated the results. Finally, in retrospect, the problem of pose computation from low resolution images was slightly artificial, and hence is not presented here.

In the case of face recognition we are addressing a very real problem — the recognition or verification of a human's identity from low-resolution facial images. In the human identification problem, it is common for a wide-field of view camera to be used and for subjects to be at varying

distances. Increasing the working range of existing systems is an ongoing research topic, and super-resolution is one potential way of achieving this. We synthetically "project" the images to produce the low resolution data. For true super-resolution, we would need multiple views of hundreds of heads and a robust facial matching techniques (since the head may rotate between frames). An experiment using multiple camera resolutions and facial matching is being planned. In this experiment, there is a large enough data space to instill confidence in the results. This allows us to quantitatively compare super-resolution using bilinear resampling with super-resolution using QRW and have statistical confidence in the hyphothesis that improved image warping (QWR) may improve super-resolution results.

5.1. Spectral Analysis of Super-Resolution from Synthetically Down-Sampled Images

For a spectral comparison, a sequence of five synthetically down-sampled images were used. The original high-resolution image provides the necessary ground truth for comparing the super-resolution results.

Fig. 6.7 shows the experimental results. The test images are generated from the high-resolution image shown in Fig. 6.7 by translation followed by down sampling. Fig. 6.7 (a) shows the original high-resolution image; Fig. 6.7 (b) the scale-up of the down-sampled version of Fig. 6.7 (a) by a factor of 4 using bilinear resampling (no super-resolution); Fig. 6.7 (c) the super-resolution result from the synthetically down-sampled image sequence using bilinear resampling with deblurring; Fig. 6.7 (d) the super-resolution result from the synthetically down-sampled image sequence using QRW with deblurring.

Table 6.3 shows the powers of the Fourier transform of the images shown in Fig. 6.7. For summary analysis, the images are divided into regions based on their relation to the sampling rate of the low-resolution (64 × 64) image. The regions are shown graphically in Table 6.3a, with the power within them shown in Table 6.3b. The column marked P_3 shows the power of the whole region (i.e., the whole spectrum). The columns marked P_2, P_1, and P_0 show, respectively, the power of the central 192 × 192 region, the power of the central 128 × 128 region, and the power of the central 64 × 64 region of the image spectrum. The column marked P_2', P_1', and P_0' show, respectively, the power of the whole region minus the power of the central 192 × 192 region, the power of the whole region minus the power of the central 128 × 128 region, and the power of the whole region minus the power of the central 64 × 64 region. As is to be expected, most of the power concentrates in the

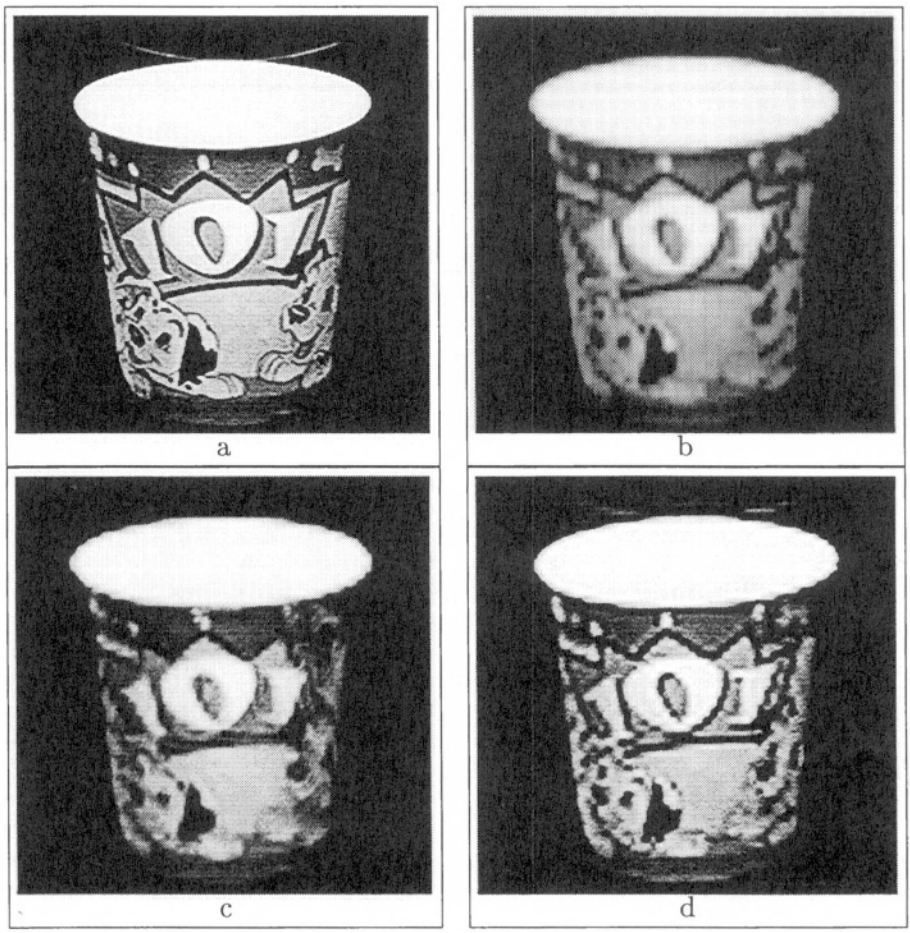
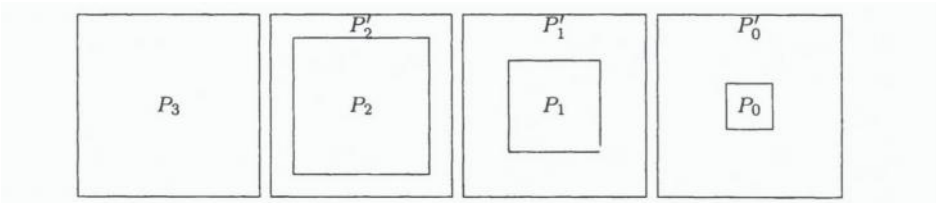


Figure 6.7. Results from a sequence of five synthetically down-sampled images. (a) the original image; (b) the scale-up of the down-sampled version of (a) by a factor of four using bilinear resampling (no super-resolution); (c)–(d) super-resolution from the synthetically down-sampled image sequence using, respectively, bilinear resampling with deblurring and QRW with deblurring.

central 64×64 region which are the frequencies that can be directly represented in the low-resolution images. Outside this region, the power is relatively small. Obviously, super-resolution using QRW with deblurring does considerably better.

Figs. 6.8 shows the difference between a slice of the 2D Fourier spectra (scanline 128 i.e. DC component in y) of the super-resolution reconstructions and the ground truth high-resolution image shown in Fig. 6.7a. Clearly this spectral analysis example shows the superiority, both quantitatively and qualitatively, of QRW warping over simple bilinear warp-



a

Method	P_3	P_2	P_1	P_0	P_2'	P_1'	P_0'
Original	11058.09	11052.15	11036.98	10928.85	5.93	21.11	129.24
Bi-linear	10722.40	10722.14	10721.59	10714.03	0.25	0.80	8.37
SRBD	10712.98	10712.14	10711.22	10698.52	0.83	1.76	14.46
SRQRWD	11060.74	11059.61	11058.35	11030.00	1.13	2.39	30.74

b

Table 6.3. Power of the Fourier transform of the images shown in Fig. 6.7. (a) Regions in the computation of the powers of the 2D Fourier transform. (b) the power within those regions with P_3 being the whole region (i.e., the whole spectrum); P_2, P_1 and P_0 being the regions inside the inner squares (192x192, 128x128, and 64x64, respectively); Primed labels indicated the complementary area of a region. Recall the original images were 64x64, so P_0 is associated with the power representable in those images, and P_0' what was gained by processing. Bi-linear is simple warping with bi-linear without super-resolution or debluring (i.e reconstruction of a single input image). SRBD is super-resolution using bi-linear warping followed by debluring, and SRQRWD is super-resolution using QRW followed by debluring.

ing or bilinear super-resolution. However, the quantitative analysis was very crude as it was based only on the power in the various regions. The next two sections show the superiority of QRW using task-based quantitative metrics.

5.2. OCR-based evaluation

In this section, we discuss our evaluation using OCR, which we refer to as *OCR-based measurement.* The fundamental idea of this approach is to use OCR as our fundamental metric (as the name suggests). The evaluation consists of three basic steps:

1 Obtain the super-resolution images using the super-resolution algorithm described in Section 4.

2 Pass the super-resolution results obtained in the previous step through a "character-oriented" OCR system.

3 Determine the number of characters recognized, i.e., the rate of recognition.

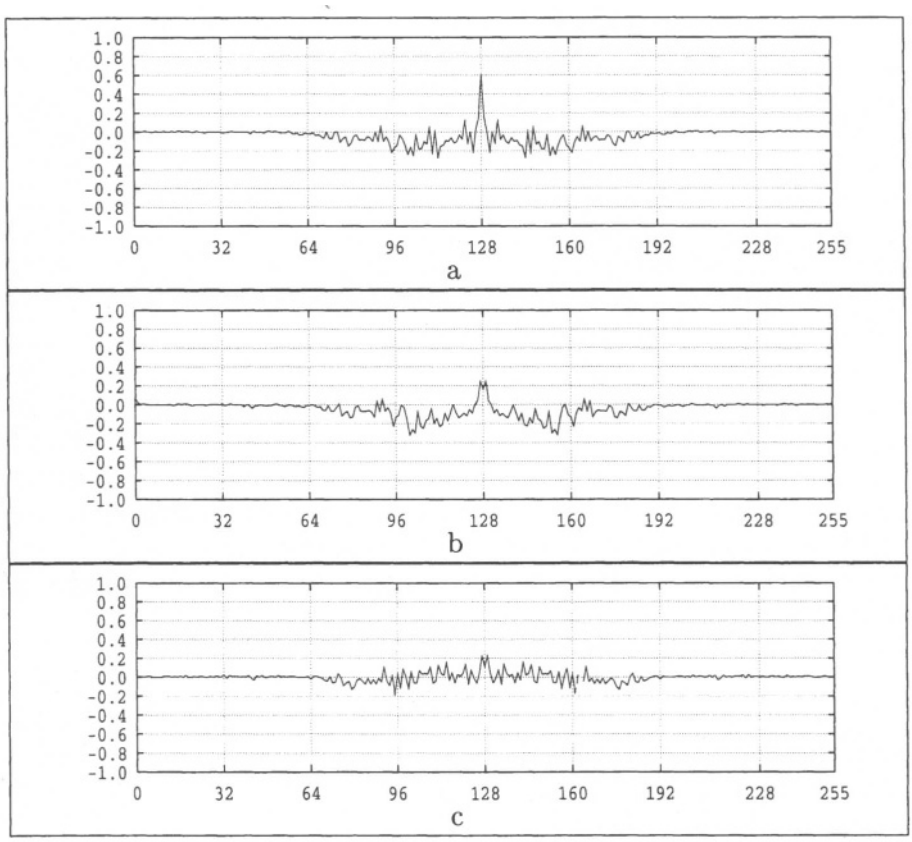


Figure 6.8. Display of $|F^S_{128}(u)| - |F^O_{128}(u)|$ where $|F^S_{128}(u)|$ and $|F^O_{128}(u)|$ are the Fourier spectra in x, sampled at the the DC component (row 128) in y) for the various super resolution approximations and the original image Fig. 6.7. (a) shows the difference using bilinear resampling (no super-resolution) and the original; (b) shows the difference between super-resolution using bilinear resampling with deblurring and the original and (c) showing the difference between super-resolution using QRW with deblurring and the original.

The goals of this evaluation is to quantify the effectiveness of super-resolution. Evaluations herein are made by comparing the super resolution results and those using bilinear warping.

While most 'state-of-the-art' OCR programs can use dictionary lookup to aid in their recognition, we chose to use a pure character based system. This ensures that the system's behavior is driven only by the image input and not significantly impacted by the grammatical context of examples. We have also chosen to use a font-independent system, i.e., one that is not trained on the font and on the resolution being used. Training the OCR system might allow the training to compensate for poor, but consistent, behavior in resampling or super-resolution.

Since the OCR program is not trained on a particular font, we break up our analysis of errors into two categories. The first error measure compares the OCR output to the ground-truth input characters. We use C_r^i to indicate the number of characters correctly recognized. We consider multiple types of errors, including C_i, C_m, C_e, and C_s which give, respectively, the number of incorrectly recognized characters, the number of missing characters, the number of extra characters, and the number of characters that are split into two or more characters, with $C_{i+m+e+s}$ being the sum of these. %Correct indicates the percentage of characters correctly recognized (C_r^i divided by $C_{i+m+e+s}$). For many fonts, some characters are so visually similar that without the use of training or context, distinguishing pairs of characters is quite difficult, e.g., `0 vs O, 1 vs l vs ! vs |`, and in some fonts, / vs *l* vs *t* and *h* vs *b* (see Fig. 6.13). In the context of our experiments, Fig. 6.9 contains three ambiguous characters, Fig. 6.11 contains four ambiguous characters and Fig. 6.13 contains seventeen ambiguous characters.

For brevity, we also use the abbreviations as shown in Table 6.4 to describe the algorithms discussed.

Abbrv.	Meaning
BRX	bilinear resampling without distortion correction and deblurring
BRDCD	bilinear warping with distortion correction and deblurring
BR	bilinear resampling without deblurring
BRD	bilinear resampling with deblurring
SR	super-resolution using QRW without deblurring
SRD	super-resolution using QRW with deblurring
SRDCD	super-resolution using QRW with distortion correction and deblurring

Table 6.4. Abbreviations of algorithms considered. Distortion correction is needed to remove radial lens distortions common in inexpensive lenses.

The OCR program used for the experiments described herein is "Direct for Logitech, Version 1.3." The images used are normalized with respect to the super-resolution image with deblurring, as follows:

$$I_n = \frac{\bar{I}_s}{\bar{I}_u} I_u$$

where I_n is the normalized image, $\bar{I}_s$ is the average of the intensity values of the super-resolution image with deblurring, and $\bar{I}_u$ is the average of the intensity values of the image to be normalized. Within each dataset, the same threshold is used to binarize all the images.

The test data shown in this section was taken using laboratory quality imaging systems, a Sony XC-77 camera, attached to either a Datacube MV200 System or a Matrox Meteor Capture Card. As is to be expected, better imaging reduces the need for super-resolution; lower quality cameras increase the significance of super-resolution imaging.

All examples herein are scaled up by a factor of four, with the distance between camera and sample being changed so that the scaled images would yield an image with character sizes within the range accepted by the OCR system. We qualitatively evaluated the approach on a wider range of fonts and imaging conditions. Note that fonts with thinned letters, such as the "v" in Fig. 6.11, tend to be broken into multiple letters. Characters in slanted serif fonts tend to connect and thus, fail to be recognized. Inter-word spacing is not handled well (and multiple spaces are ignored in our measures). The ease of finding a good threshold depends on the uniformity of the lighting, image contrast, and lens quality. Better OCR algorithms may remove most or all of these difficulties. The quantitative examples show a few of these features, but in general, we choose examples that are not dominated by these artifacts.

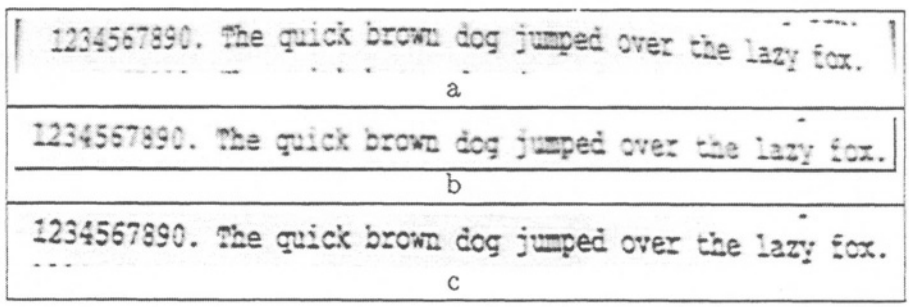


Figure 6.9. Super-resolution results from a sequence of 32 391x19 images taken by a Sony XC-77 Camera. (a) one of the original images scaled up using bilinear resampling and without distortion correction; (b) and (c) the results after distortion correction, with (b) showing bilinear warping with deblurring and (c) showing super-resolution using QRW with deblurring.

Fig. 6.9 shows the super-resolution results from a sequence of 32 391 × 19 images taken by a Sony XC-77 camera attached to a Datacube MV200 System before they are passed through OCR for recognition. Fig. 6.9a shows one of the original images scaled up using bilinear resampling and without distortion correction. Fig. 6.9b and Fig. 6.9c show the results after distortion correction, with Figs. 6.9b showing bilinear warping with deblurring and Figs. 6.9c showing super-resolution using QRW with deblurring.

Method	Output of OCR
BRX	1234561990. te gu<clc bow dog jumped over te sazs
BRDCD	:23*;56?990. X:e quick brown dog jt:Wed over the Lazf for
SRDCD	2234567890. The quick brown dog jumped over the lazes 'ox.

Method	%Correct	C_r^1	C_i	C_m	C_e	C_s	$C_{i+m+e+s}$
BRX	67	32	7	8	0	1	16
BRDCD	71	35	11	1	0	1	14
SRDCD	94	45	2	0	0	1	3

Figure 6.10. Output of OCR for the first example, the text shown in Fig. 6.9. The smaller size of text and more significant distortions make the impact of super-resolution using QRW very dramatic.

Figs. 6.10 summarizes the results of passing the super-resolution results shown in Fig. 6.9 through OCRThis example did not contain any font-related ambiguities. The original text (see Fig. 6.9) consists of total 48 characters, including the two periods but excluding whitespace. Columns marked C_i, C_m, C_e, and C_s give, respectively, the number of incorrectly recognized characters, the number of missing characters, the number of extra characters, and the number of characters that are split into two or more characters.

Because of the nonuniformity of the lighting in this example, each image had its own threshold which was chosen to maximize its recognition rate. Using bilinear resampling without distortion correction, and deblurring (BRX), 32 out of the 48 characters (67%) were recognized. Using bilinear warping with distortion correction and deblurring (BRDCD), 35 out of the 49 characters (71%) were recognized. Using the super-resolution algorithm given in Section 4 with deblurring (SRDCD), 45 out of the 48 characters (94%) are recognized. Compared to bilinear resampling without distortion correction, super-resolution using QRW recognizes 27% more characters. Compared to bilinear with distortion correction and deblurring, super-resolution using QRW recognizes 21% more of characters. With text consisting of thousands of characters, this is definitely a significant improvement.

Qualitatively, one might note the errors are concentrated on the outer edges of the example where there was the most significant distortions and the worst lighting.

Fig. 6.11 shows the super-resolution results for an example with larger characters (almost twice the size) taken with a better lens (much less distortion and less blur). The input was a sequence of 8 430 × 75 images taken by a Sony XC-77 camera attached to a Matrox Meteor Capture

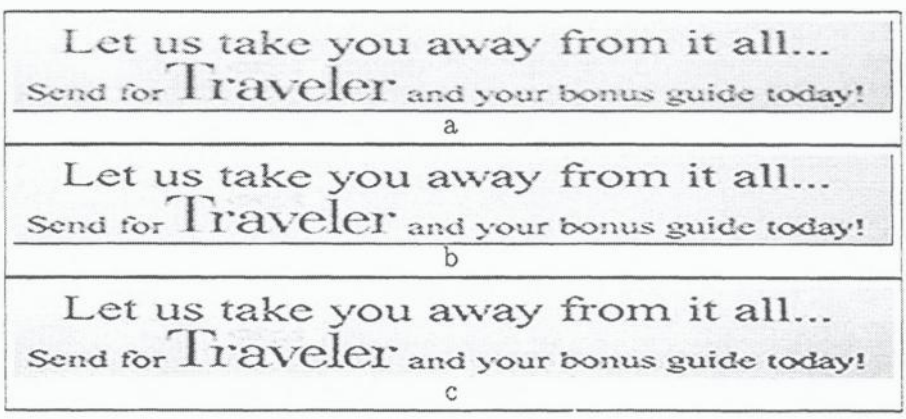


Figure 6.11. Super-resolution results from a sequence of 8 430×75 images taken by XC-77. (a) one of the original images scaled up using bilinear warping; (b) (a) deblurred; (c) super-resolution using QRW with deblurring.

Card before they are passed through OCR for recognition. The original text consists of a total of 66 characters including three periods and one exclamation mark.

Method	Output of OCR	C_i	C_m	C_e	C_s
BR	Let us take you assay from it all...	0	0	0	1
	Send fior TI aveler and your bonus gtude todyl	1	2	2	1
BRD	Let us take you assay from it all...	0	0	0	1
	Send for Tl aveler and your bonus guide today!	0	0	0	1
SR	Let us take you away from it all...	0	0	0	0
	Send fior Tl alreler and your bonus guide today	0	1	1	2
SRD	Let us take you away from it all...	0	0	0	0
	Send for Tra@&eler and your bonus guide today!	0	0	0	1

Method	%Correct	C_r^i	C_i	C_m	C_e	C_s	$C_{i+m+e+s}$
BR	89.7	61	1	2	2	2	7
BRD	97.0	64	0	0	0	2	2
SR	94.0	63	0	1	1	2	4
SRD	98.5	65	0	0	0	1	1

Figure 6.12. Results of the OCR test for the second example, shown in Fig. 6.11. Again debluring helped in both cases and super-resolution using QRW with debluring was the best algorithm.

Fig. 6.12 shows the experimental results of passing the super-resolution results shown in Fig. 6.11 through OCR. While the bilinear and QRW super-resolution (SRD) images look nearly identical, the quantitative

OCRanalysis shows a difference. Compared to bilinear resampling without deblurring (BR), 7% more of the characters are recognized with QRW. Compared to bilinear warping with deblurring (BRD), 2% more of the characters are recognized. 2% may mean a lot depending on applications. Compared to bilinear resampling without deblurring (BR), super-resolution with deblurring (SRD) reduces the number of incorrectly recognized, missing, extra, and split characters from 7 to 1.

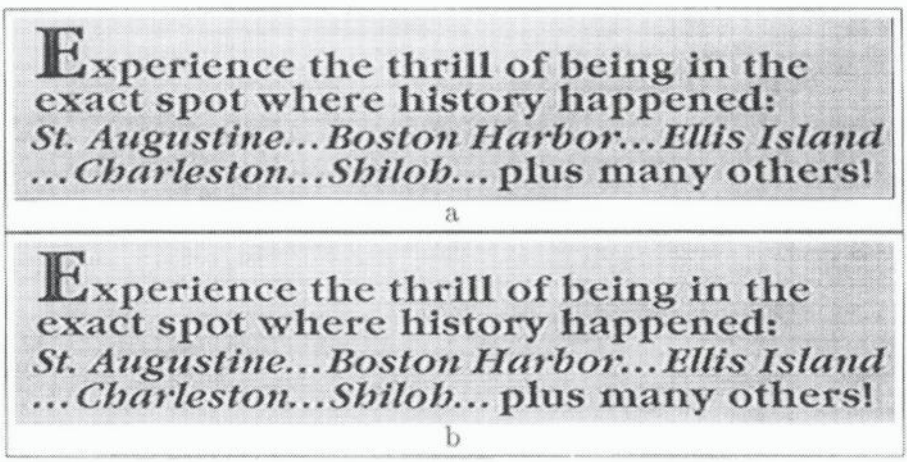


Figure 6.13. The third example sequence of 8 490×140 images taken by XC-77. The top (a)shows the results using bilinear warping with debluring and the bottom (b) shows the results of super-resolution using QRW with deblurring. While the images look nearly identical, the quantitative OCRanalysis shows a difference.

Fig. 6.14 shows the analysis for a third quantitative experiment. The original text, Fig. 6.13 consists of total 142 characters with mixed fonts and a large number of ambiguous characters. Compared to bilinear resampling without deblurring (BR), 5% more of the characters are recognized. Compared to bilinear resampling with deblurring (BRD), 3% more of the characters are recognized. Again, 3% could mean a lot depending on applications – if there were 2000 characters on a page, it is a difference of 60 characters. Discounting the ambiguous characters will increase the rate of recognition by 2% for all methods except bilinear (BR) which increases by only 1%. Compared to bilinear resampling with or without deblurring (BR or BRD), super-resolution with deblurring (SRD) reduces the number of incorrectly recognized, missing, extra, and split characters from 18 to 12. Looking at the details one can see that the italics was handled poorly. While a multi-font OCRsystem might do much better overall, the example does show that super-resolution improves the results.

Method	Output of OCR	C_i	C_m	C_e	C_s
BR	Experlence the thrill of being in the	3	0	0	0
	exact spot where history happened	0	1	0	0
	St. Augustfne...Boston Hahbor...gWs Istand	5	2	0	0
	...Cbarleston@..SbSlob... plus maay othersl	0	7	0	0
BRD	Experience the thrill of being in the	0	0	0	0
	exact spot where history happened:	0	0	0	0
	St. Augustfne. . .Boston Hahbor...E s Island	3	2	2	0
	... CbarZeston...Sbilob... plus Expeothersl	9	1	1	0
SR	Experience the thrill of being in the	2	0	0	0
	exact spot where history happened	0	1	0	0
	Stg Augustine...Boston Hahbor...SWs Island	4	2	0	0
	...Cbarlesto@t...Sbf/ob.. plus many othersl	6	1	0	1
SRD	Experience the thrill of being in the	0	0	0	0
	exact spot where history happened:	0	0	0	0
	Sf. Augustxne...Boston Harbor...Sis Istand	4	2	0	0
	...Cbar/eston...Sbi/oh.. plus many othersl	5	1	0	0

Method	%Correct	C_r^i	C_i	C_m	C_e	C_s	$C_{i+m+e+s}$
BR	87.3	124	8	10	0	0	18
BRD	87.7	128	12	3	3	0	18
SR	88.1	126	12	4	0	1	17
SRD	91.5	130	9	3	0	0	12

Figure 6.14. Performance of OCR testing for the third experiment. As you can see by looking at the recovery detail, the italics was poorly handled. Overall super-resolution using QRW and deburing performed the best.

5.3. OCR experiments summary

The qualitative aspects of our experimental results can be summarized as follows:

- Naturally, the better the quality of the original images and the larger the input characters, the smaller the impact of super resolution on OCR. But even for large clear text, it did have a measurable impact.
- If there is no motion, minimal warping and good large text, super-resolution will not help much more than simple temporal averaging with bilinear warping.
- Type style has a strong impact on the rate of recognition.

This section has shown how OCRcan be used used for super-resolution evaluation. The advantages of this approach is that it is very straight

forward; there are a large number of both commercial and free OCR packages, and data collection is also straightforward. The use of OCR is well suited for evaluation of super-resolution tasks which will be similar in nature to OCR, e.g., license plate recognition, 2D pattern matching, and handwritting recongition.

In general, the difficulties of using OCR as a measurement for super-resolution can be summarized as follows:

- The rate of recognition depends, to a large extent, on the quality of the OCR programs. Better OCRprograms, especially those that use dictionary lookup, would reduce the impact of low level processing. But in general, better low level processing would provide better results.
- If binary images are required for recognition, as most of the OCR programs do even if implicitly converted internally, then the rate of recognition is sensitive to the thresholding process used to convert gray-scale or color images to binary images. This evaluation used external thresholding and different thresholds may give different results. While localized thresholding would help increase the rate of recognition, we have not used them here.
- Many OCR programs treat their inputs as "binary". Thus as an evaluation for super-resolution techniques, it may seem to downplay the importance of accurate grayscale production, especially at middle intensity levels. On the other hand, these intermediate levels do occur on character boundaries and may, in fact, be the driving factor in the superiority of super-resolutions. However the nearly binary nature of the data may suggest that over enhancement might do better.

6. Face-Based evaluation

In this section, we evaluate how SR can be used to improve the performance of a face recognition (FR) system. A formal description of the FR problem is given first, followed by an evaluation of a simulated SR-enhanced FR sytem.

Face Recognition Systems

One view of an FR system is a facility that provides a mapping from facial images to labels that uniquely identify the subject of the image. We assume that given an FR system, there exists some image set of known subjects, also known as a *gallery*, which we denote as G. In addition, there exists some *probe set* $P = \{p_1, p_2, \ldots, p_{|P|}\}$ where each

Algorithm	Description
s_0	FaceIt *Image to Scan Template*
s_1	FaceIt *Image to Full Tempalte* (Normal mode)

Figure 6.15. **Comparison operators.** The precise comparison operators used in our evaluation and their names as described by the FaceIt documentation.

image, $p_i \in P$ and $p_i \notin G$, is an image of some subject the FR system has to recognize. In the system considered here, assume that the gallery G and the probe set P are non-identical, but have been created from the same population.

Most FR systems, presented with a gallery image g and probe image p_i, have the capability of computing some bounded similarity measure $s(g, p_i)$ representing the "strength" of the match between the images. Without loss of generality, assume that a score $s(g, p)$ of 10.0 indicates the highest possible system confidence that the subject in image g is the same as the subject in image p, and that a score $s(g, p)$ of 0.0 indicates the lowest possible system confidence that the subjects are the same (or highest confidence that they are different).

Let id(x) represent the true identity of the subject in image x, and g_h represent the gallery image of subject h. Given a probe p, a vector of similarity scores $s(g, p)$ can be calculated from all images $g \in G$. Sorting the similarity vector and finding the subject's relative position along it, determines the probe's *rank*. Specifically, a probe has a rank of n over gallery set G if in the similarly vector, there exist exactly n scores greater than or equal to $s(g_{\mathrm{id}(p)}, p)$.

The algorithm used in the evaluation was Visionics' FaceIt. FaceIt is a commercial, high-quality, face detection and recognition system based on Local Feature Analysis. FaceIt requires no explicit "training" stage; i.e. no internal component of the system incorporates information across the entire gallery. This is different from many linear subspace classifiers (such as Eigenface approaches) which must be trained with respect to the entire gallery. With FaceIt, adding an image to the gallery does not require a retraining, and has no side effects on the probe to gallery image comparison. In other words, given a particular gallery and probe image g and p, the addition of a new gallery image g' has no effect on the FaceIt similarity measure $s(g, p)$. Naturally, the rank could be effected. For our evaluation, we selected two different similarity measures, described in Figure 6.15.

Experimentation

The context of the evaluation consisted of a subset of the Essex database packaged with FaceIt. Due to constraints in the available data, we selected 352 image pairs. Each pair consists of two images of the same subject with approximately the same pose, but slightly different expression. These images, in FERET nomenclature, are referred to as *A* and *B* images. All subjects selected were front facing, imaged in front of simple (essentially monochromatic) backgrounds, and diffusely illuminated. Since we were more interested in the effects of the image preprocessing stages, we presented favorable data to the system.

More formally, let $q^i = \{q_{(i,A)}, \ldots, q_{(i,D)}\}$ represent the set of four images of subject *i* and $Q^c = \{q_{(1,c)}, \ldots, q_{(L,c)}\}$ represent the set of all *c* images — in our case, $c \in \{A, B, C, D\}$. Let $f(q)$ represent some image processing function on image *q* (this will soon be replaced with a super-resolution algorithm), *s* represent a similarly measure, and $\mathbf{rank}_s(G, p)$ represent the rank of probe *p* over gallery *G* via similarity measure *s*. Then, using *G*, *P*, *s*, and *f*, an evaluation that obtains a set of ranks (one set per probe image) can be denoted as

$$eval(G, P, s, f) = \bigcup_{p \in P} rank_s(G, f(p)). \tag{6.7}$$

In order that confidence intervals could be obtained, an evaluation framework based on population stratification and methods of replicate statistics (*balanced repeated replicates* or BRR, specifically) was used. Letting each subject correspond to a stratum, three samples per stratum (or PSUs) are obtained by fixing one set of images as the gallery and probing the gallery with the remaining data sets. That is, one set of samples was obtained with $eval(Q^A, Q^B, s, f)$, another with $eval(Q^A, Q^C, s, f)$ and so on.

Dataset Generation In this section we describe the probe and gallery sets used in our evaluation. The notation used here will also be used to describe the results of the evaluation.

A series of low-resolution images to serve as input to the SR algorithm is generated first. Based on previous results [Chiang, 1998], it was decided that four low-resolution input images would be used. To simluate a low-resoultion face from slightly different views, a perspective projection was used. Let *m* represent a random, but known, perspective and scalar pixel-to-pixel mapping, where the image width and height are reduced to 25 percent of their original size. In this evaluation, the perspective projection was limited so that a pixel's horizontal coordinate would be displaced by at most 10% of the original image width. Let *M* represent

the inverse operation — a 4 times dilation and perspective "correction." Since four low-resolution images were generated, four such m mappings needed. A set of four (distinct) mappings generated from a random seed k is denoted as $rvec(m)_k = \{m_k^1, m_k^2, m_k^3, m_k^4\}$. It follows that the set of complimentary mappings is denoted as $rvec(M)_k$.

Given a map m, image p, and warping algorithm a, a new low-resolution image p' can be denoted as a function of m and p, or $p' = a(m, p)$. Let r and b represent QRR and bilinear warping algorithms respectively. Then, for each probe image $p_i \in P$, a set of four low-resolution images generated from warping algorithm a is

$$P_i^a = \bigcup_{m_i^k \in rvec(m)_i} a(m_i^k, p_i) = \{a(m_k^1, p_i), \ldots, a(m_i^4, p_i)\}. \tag{6.8}$$

If we denote a SR algorithm using warp a as SR_a, then we may modify Equation (6.7), our previous definition of *eval*, to reflect SR_a as

$$eval(G, P, s, a) = \bigcup_{p \in P} rank_s(G, SR_a(P_i^a)). \tag{6.9}$$

In the ideal case, the SR algorithm generates the corrective maps by directly inverting each distortion m. Naturally, this is only possible when m is known a priori. Note that other phenomena, such as sensor noise, deformations in expression, and changes in illumination, are not incorporated into the evaluation. By using ideal conditions, such as nearly ideal inverse mappings, our evaluation is not only significantly simplified, but better reflects a more "upper bound"-like measure of the effects of SR.

As reflected in Equation (6.9), changing a SR algorithm has no effect on the particular maps (m and M) used. This critical constancy ensures that differences in the resulting ranking can be attributed only to the change in warping algorithm.

6.1. Experimentation and Results

To establish a peformance gain (or loss) for SR, the raw low-resolution images will be used as the baseline evaluation. In a system without SR, these raw images are what would be used as probe images. The ideal SR images give an indication of the "best" possible performance gain due to SR.

Unfortunately, *rank*, as defined previously, can be a non-robust measure. The penalty incurred by a misdetection (defined as a rank greater than t, where t is some predefined threshold) is linearly related. For example, suppose a probe has a very high (poor) rank — 400, for instance.

Figure 6.16. **Low-resolution images of Subject 202.** Example of a set of four, perspective projection distored, low resolution images which serve as SR algorithm input. All images were generated from the higher-resolution *A* image of subject 202. Closer inspection reveals subtle difference between the images: the first face appears narrower and slighly elongated with respect to the other three faces, the fourth appears slighly smaller, lower, and wide than the other three, etc. The black stripes on the left and right side of the images are artifacts resulting from the projection. Because each subject has their own set of unique maps, some low-resolution image sets will show more or less variation between images.

The penalty incurred in the mean rank due to this 400-rank misdetect is much greater than that due to a probe with a rank of, say 50. In both cases, however, if our threshold $t = 10$, then they are both clearly misdetects. Therefore, for our evaluation, we perform the BRR mean and variance over the following statistic, where r represents a probe's rank and α is some threshold rank

$$\theta_n = 1 \text{ if } r < \alpha, \quad \theta_n = 0 \text{ otherwise.} \tag{6.10}$$

Another view of θ_n is the expected value of the fraction of probes within the top α match candidates. For this particular evaluation, we used *alpha* = 0.01. In other words, a probe scored a 1 if its subject was within the top α of candidate images (just the top image in our case).

The generation of multiple low-resolution images provides a broad baseline. Let $Q^c_{(a,i)}$ represent the set of all ith low-resolution images generated from Essex databaset set Q^c using warping algorithm a. In other words, the set $Q^D_{(b,3)}$, or

$$\{b(m_1^3, q_{(1,D)}), b(m_1^3, q_{(2,D)}), \ldots, b(m_1^3, q_{(L,D)})\} \tag{6.11}$$

is the set of all low-resolution images from set Q^C generated from a map of index 3. Note that this partitioning is somewhat arbitrary, and only dependent on particular indexes. This notation will be used again shortly. Similarly, we let

$$Q^A_{(r,\mathrm{SR})}, \ldots, Q^D_{(r,\mathrm{SR})}, Q^A_{(b,\mathrm{SR})}, \ldots . Q^D_{(b,\mathrm{SR})}. \tag{6.12}$$

denote the super-resolution image sets generated from their respective low-resolution images. Finally, the BRR mean estimates are generated from each $(Q^A_{(a,i)}, \ldots, Q^D_{(a,i)})$ set. These are denoted by replacing the set index with the symbol μ. For example

$$Q^{\mu}_{(b,1)}, Q^{\mu}_{(r,1)}, \ldots, Q^{\mu}_{(b,4)}, Q^{\mu}_{(r,4)}, Q^{\mu}_{(b,\mathrm{SR})}, Q^{\mu}_{(r,\mathrm{SR})} \tag{6.13}$$

These BRR estimators incorporate rank information across all *A*, …, *D* sets (appropriately).

The raw results of the evaluation are shown in Figure 6.17. For each row, the table shows the expected value of the fraction of probes that produce ranks of 0. The BRR estimated standard error of these means is shown in parenthesis. As shown in the figure, both QRW and bilinear based super-resolution improves the fraction of recognized faces with a statistical significance.[1] For all experiments, QRW super-resolution produced better fractions then bilinear. This is not always the case for the component fractions. This indicates that in certain cases, it is possible that a particular low-resolution image may be better than a super-resolution image. In a real face recognition system, however, ground truth is not available. Therefore, it would be impossible to know *which* low-resolution produces the correct result. Nevertheless, this phenomenon is more of a face-recognition system issue than a super-resolution issue. It should be noted that increasing α (for $\alpha < 10$) does not dramatically change the fact that QRW produces statistically significant higher fractions.

Figure 6.17 shows the results of both the FaceIt algorithms. In the first algorithm, the overall peformance is much better and we see that super-resoltion helps for both bilinear warping and for QRW. It also shows that QRW is statistically superior, being at least three standard deviations above bilnear-based super-resolution. The second algorithm, overall, did not perform as well. The behavior of QRW is consistent — again showing a statistically significant improvement for super-resolution over the individual low-resolution inputs. However, this algorithm had cases where bilinear outperformed QRW on the individual examples. However, this example shows that unlike QRW, with bi-linear warping, super-resolution was not better than the individual inputs.

7. Conclusion

This chapter discussed techniques for image consistent reconstruction and warping using the integrating resampler. By coupling the degrada-

[1]This is a statistically sound statement which is dependent on the unique properties of BRR.

probe sets	alg s_0	alg s_1
$Q^{\mu}_{(b,1)}$	0.8277 (0.0083)	0.6591 (0.0114)
$Q^{\mu}_{(r,1)}$	0.8390 (0.0087)	0.6382 (0.0119)
$Q^{\mu}_{(b,2)}$	0.8438 (0.0085)	0.6591 (0.0112)
$Q^{\mu}_{(r,2)}$	0.8570 (0.0086)	0.6705 (0.0125)
$Q^{\mu}_{(b,3)}$	0.8219 (0.0082)	0.6609 (0.0113)
$Q^{\mu}_{(r,3)}$	0.8589 (0.0081)	0.6430 (0.0115)
$Q^{\mu}_{(b,4)}$	0.8286 (0.0088)	0.6866 (0.0112)
$Q^{\mu}_{(r,4)}$	0.8589 (0.0081)	0.6335 (0.0120)
$Q^{\mu}_{(b,\mathrm{SR})}$	0.8494 (0.0079)	0.6647 (0.0115)
$Q^{\mu}_{(r,\mathrm{SR})}$	0.8731 (0.0078)	0.6875 (0.0109)

Figure 6.17. **Mean Rank Estimates**. The results of the evaluation, showing the expected value of the fraction of the probes that have rank 0 (most likely candidate). Standard devations are shown in parenthesis.

tion model of the imaging system directly into the integrating resampler, we can better approximate the reconstructed image and the warping characteristics of real systems. This, in turn, significantly improves the quality of super-resolution images. Examples of super-resolutions for gray-scale images show the usefulness of the integrating resampler in applications scaling by a factor of upto 4 using 8-32 images. We disussed three quantitative evaluation approaches and in each case, we saw that super-resolution using QRW was superior to bilinear approaches. Even in those cases where the super-resolution images were visually similar, we had measurable quantitative improvements.

References

[Andrews and Hunt, 1977] Andrews, H. and Hunt, B. (1977). *Digital Image Restoration.* Prentice-Hall, Englewood Cliffs, NY.

[Andrews and Patterson, 1977] Andrews, H. and Patterson, C. (1977). Digital interpolation of discrete images. *IEEE Trans. Computers*, C-25:196–202.

[Bascle et al., 1996] Bascle, B., Blake, A., and Zisserman, A. (1996). Motion deblurring and super-resolution from an image sequence. *Computer Vision—ECCV*, pages 573–581.

[Boult and Wolberg, 1992] Boult, T. and Wolberg, G. (1992). Correcting chromatic abberations using image warping. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.*

[Boult and Wolberg, 1993] Boult, T. and Wolberg, G. (1993). Local image reconstruction and subpixel restoration algorithms. *CVGIP: Graphical Models and Image Processing*, 55(l):63–77.

[Chiang, 1998] Chiang, M. (1998). *Imaging-Consistent Warping and Super-Resolution.* PhD thesis, Department of Computer Science, Columbia University.

[Chiang and Boult, 1996] Chiang, M. and Boult, T. (1996). The integrating resampler and efficient image warping. *Proceedings of the DARPA Image Understanding Workshop*, pages 843–849.

[Chiang and Boult, 2000] Chiang, M. and Boult, T. (2000). Super-resolution via imaging warping. *Image & Vision Computing Journal Special Issue.*

[Drago and Granger, 1985] Drago, F. and Granger, E. (1985). Optics requirements for making color fiche transparencies of maps, charts, and documents. *Image Quality: An Overview Proc. SPIE*, 549:25–33.

[Fant, 1986] Fant, K. (1986). A nonaliasing, real-time spatial transform technique. *IEEE Computer Graphics and Applications*, 6(1):71–80. See also "Letters to the Editor" in Vol.6 No.3, pp. 66-67, Mar 1986 and Vol.6 No.7, pp. 3-8, July 1986.

[Gonzalez and Wintz, 1987] Gonzalez, R. and Wintz, P. (1987). *Digital Image Processing.* Addison-Wesley, Reading, MA.

[Granger, 1974] Granger, E. (1974). Subjective assessment and specification of color image quality. *SPIE Proc. Image Assessment and Specification*, page 86.

[Gross, 1986] Gross, D. (1986). Super-resolution from sub-pixel shifted pictures. Master's thesis, Tel-Aviv University.

[Hou and Andrews, 1987] Hou, H. and Andrews, H. (1987). Cubic splines for image interpolation and digital filtering. *IEEE Trans. Acoust. Speech, Signal Process.*, ASSP-26:508–517.

[Huck et al., 1991] Huck, F., Alter-Gartenberg, R., and Z.-U. Rahman (1991). Image gathering and digital restoration for fidelity and visual quality. *CVGIP: Graphical Models and Image Processing*, 53:71–83.

[Irani and Peleg, 1991] Irani, M. and Peleg, S. (1991). Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 53(3):231–239.

[Irani and Peleg, 1993] Irani, M. and Peleg, S. (1993). Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335.

[Jain, 1989] Jain, A. (1989). *Fundamentals of Digital Image Processing.* Prentice-Hall, Englewood Cliffs, NY.

[Keren et al., 1988] Keren, D., Peleg, S., and Brada, R. (1988). Image sequence enhancement using sub-pixel displacements. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 742–746.

[Keys, 1981] Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech, Signal Process.*, ASSP-29:1153–1160.

[Maeland, 1988] Maeland, E. (1988). On the comparison of interpolation methods. *IEEE Trans. Medical Imaging*, MI-7(3):213–217.

[Mitchell and Netravali, 1988] Mitchell, D. and Netravali, A. (1988). Reconstruction filters in computer graphics. *Computer Graphics (SIGGRAPH '88 Proceedings)*, 22(4):221–228.

[Nene et al., 1994] Nene, S., Nayar, S., and Murase, H. (1994). SLAM: Software Library for Appearance Matching. Tech. Rep. CUCS-019-94, Columbia University, Department of Computer Science.

[Oakley and Cunningham, 1990] Oakley, J. and Cunningham, M. (1990). A function space model for digital image sampling and its application in image reconstruction. *CVGIP: Graphical Models and Image Processing*, 49:171–197.

[Park and Schowengerdt, 1982] Park, S. and Schowengerdt, R. (1982). Image sampling, reconstruction, and the effect of sample-scene phasing. *Applied Optics*, 21(17):3142–3151.

[Park and Schowengerdt, 1983] Park, S. and Schowengerdt, R. (1983). Image reconstruction by parametric cubic convolution. *CVGIP: Graphical Models and Image Processing*, 23:258–272.

[Parker and and D.E. Troxel, 1983] Parker, J. and and D.E. Troxel, R. K. (1983). Comparison of interpolating methods for image resampling. *IEEE Trans. Medical Imaging*, MI-2(l):31–39.

[Pavlidis, 1982] Pavlidis, T. (1982). *Algorithms for Graphics and Image Procesing.* Computer Science Press, Rockville, MD.

[Peleg et al., 1987] Peleg, S., Keren, D., and Schweitzer, L. (1987). Improve image resolution using subpixel motion. *Pattern Recognition Letter*, pages 223–226.

[Pratt, 1978] Pratt, W. (1978). *Digital Image Processing.* John Wiley & Sons, New York, NY.

[Pratt, 1990] Pratt, W. (1990). *Digital Image Processing.* John Wiley & Sons, New York, NY.

[Reichenbach and Park, 1989] Reichenbach, S. and Park, S. (1989). Two-parameter cubic convolution for image reconstruction. *Proc. SPIE Visual Communications and Image Processing*, 1199:833–840.

[Simon, 1975] Simon, K. (1975). Digital image reconstruction and resampling for geometric manipulation. *Proc. IEEE Symp. on Machine Processing of Remotely Sensed Data*, pages 3A–1–3A–11.

[Traub et al., 1988] Traub, J., Wasilkowski, G., and Wozniakowski, H. (1988). *Information-Based Complexity*. Academic Press, New York.

[Wolberg, 1990] Wolberg, G. (1990). *Digital Image Warping*. IEEE Computer Society Press, Los Alamitos, California.

[Wolberg and Boult, 1989] Wolberg, G. and Boult, T. (1989). Image warping with spatial lookup tables. *Computer Graphics (SIGGRAPH '89 Proceedings)*, 23(3):369–378.

This page intentionally left blank

Chapter 7

RESOLUTION ENHANCEMENT USING MULTIPLE APERTURES

Takashi Komatsu
Department of Electrical Engineering, Kanagawa University
3-27-1 Rokkakvbashi.Kanagwa-ku, Yokohama, 221-8686, JAPAN
komatt01@kanagawa-u.ac.jp

Kiyoharu Aizawa
Department of Electrical Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
aizawa@hal.t.u-tokyo.ac.jp

Takahiro Saito
Department of Electrical Engineering, Kanagawa University
3-27-1 Rokkakubashi, Kanagwa-ku, Yokohama, 221-8686, JAPAN
saitot01@kanagawa-u.ac.jp

Abstract This chapter develops very high definition image acquisition system which is based on the signal-processing approach with multiple cameras. The approach produces an improved resolution image with sufficiently high signal-to-noise ratio by processing and integrating multiple images taken simultaneously with multiple cameras. Originally, in this approach, we used multiple cameras with the same pixel aperture, but in this case there needs to be severe limitations both in the arrangement of multiple cameras and in the configuration of the scene, in order to guarantee the spatial uniformity of the resultant resolution. To overcome this difficulty completely, this chapter presents the utilization of multiple cameras with different pixel apertures, and develops a new, alternately iterative signal-processing algorithm available in the different aperture case. Experimental simulations and results are also presented. These results show that the utilization of multiple different-aperture

cameras prospects well and that the alternately iterative algorithm behaves satisfactorily.

Keywords: CCD camera, shot noise, registration, image acquisition, high-resolution imaging

1. Introduction

Solid-state imaging device technology is considered promising as a high-resolution imaging device. Although a CCD camera with two million pixels has been developed for HDTV, spatial resolution should be enhanced further for the development of super high resolution visual media such as the digital cinema. The most straightforward approach for enhancing the spatial resolution is to apply the production technology to reducing the pixel size, viz., the area of each photo detector, in order to increase the number of pixels. As the pixel size decreases, however, so does the amount of light available for each pixel. Hence, the picture quality is degraded because the existence of shot noise, viz., variation of input, is unavoidable in principle. To keep shot noise invisible on a monitor, there needs to be a limitation in the pixel size reduction, and the limitation is estimated at approximately $50\mu m^2$ [2]. The current solid-state imaging device technology has almost reached this limit. Therefore, a new approach is required to enhance the spatial resolution further beyond this limit.

One promising approach towards improving spatial resolution further is to incorporate signal processing techniques into the imaging process. Along the lines, we presented a new signal processing based method for acquiring an improved resolution image with sufficiently high signal-to-noise ratio (SNR) by processing and integrating multiple images taken simultaneously with multiple cameras[1],[9],[11]. Originally we used multiple cameras with the same pixel aperture, where the term of the 'pixel aperture' means the spatial shape of each photo detector of a solid-state imager. Afterwards, we have found that to obtain the spatially uniform resolution improvements in the same-aperture case there must be limitations both in the arrangement of multiple cameras and in the geometrical configuration of the scene[10]. In the same aperture case, if and only if multiple cameras are coplanar and the object of imaging is a two-dimensional plate perpendicular to their optical axes, the spatial uniformity of the resultant resolution will be guaranteed. The limitations are considered to be very severe, and lower the applicability of the signal-processing based imaging scheme.

The utilization of multiple different-aperture cameras frees the signal processing based imaging scheme from the above limitations completely. The signal processing based imaging scheme will work well on the assumption that imaged areas of pixels of multiple cameras do not coincide with each other. In the different aperture case, imaged areas of pixels do not coincide with each other, irrespective of the arrangement of multiple cameras and the geometrical configuration of the scene. In the different aperture case, however, the aliasing artifacts included in low-resolution images taken with multiple different cameras disturb the smooth performance of the signal processing for integrating the multiple low-resolution images into an improved resolution image, more heavily than in the same aperture case because each camera produces its own aliasing artifacts corresponding to its pixel aperture. In the different-aperture case, it is especially important to render the signal processing robust under the condition that aliasing is occurring severely and in various ways. To solve this problem, we incorporate a new, alternately iterative algorithm for integrating multiple low-resolution images into the signal processing based imaging scheme.

2. Original Concept

2.1. Concept

Figure 7.1 illustrates the concept of the signal processing based image acquisition scheme using multiple cameras. It trades computational complexity of signal processing for improvement in spatial resolution. The concept of this image acquisition method is to unscramble the within-passband and aliased frequency components, which are weighted differently in undersampled images, by integrating multiple images taken simultaneously with multiple cameras, and then to restore the frequency components up to high frequencies so as to obtain an improved-resolution image with sufficiently high SNR. The signal-processing based approach using multiple cameras consists of the two stages.

2.1.1 Estimation of discrepancies and integration of sampling points (Registration). Firstly, we estimate the relative discrepancies of pixels between two different images chosen from among multiple images with fractional pixel accuracy, and then we combine together sampling points in multiple images, according to the estimated discrepancies, to produce a single image plane where sampling points are spaced nonuniformly. The estimation of discrepancies is referred to as registration, and is extensively studied in various fields of image processing. As for the estimation of discrepancies in fractional pixel accuracy,

the block-matching technique and the least square gradient technique may be used[6], but the work presented here employs the block-matching technique to measure relative shifts.

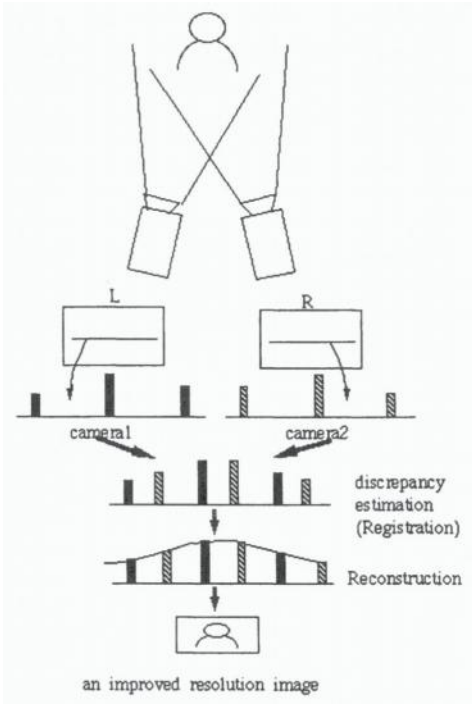


Figure 7.1. Schematic diagram of the signal processing based image acquisition using multiple cameras.

2.1.2 Reconstruction of an improved-resolution image (Reconstruction). In the reconstruction stage, an improved resolution image with uniformly-spaced samples is reconstructed from the nonuniformly spaced samples composed of samples of multiple images. This stage involves a two-dimensional reconstruction method for producing uniformly spaced sampling points from nonuniformly spaced sampling points. Reconstruction is an extensively treated problem, and for this purpose we might use various reconstruction methods, e.g., a polynomial interpolation method[5], a coordinate transformation method[4], a DFT-based method[14], an iterative method[3],[7],etc. The methods suggested so far have individual limitations in the two-dimensional case. Presently,

we believe that the iterative method using the Landweber algorithm[12] is fairly flexible, it is suited to the present problem, and thus is a promising candidate for the two-dimensional reconstruction method relevant to the VHD image acquisition. Hence we use the Landweber-type iterative method[12].

The achievable passband, when using a camera of solid-state-type imaging device, depends on the aperture impulse response function of a photo detector, provided the degradation by an optical lens is small enough. In other words, the aperture effect determines the upper bound in the performance of an imaging system.

Our theoretical analysis based on the aperture effect, which determines the upper bound of the performance of this method, has yielded the following conclusions:

1 If we use two cameras with the same pixel aperture and the aperture ratio is 100%, the resolution improvement is limited to twice the original sampling frequency. (We define the aperture ratio as the ratio of the total area of all the photo detectors of a solid-state imager to the area of the imaging chip surface of the solid-state imager.)

2 If we use more than two cameras with the same pixel aperture, SNR of the reconstructed improved-resolution image will be enhanced even in the case of the unit aperture ratio. This means that the pixel size can be further reduced in advance, because the photo detector size limitation is determined by SNR. Therefore, the spatial resolution can be further increased in advance.

2.2. Registration

The block-matching technique firstly interpolates given input images N times, compares image blocks of one magnified image to the other magnified image, and then determines the displacement of 1/N pixel accuracy which gives the best similarity between the two blocks. In most cases, the block-matching method works well to provide stable estimation of relative shifts, but involves large computational efforts. The least square gradient technique computes the displacement of fractional pixel accuracy in itself, and can handle a linear transformation as well as translation; but occasionally leads to unstable erroneous estimation of discrepancies. The work presented here employs the block-matching technique to measure relative shifts.

Undersampled images include aliased frequency components, which might cause errors in the estimation process. It is important to render the discrepancy estimation robust under the condition that aliasing is

occurring severely. Because of this, the work reported here limits the estimation accuracy to one-fourth pixel accuracy.

2.3. Reconstruction

Fig.7.2 illustrates the reconstruction method based on the Landweber algorithm. The relevant reconstruction problem is to find uniformly spaced samples g using the observed nonuniformly spaced samples f. The Landweber algorithm solves the reconstruction problem by starting with initial uniformly spaced samples $g^{(0)}$ and then updating the value of the uniformly spaced samples $g^{(n)}$ iteratively by the recurrence formula

$$g^{(n+1)} = B \circ g^{(n)} = g^{(n)} + \alpha \cdot A^* \circ (f - A \circ g^{(n)}) \qquad (7.1)$$

Where A is the nonuniform sampling process, A^* is the adjoint operator of the operator A and the value of the parameter a should be set small so that the operator B is a contraction mapping to guarantee the convergence. Prom the fixed point theorem [13], it follows that $g^{(n)}$ converges to an attracting fixed point $g^{(\infty)} = A^+ \circ f$, where A^+ is the Moore-Penrose pseudo inverse operator of the operator A, irrespective of an initial entity $g^{(0)}$ as n grows to infinity. Hence we can choose an initial entity $g^{(0)}$ arbitrarily.

In Fig.7.2, given the observed nonuniformly spaced samples f, the method starts with initial uniformly spaced samples $g^{(0)}$ and produces uniformly spaced samples $g^{(n)}$, approximating the original unknown uniformly spaced samples g better than $g^{(n-1)}$ by the iterative application of eq.7.1. Assuming that the pixel aperture of a camera is square and its sensitivity is uniform within the aperture, the operator A represents the nonuniform sampling process, as illustrated in Fig.7.2, which produces the low-resolution nonuniformly spaced sample S_a by averaging the corresponding pixels within the region R_a of the high resolution uniformly sampled image. On the other hand, as illustrated in Fig.7.2, the operator αA^* is used to distribute the error $f - A \circ g^{(n)}$ of the low-resolution nonuniformly spaced sample S_a onto the region R_a with the 0th order interpolation method. In the high-resolution uniform sample domain, all interpolated errors are combined by summing them. Irani and Peleg's (intuitively deriving) approach[7] belongs to this general formulation, although they have not dealt with the nonuniform displacement case.

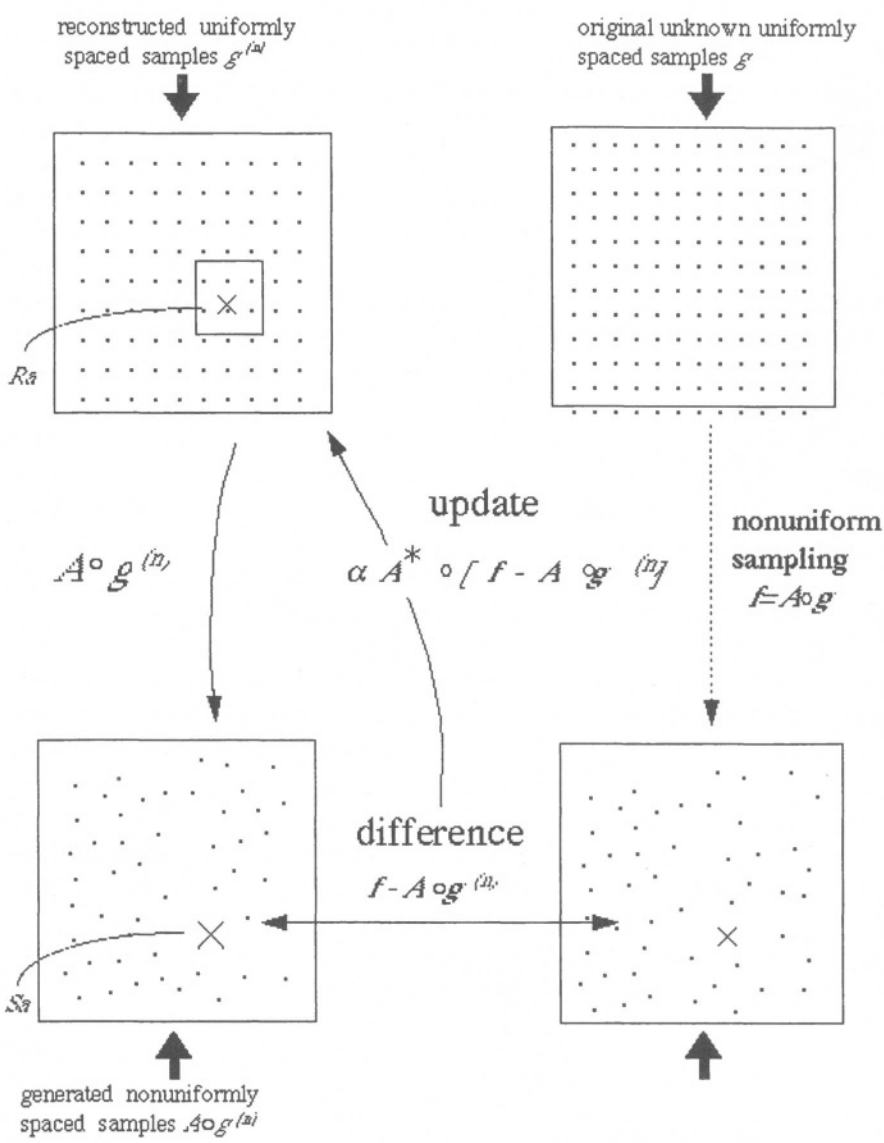


Figure 7.2. Reconstruction method based on the Landweber algorithm.

3. Image Acquisition with Multiple Different-Aperture Cameras

3.1. Weak Points in the Imaging Scheme with Multiple Same-Aperture Cameras

When we use multiple cameras with the same pixel aperture, the improvements of spatial resolution depend both on the arrangement of multiple cameras and on the geometrical configuration of the scene. Hence, in general, we might not expect that the resolution improvements will be spatially uniform everywhere in the reconstructed image. If and only if multiple cameras are coplanar and the object of imaging is a two-dimensional plate perpendicular to their optical axes, the spatial uniformity of the resultant resolution will be guaranteed. Let us consider two typical cases where the spatial uniformity of the resultant resolution cannot be guaranteed.

One typical case is as follows: two cameras are located with convergence; their optical axes intersect at a point in the middle of the two-dimensional object plate. In this case, in the portion of the two-dimensional object plate where the optical axes of the two imagers intersect, a pixel of one image taken with one camera almost coincides with some pixel of another image taken with another camera. Hence, in this portion, spatial resolution is not improved satisfactorily.

On the other hand, at a distance from the middle of the two-dimensional object plate there lies a portion where a pixel of one image does not coincide with any pixel of another image completely. Hence, in this portion, spatial resolution is improved to some extent. Therefore, in this case, the resolution improvements are never spatially uniform.

Another typical case is as follows: two cameras are coplanar, but the object of imaging is not a two-dimensional plate perpendicular to their optical axes. Fig.7.3 illustrates this case. Let us imagine that a point $P(X, Y, Z)$ on the object surface is projected onto the two cameras. On the image plane of a reference camera, the projected location is

$$x_0 = \frac{f}{z}X, \tag{7.2}$$

$$y_0 = \frac{f}{z}Y, \tag{7.3}$$

where f is the focal length. On the image plane of another camera, the projected location is

$$x_1 = \frac{f}{z}(X - a) = x_0 - \frac{f}{z}a, \tag{7.4}$$

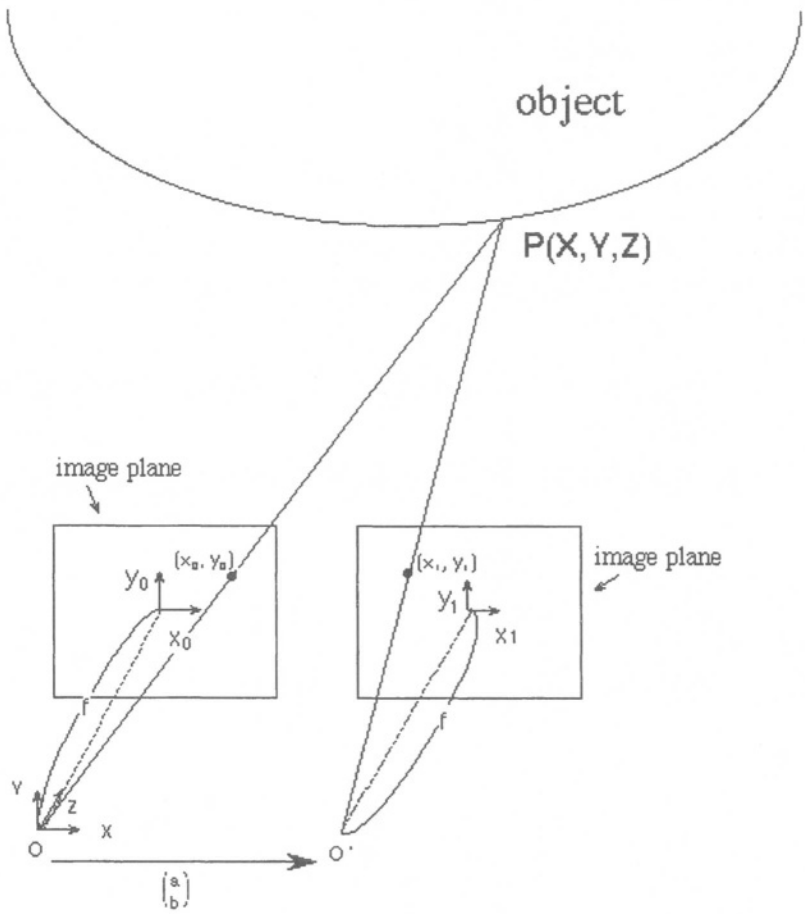


Figure 7.3. Arrangement of the two cameras. The two cameras are coplanar, but the object of imaging is not a two-dimensional plate perpendicular to their optical axes.

$$y_1 = \frac{f}{z}(Y - b) = y_0 - \frac{f}{z}b, \tag{7.5}$$

where a and b are the baseline distances in the x and y directions, respectively, between the two cameras. Let us suppose that the parameters (a, b, f) are known in advance. The relative shift of each sampling point position between the two images is

$$d_x = x_1 - x_0 = -a f \frac{1}{z} \propto \frac{1}{z}, \tag{7.6}$$

$$d_y = y_1 - y_0 = -b f \frac{1}{z} \propto \frac{1}{z}. \tag{7.7}$$

The relative shift (d_x, d_y) changes according to the depth of the point on the object surface. Hence, on the image plane there exists a portion where a pixel of one image almost coincides with some pixel of another image. Therefore, in this case also, the resolution improvements are never spatially uniform.

3.2. Utilization of Multiple Different-Aperture Cameras

As shown in Fig.7.4, if we use multiple cameras with different pixel

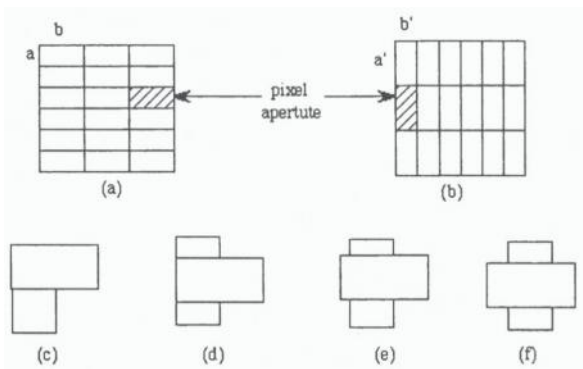


Figure 7.4. Different pixel apertures and their overlapping patterns: (a) horizontally wide rectangular pixel aperture; (b) vertically long rectangular pixel aperture; (c) overlapping pattern, No.l; (d) overlapping pattern, No.2; (e) overlapping pattern, No.3; (f) overlapping patern,No.4.

apertures, a pixel of one image taken with one camera would never coincide with any pixel of another image taken with another camera, irrespective of the arrangement of multiple cameras and/or the geometrical

configuration of the scene. Hence, we might expect that the resultant resolution will be spatially uniform.

Fig.7.5 compares a reconstructed image in the different aperture case with that in the same aperture case. Fig.7.6 illustrates the different two

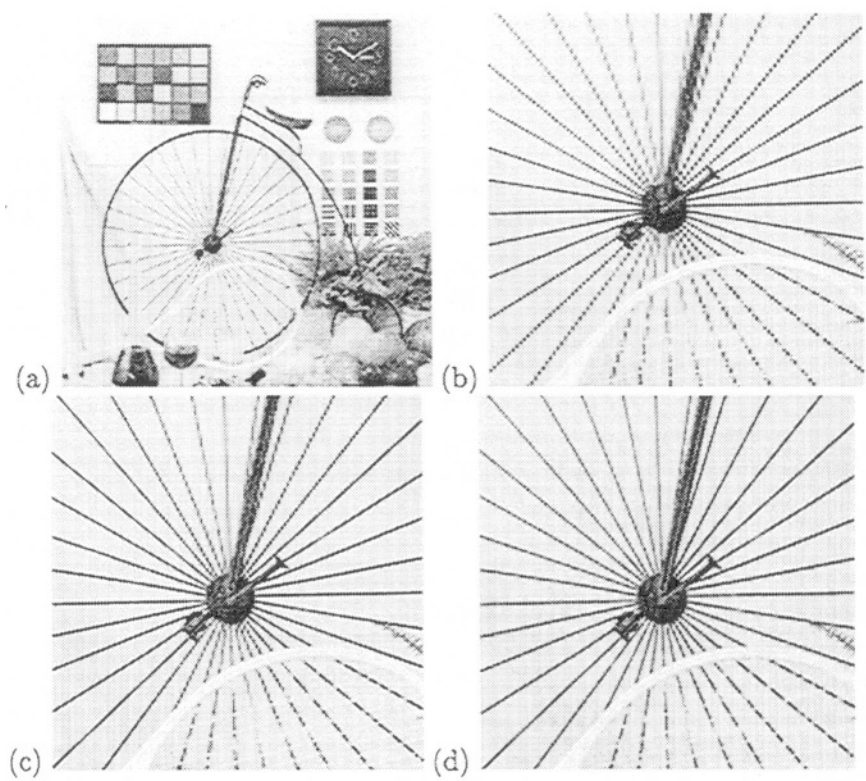


Figure 7.5. Reconstructed image in the different-aperture case versus that in the same-aperture case; two cameras are located horizontally with 1° convegence and the relative displacement is assumed to be known perfectly: (a) two-dimensional test plate, (b) part of the image captured by a single camera, (c) part of the image reconstructed with two cameras having the same pixel aperture, (d) part of the image reconstructed with two cameras having different pixel apertures.

pixel apertures used in the different-aperture case. In the simulation of Fig.7.5, we used a high resolution printing digital image data with 2048 pixels x 2048 pixels as a two-dimensional test plate, and we assume that two cameras are located horizontally with 1° convergence; their optical axes intersect at a point in the middle of the two-dimensional test plate and the angle between their optical axes is 1°. Moreover, we suppose that the aperture ratio is 100%. Under these conditions,

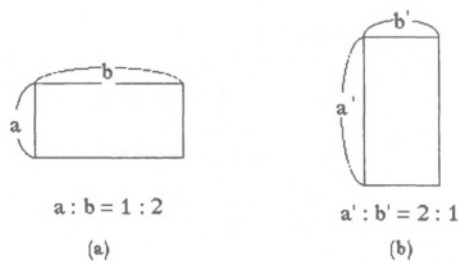


Figure 7.6. Different pixel apertures for two cameras used in the experimental simulations of Fig.7.5, Figs.7.8-7.11:(a) horizontally wide rectangular pixel aperture used for a left camera, (b) vertically long rectangular pixel aperture used for a right camera.

we simulate the imaging process of each camera with the horizontally wide rectangular aperture or the vertically long rectangular aperture of Fig.7.6; viz., we simulate the projection of the two-dimensional test plate onto the image plane of each camera with low resolution, and thus produce projected low-resolution images. In the simulation of Fig.7.5, we use the computationally projected low-resolution images instead of a real image captured with each camera. To avoid some unfavorable effect of using the computationally projected low-resolution images, we limit the spatial resolution, viz., the number of pixels, of each camera to one-tenth of the spatial resolution of the test digital image data. Furthermore, we assume that the relative shifts of pixels between the two computationally projected low-resolution images are known perfectly, and we simulate the imaging scheme with two cameras.

Fig.7.5(a) shows a high-resolution image used as a two-dimensional test plate. Fig.7.5(b) shows part of the image captured by a single camera with the horizontally wide rectangular pixel aperture of Fig.7.6(a), but in this figure, for ease of comparison, the image is interpolated and magnified with the sinc function in the horizontal direction. Fig.7.5(c) shows part of the image reconstructed by the imaging scheme with two cameras having the same pixel aperture, viz., the horizontally wide rectangular pixel aperture of Fig.7.6(a). Fig.7.5(d) shows part of the image reconstructed by the imaging scheme with two cameras having different pixel apertures; one camera has the horizontally wide rectangular pixel aperture of Fig.7.6(a) and another camera has the vertically long rectangular pixel aperture of Fig.7.6(b). In the same-aperture case of Fig.7.5(c), the aliasing artifacts are not fully eliminated, and the resultant resolution does not seem to be spatially uniform; around the vertical spokes of the bicycle the aliasing artifacts are not eliminated

well, but in the portion of the spherical and square test charts the resolution improvements are clearly visible. On the other hand, in the different-aperture case of Fig.7.5(d), the aliasing artifacts are eliminated well, and the resultant resolution seems to be spatially uniform.

As is obvious from this instance, the utilization of multiple different-aperture cameras will be advantageous on condition that the displacement estimates are derived in the registration stage with sufficiently high accuracy.

3.3. Frequency Response of the Imaging Scheme with Multiple Different-Aperture Cameras

In order to estimate the frequency response of the imaging scheme with two cameras having the different pixel apertures of Fig.7.6 by experimental simulations, we used a digital image data with a single spatial frequency component as shown in Fig.7.7 as a two-dimensional test plate. We estimate the frequency response by changing the spatial fre-

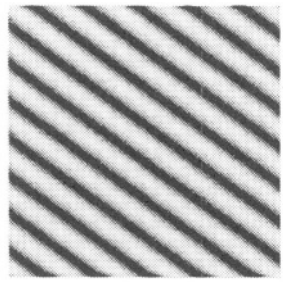


Figure 7.7. Image data used for estimation of a frequency responce of an imaging scheme.

quency of the digital image data used as the test plate and by using the projected low-resolution images produced from the digital image data computationally in the same way as described in Section 3.2. We define the frequency response as the modulation transfer function, viz., the ratio of the output power of the frequency component contained in the image reconstructed by the imaging scheme with the two different-aperture cameras to the input power of the digital image data with the single spatial frequency component, under the condition that the two cameras are located horizontally in parallel and the test plate is located perpendicularly to the optical axes of the two cameras. Moreover, we suppose that the aperture ratio is 100 %. In this case, the relative shift of each sampling point position between the two images captured with

the two cameras is spatially uniform and assumed to be known perfectly in the simulation. However, as shown in Figs.7.4(c)-7.4(f) there exist various overlapping patterns between the two different pixel apertures. In the simulation, however, we apply the experimental analysis of the frequency response only to the cases of Figs.7.4(c) and 7.4(f), because the cases of Figs.7.4(d) and 7.4(e) are considered to provide intermediate frequency responses between those in the cases of Figs.7.4(c) and 7.4(f).

Figure 7.8 shows the estimated frequency responses. In Fig.7.8, in the spatial frequency domain normalized in the radius direction, the contour lines of -1dB, -3dB and -6dB are shown for each estimated frequency response. Figures 7.8(a) and 7.8(b) show the estimated frequency responses of the imaging scheme with the two different-aperture cameras, and Figs.7.8(a) and 7.8(b) correspond to the estimated frequency responses in the cases of Figs.7.4(c) and 7.4(f), respectively. Figure 7.8(c) shows the estimated frequency response of a single camera with the vertically long rectangular pixel aperture of Fig.7.6(b). Figure 7.8(d) shows the estimated frequency response of a single camera with the square pixel aperture with the area equal to that of the vertically long rectangular pixel aperture of Fig.7.6(b). In the case of a single camera such as Fig.7.8(c) and 7.8(d), the estimated frequency response is identical to the usual modulation transfer function.

As is obvious from Fig.7.8, the imaging scheme with multiple different-aperture cameras provides an improved frequency response, which approximates the logical sum of frequency responses of multiple cameras used in the imaging scheme and is almost kept unchanged irrespective of overlapping patterns shown in Fig.7.4.

3.4. Alternately Iterative Algorithm for Registration and Reconstruction

In the different-aperture case, there still remains an unsolved problem about the registration stage. In the different-aperture case, all the image contents such as spatial frequency components contained in one image taken with one camera do not necessarily appear in another image taken with another camera; the aliasing artifacts produced by each camera change according to its pixel aperture. Therefore, it would seem that separately from the reconstruction stage we cannot render only the registration stage robust under the condition that aliasing is occurring severely, and in various ways.

On the other hand, however, once we reconstruct an improved resolution image by integrating low-resolution images taken with multiple different-aperture cameras, then we might utilize the improved-resolution

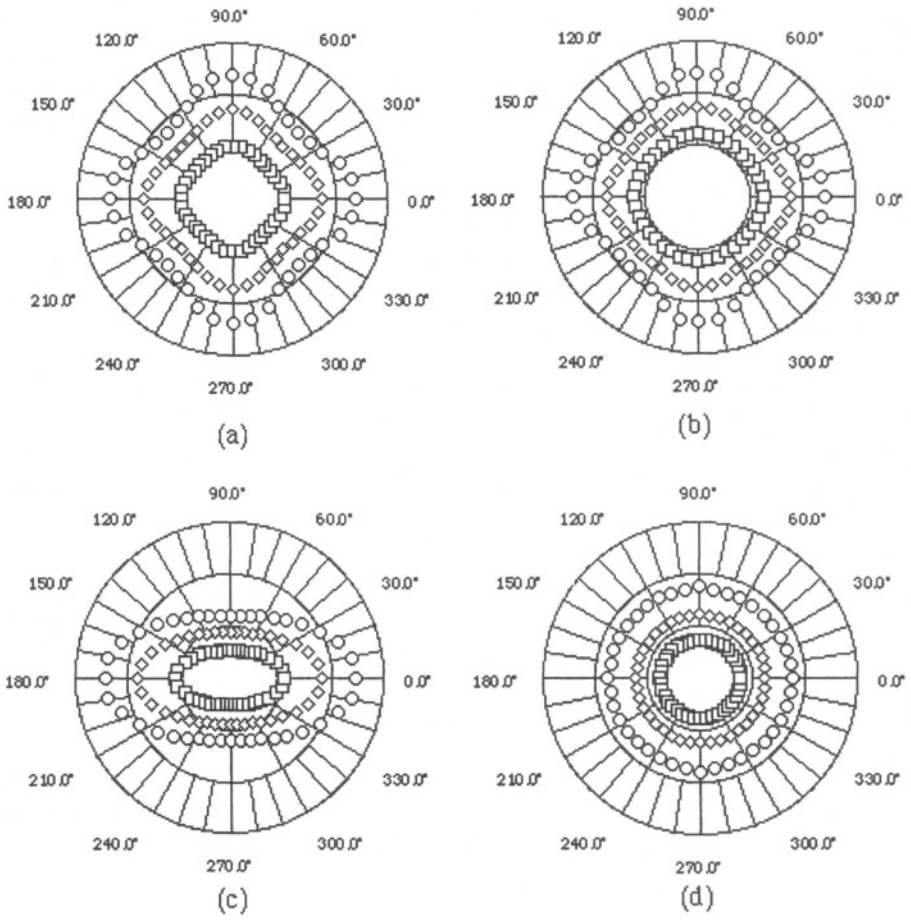


Figure 7.8. Estimated frequency responses of imaging schemes [(□) -1dB,(◇) -3dB,(○) -6dB]:(a)imaging schemes with two different-aperture cameras in the case of the overlapping pattern of Fig.7.4(c),(b)imaging scheme with two different-aperture cameras in the case of the overlapping pattern of Fig.7.4(f),(c)single camera with the vertically long rectangular pixel aperture of Fig.7.6(b),(d) single camera with the square pixel aperture with the same area as that of the vertically long rectangular pixel aperture of Fig.7.6(b).

image as a reference image in the registration stage so as to enhance its estimation accuracy; we might estimate the relative shift of each sampling-point position of one image compared to the reference improved-resolution image with higher accuracy to some extent, because the reference improved-resolution image holds most of the image contents lying in a low resolution image taken with each camera. Prom this viewpoint, we might conjecture that for the increase of the estimation accuracy in the registration stage it is indispensable to combine the registration stage with the reconstruction stage. Along the lines, we develops an alternately iterative algorithm for performing the registration operation together with the reconstruction operation.

The alternately iterative algorithm is organized as follows:

Alternately Iterative Algorithm

1 Apply the block-matching technique to images $I_0 \sim I_{M-1}$ taken with multiple different-aperture cameras $C_0 \sim C_{M-1}$, and then estimate the relative discrepancies of pixels in each of the observed images $I_1 \sim I_{M-1}$ compared to the reference observed image I_0. In this step, firstly interpolate and magnify each observed image I_i horizontally and vertically in the ratio according to its pixel aperture with the 0th order interpolation technique so as to obtain the magnified image $\hat{I}_i$ of the same size prescribed in advance, then compare image blocks in each of the magnified images $\hat{I}_1 \sim \hat{I}_{M-1}$ to the reference magnified image $\hat{I}_0$, and finally determine the displacement with fractional pixel accuracy.

2 Integrate all the pixels of the observed images $I_1 \sim I_{M-1}$ into the reference observed image I_0 according to the displacement estimates derived in the step 1, and then reconstruct an improved resolution image I with the Landweber algorithm.

3 This step is similar to the step 1 except that the improved resolution image I, instead of the observed image I_0, is used as a reference image. In this step, firstly compare image blocks in each of the magnified images $\hat{I}_1 \sim \hat{I}_{M-1}$ to the reference improved-resolution image I, and then estimate the relative discrepancies.

4 This step is similar to the step 2 except that the displacement estimates derived in the step 3, instead of those derived in the step 1, are used. In this step, reconstruct an improved-resolution image I anew by integrating the observed images $I_1 \sim I_{M-1}$ into the reference observed image I_0 according to the displacement estimates derived in the step 3. If the improved resolution image I

is almost unchanged compared to the preceding one, then halt the algorithm; otherwise, return to the step 3.

4. Experimental Simulations

In order to evaluate the alternately iterative algorithm for the whole process of registration and reconstruction in the different-aperture case, first we used a two-dimensional zone plate chart as shown in Fig.7.9(a).To make a projected low resolution image from a two-dimensional test plate, which simulates the imaging process with each camera, we modeled the imaging process as follows:

1 The projection is perspective.

2 Two cameras are located horizontally in parallel; their optical axes are parallel.

3 The two-dimensional test plate is located perpendicularly to the optical axes of the two cameras.

4 The two cameras have different pixel apertures as shown in Fig.7.6; the aperture ratio of each camera is 100%.

We used a two-dimensional zone plate chart as a two-dimensional test plate. The zone plate chart has the special property that the magnitude of its Fourier transform is the same as its own magnitude. Hence, we can clearly see the aliased frequency components produced by the imaging process as a moire pattern, directly on the projected image of the zone plate chart. Figures 7.9 and Fig.7.10 show the simulation results. Figure 7.9(a) shows an original zone plate chart used as a two-dimensional test plate. Figure 7.9(b) shows part of the image captured by the left camera with the horizontally wide rectangular pixel aperture of Fig.7.6(a), but in this figure, for ease of comparison, the image is interpolated and magnified with the sinc function in the horizontal direction. Figure 7.9(c) shows part of the image captured by the right camera with the pixel aperture of Fig.7.6(b), but in this figure the image is interpolated and magnified in the vertical direction. Figure 7.9(d) shows part of the improved resolution image which is reconstructed under the ideal condition that the relative displacement between the two images of Fig.7.9(b) and 7.9(c) is known perfectly. Figures 7.9(e) ~7.9(h) shows part of the improved resolution image reconstructed at each iteration step of the alternately iterative algorithm.

Figure 7.10 illustrates the accuracy chart of the displacement estimates derived at each iteration step of the alternately iterative algorithm, in the fpllowing manner. The white area of the accuracy chart

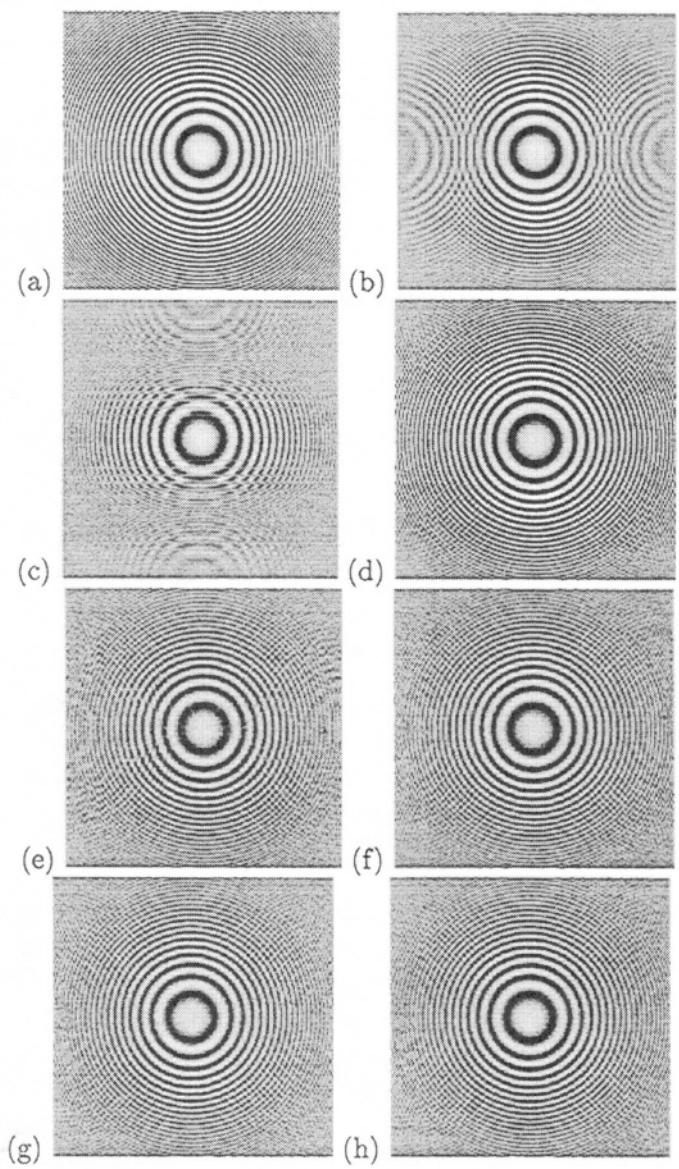


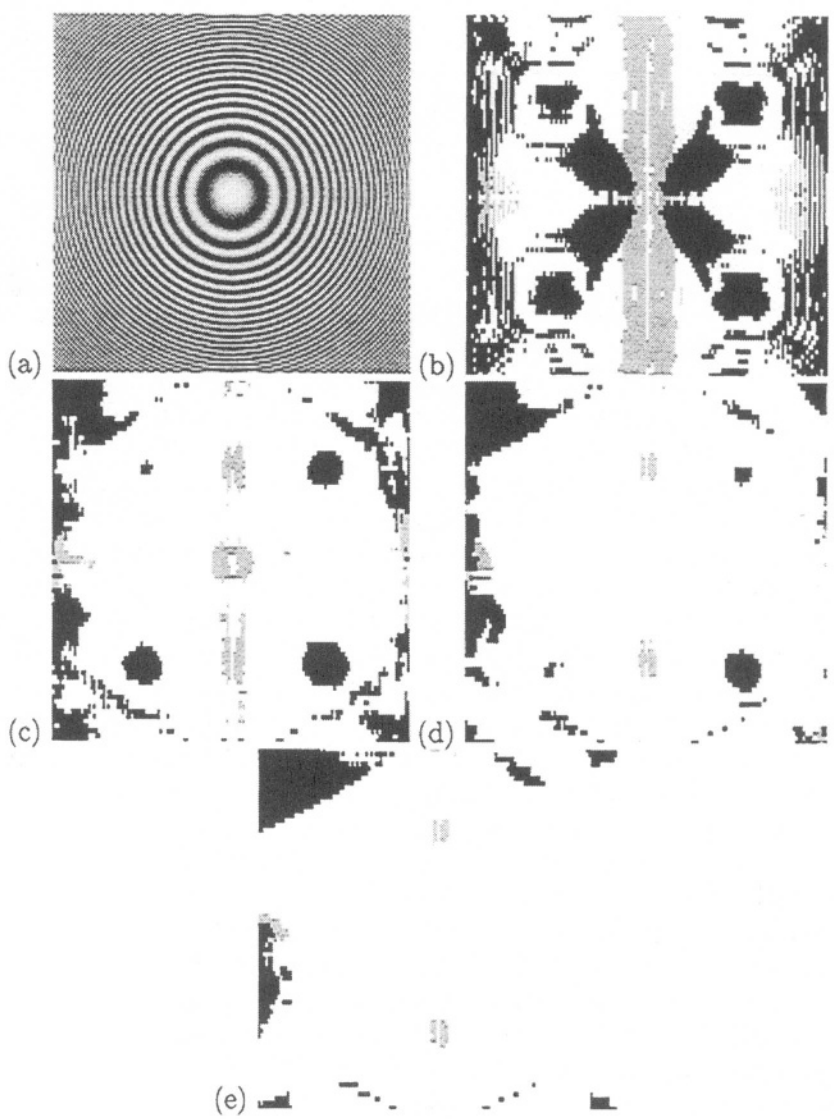
Figure 7.9. Results of experimental simulations conducted on a two-dimentional zone plate chart; reconstructed images provided by the alternately iterative algorithm: (a) original zone plate chart, (b) part of the image captured by the left camera, (c) part of the image captured by the right camera, (d) part of the image reconstructed under the ideal condition that the relative displacement is known perfectly, (e) part of the image reconstructed at the 1st iteration step, (f) part of the image reconstructed at the 2nd iteration step, (g) part of the image reconsturucted at the 4th iteration step, (h) part of the image reconstructed at the 10th iteration step.

represents pixel positions where the displacement estimate is provided within one-fourth pixel accuracy; the black area denotes pixel positions where the difference between the displacement estimate and the true value exceeds a half pixel; the gray area corresponds to pixels for which the displacement estimate is not provided within one-fourth pixel accuracy, but within one-half pixel accuracy.

In the image of Fig.7.9(e) reconstructed at the initial iteration step of the alternately iterative algorithm, which image is identical to the image reconstructed by the conventional non-iterative algorithm[9], because of insufficient accuracy of the displacement estimates, the resolution improvements do not reach the improvement level of the image of Fig.7.9(d), which is reconstructed under the ideal condition that the displacement is known perfectly. At the early iteration steps (such as the 2nd or the 3rd iteration step) only in the relatively narrow, low-frequency region lying around about the middle of the zone plate chart, the displacement is estimated with satisfactorily high accuracy. With the advance of the iteration, however, as is obvious from Fig.7.10, the region where the highly accurate displacement estimates are provided is extended gradually to the peripheral higher frequency region lying at a distance from the middle of the zone plate chart. Furthermore, after a few iterations the algorithm provides an ideally improved resolution image; we cannot distinguish between the image of Fig.7.9(h) reconstructed at the 4th iteration step and the image of Fig.7.9(d) reconstructed under the ideal condition that the displacement is known perfectly.

In the second simulation, we used two real images of Fig.7.11(a) and Fig.7.11(b). These two images are generated from the images taken with one stationary video camera. We produce the image of Fig.7.11(a) by computing average intensity of horizontally adjacent two pixels, whereas we generate the image of Fig.7.11(b) by computing average intensity of vertically adjacent two pixels. These averaging operations are meant to simulate imaging the scene with the two cameras having different pixel apertures as shown in Fig.7.6. Figure 7.11(c) shows the improved-resolution image that is reconstructed at the 10th iteration step of the alternately iterative algorithm. In Fig.7.11, after several iterations, the algorithm provides a fairly improved resolution image, which is analogous to the case of Fig.7.9.

From these results, we might conclude that the alternately iterative algorithm is a very potential technique for the signal-processing based image acquisition scheme with multiple different-aperture cameras.



	Accuracy Δd
white	$\Delta d \leq 1/4(pixel)$
gray	$1/4(pixel) < \Delta d \leq 1/2(pixel)$
black	$1/2(pixel) < \Delta d$

Figure 7.10. Accuracy chart of the displacement estimates derived at each iteration step for a two-dimensional zone plate chart: (a) original zone plate chart, (b) accuracy chart at the 1st iteration step, (c) accuracy chart at the 2nd iteration step, (d) accuracy chart at the 4th iteration step, (e) accuracy chart at the 10th iteration step.

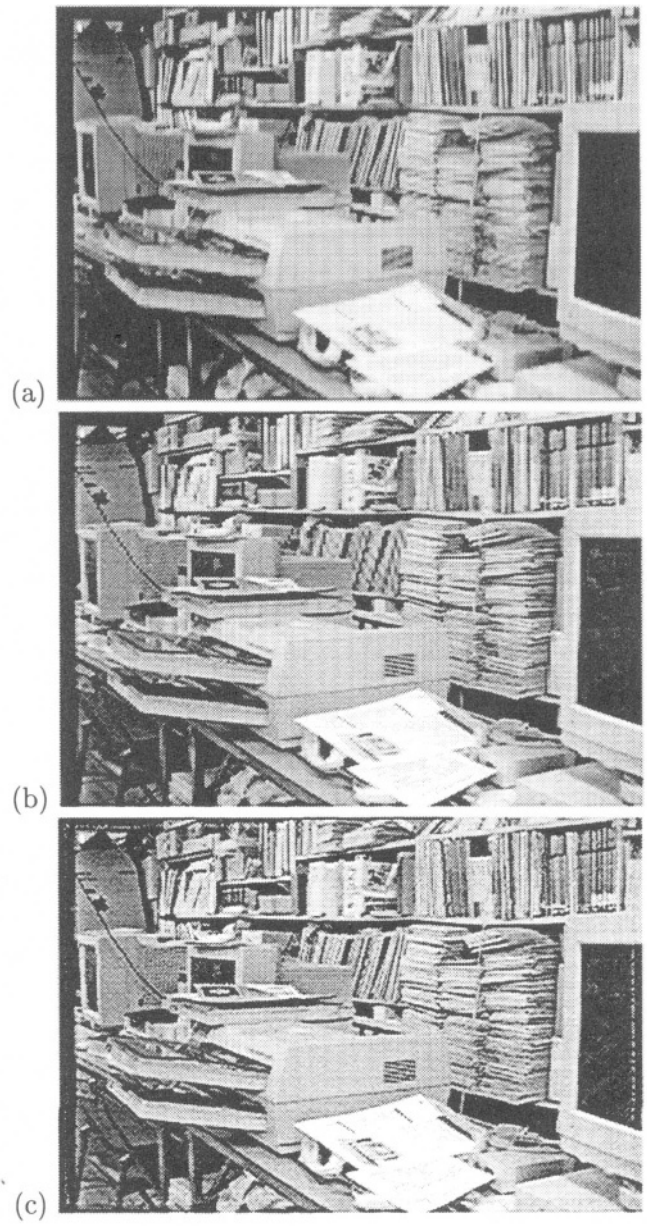


Figure 7.11. Result of the experimental simulation conducted on real images; a reconstructed image provided by the alternately iterative algorithm: (a) the image captured by the left camera, (b) the image captured by the right camera, (c) the image reconstructed at the 10th iteration step.

5. Conclusions

The signal processing based imaging scheme with multiple cameras is to produce an improved resolution image by integrating multiple images taken simultaneously with multiple cameras, and the imaging scheme is very promising for improving spatial resolution further beyond the physical limit of existing image acquisition devices. If imaged areas of pixels of multiple cameras do not coincide with each other, the signal processing based imaging scheme will work well. When we use multiple cameras with the same pixel aperture, there are severe limitations both in the arrangement of multiple cameras and in the configuration of the scene, in order to guarantee the spatial uniformity of the resultant resolution. The utilization of multiple different-aperture cameras frees the signal processing based imaging scheme from the above limitations completely. In this case, however, the registration problem becomes more difficult to solve than in the same aperture case. To solve the registration problem, we incorporate a new, alternately iterative algorithm into the signal processing based imaging scheme. The experimental simulations demonstrate that the utilization of multiple different-aperture cameras is a promising method and that the alternately iterative algorithm is a very potential technique for the signal processing based image acquisition scheme with multiple different-aperture cameras.

References

[1] AIZAWA K., KOMATSU T., and SAITO T., "A scheme for acquiring very high resolution images using multiple cameras", *Proc. of IEEE 1992 Int. Conf. Accoust., Speech and Signal Process.*, 1992, pp.III:298-III:292.

[2] ANDO T., "Trends of high-resolution and high-performance solid state imaging technology" (in Japanese), *Journal of ITE, Japan*, 1990, *44*, pp.105-109

[3] CARLOTTO M. J., and TOM V.T., "Interpolation of two-dimensional surfaces using Gerchberg algorithm", *SPIE*, 1982, *359*, pp.226-230.

[4] CLARK J. J., PALMER M. R., and LAURENCE P. D., "A transformation method for the reconstruction of functions from nonuniformly spaced samples", *IEEE Trans. on Accoust., Speech and Signal Proc.*, 1985, *vol. 33*, pp.1151-1165.

[5] FRANKE R. W., "Scattered data interpolation: tests of some methods", *Math.Comput.*, 1982, *38*, pp.181-200.

[6] HORN B. K. P., *Robot vision*, (The MIT Press,MA.1986).

[7] IRANI M., and PELEG S., "Improving resolution by image registration", *CVGIP: Graphical Models and Image Processing*, 1990, *53*, pp.231-239.

[8] KIM S. P., and BOSE N. K., "Reconstruction of 2-D bandlimited discrete signals from nonuniform samples", *Proc. IEE-F*, 1990, *137*, pp.197-204.

[9] KOMATSU T., AIZAWA K., and SAITO T., "A proposal of a new method for acquiring super high definition pictures with two cameras" (in Japanese), *Journal of ITE Japan*, 1991, *45*, pp.1256-1262.

[10] KOMATSU T., AIZAWA K., and SAITO T., "Very high resolution image acquisition via reconstruction using multiple lower-resolution cameras various in their pixel apertures", *Proc.Int. Workshop on HDTV'92*, 1992, pp.73:l-73:8.

[11] KOMATSU T., AIZAWA K., IGARASHI T., and SAITO T., "Signal-processing based method for acquiring very high resolution images with multiple cameras and its theoretical analysis", *Proc. IEE-I*, 1993, *140*, pp.19-25.

[12] LANDWEBER L., "An iteration formula for Fredholm integral equations of the first kind", *American J. Math.*, 1951, *73*, pp.615-624.

[13] STARK H.(Ed.), *Image recovery: theory and application* (Academic Press,NY,1977).

This page intentionally left blank

Chapter 8

SUPER-RESOLUTION FROM MULTIPLE IMAGES HAVING ARBITRARY MUTUAL MOTION

Assaf Zomet and Shmuel Peleg
School of Computer Science and Engineering
The Hebrew University of Jerusalem
91904 Jerusalem, Israel
{zomet,peleg}@cs.huji.ac.il

Abstract Normal video sequences contain substantial overlap between successive frames, and each region in the scene appears in multiple frames. The super-resolution process creates high-resolution pictures of regions that are sampled in multiple frames, having a higher spatial resolution than the original video frames.

In this chapter existing super resolution algorithms are analyzed in the framework of the solution of large sparse linear equations. It is shown that the gradient of the function which is minimized when solving these equations can be computed by means of image operations like warping, convolutions, etc. This analysis paves the way for new algorithms, by using known gradient-based optimization techniques. The gradient is computed efficiently in the image domain instead of multiplying large sparse matrices.

This framework allows versatile imaging conditions, including arbitrary motion models, camera blur models, etc. Prior knowledge can also be combined efficiently for obtaining a MAP super resolution solution.

As an example, super resolution is implemented with the conjugate-gradient method, and a considerable speedup in the convergence of the algorithm is achieved compared to other methods.

Keywords: Super-Resolution MAP

1. Introduction

Video sequences usually contain a large overlap between successive frames, and regions in the scene are sampled in several images. This multiple sampling can sometime be used to achieve images with a higher spatial resolution. The process of reconstructing a high resolution image from several images covering the same region in the world is called *Super Resolution.* Additional tasks may include the reconstruction of a high resolution video sequence [1], or a high resolution 3D model of the scene [2].

A common model for super resolution presents it in the following way: The low resolution input images are the result of projection of a high resolution image onto the image plane, followed by sampling. The goal is to find the high resolution image which fits this model. Formulating it in mathematical language:

Given K images $\{X_L^{(n)}\}_{n=1}^{K}$ of size $M_1 \times M_2$, find the image X_H of size $N_1 \times N_2$, which minimizes the Error function:

$$E(X_H) = \sum_{n=1}^{K} \| P_n(X_H) - X_L^{(n)} \|^2$$

where:

1 $\| \cdot \|$ - Can be any norm, usually ℓ^2.

2 $P_n(X_H)$ is the projection of X_H onto the coordinate system and sampling grid of image $X_L^{(n)}$.

When this optimization problem does not have a single solution, additional constraints may be added, expressing prior assumptions on X_H, such as smoothness.

The projection $P_n(X_H)$ is usually modeled by four stages:

1 *Geometric Transformation*

2 *Blurring*

3 *Subsampling*

4 *Additive Noise*

The major differences between most modern algorithms are in the optimization technique used for solving this set of equations, the constraints on X_H which are added to the system, and the modeling of the geometric transformation, blur and noise.

1.1. Modeling the Imaging process

1.1.1 Geometric Transformation. In order to have a unique super resolved image X_H, the coordinates system of X_H should be determined. A natural choice would be the coordinate system of one of the input images, enlarged by factor *q*, usually by two. The geometric transformation of X_H to the coordinates of the input images is computed by finding the motion between the input images.
Motion computation and image registration are beyond the scope of this paper. It is important to mention that high accuracy of registration is crucial to the success of super resolution. This accuracy can be obtained when and assumption on the motion model holds, such as an affine or a planar-projective transformation. For highly accurate model-based methods, see [3, 4, 5].

1.1.2 Blur. Image blur can usually be modeled by a convolution with some low-pass kernel. This space-invariant function should approximate both the blur of the optics, and the blur caused by the sensor. The spectral characteristics of the kernel determine whether the super resolution problem is uniquely solvable: If some (high) frequencies of the kernel vanish, then there is no single solution [6]. In this case, constraints on the solution may be added [7].
The digitized result of the camera blur is called "The *PSF - Point Spread Function*". Several ways to estimate it are:

1 Use camera manufacturer information (Which is hard to get).

2 Analyze a picture of a known object [8, 9].

3 Blind estimation of the PSF from the images [10].

Some algorithms can handle space-variant blur, such as space-variant-motion blur, air turbulence, etc.

1.1.3 Subsampling. Subsampling is the main difference between the models of super resolution and image restoration. Sometimes the samples from different images can be ordered in a complete regular grid, for example when the motion of the imaging device is precisely controlled. In this case image restoration techniques such as inverse-filtering and De-convolution can be used to restore a high resolution image. In the general case of a moving camera, the super resolution image is reconstructed from samples which are not on a regular grid. Still, image restoration techniques inspire some of the super resolution algorithms [11].

1.1.4 Additive Noise. In super resolution, as in similar image processing tasks, it is usually assumed that the noise is additive, normally distributed with zero-mean. Under this assumption, the maximum likelihood solution is found by minimizing the error under Mahalanobis Norm (using estimated autocorrelation matrix), or ℓ^2 norm (assuming uncorrelated "white noise"). The minimum is found by using tools developed for large optimization problems under these norms, such as approximated Kalman-filter [1], linear-equations solvers [11, 12], etc. The assumption of normal distribution of the noise is not accurate in most of the cases, as most of the noise in the imaging process is non-gaussian (quantization, camera noise, etc.), but modeling it in a more realistic way would end in a very large and complex optimization problem which is usually hard to solve.

1.2. Historical Overview

The theoretical basis for super resolution was laid by Papoulis [6], with *The Generalized Sampling Theorem* . It was shown that a continuous band-limited signal *G* can be reconstructed from samples of convolutions of *G* with different filters, assuming some properties of these filters (see 1.1.2-blur). This idea is simple to generalize to 2D signals (images).

A pioneering algorithm for super resolution for images was presented by Huang & Tsai [13], who made explicit use of the aliasing effect, assuming the image is band limited, and the images are noise-free. Kim et. al. generalized this work to noisy and blurred images, using least square minimization [14]. Spatial domain algorithm was presented by Ur & Gross [15]. Assuming a known 2D translation, a fine sample grid image was created from the input images, using interpolation, and the camera blur was canceled using deblurring technique. The above methods assumed blur function which is uniform over all the images, and identical on different images. They were also restricted to global 2D translation.

A different approach was suggested by Irani & Peleg [8, 16], based on previous work by Peleg et al. [17]. The basic idea, *Iterative Backward Projecting - IBP*, was adopted from computer-aided Tomography (CAT). The algorithm starts with an initial guess $X_H\{0\}$, and iteratively simulate the imaging process, reprojecting the error back to the super resolution image. This algorithm can handle general motion and non-uniform blur function, assuming they can be approximated accurately.

To reduce noise and solve singular cases, several algorithms incorporate prior knowledge into the computation by constraining the solution. Stark & Oskoui [18] and later Pati et. al. [12] base their algorithm on a set theoretic optimization tool called POCS (Projection Onto Convex Sets). It is assumed that convex constraints on the solution are known, so that their intersection is also a convex set. The implementation of the algorithm is similar to the IBP algorithm of Irani & Peleg, with a modification in the backprojection stage: The errors in the imaging are projected onto the solution image, while keeping the solution in the convex set defined by the constraints. Pati et. al. also added motion blur to the imaging process model.

Markov Random Field was also used to regularize super resolution. Shekarforoush et. al. [19] formulated the super resolution problem in probabilistic bayesian framework, and used MRF for modeling the prior, and finding the solution. Similar formulation was presented by Schultz & Stevenson [20], who use prior on the edges and smoothness of the image to compensate for bad motion estimation, based on Huber-Markov Random Field formulation.

There is a great similarity between super resolution and image restoration, and indeed many of the super resolution techniques are adopted from image restoration. A unifying framework for super resolution as a generalization of image restoration was presented by Elad & Feuer [11]. Super resolution was formulated using matrix-vector notations, and it was shown that existing super resolution techniques are actually variations of standard quadratic minimization techniques for solving linear equations sets. Based on this analysis, they proposed other sparse matrix optimization methods for the problem.

Finally, an analytical probabilistic method was recently developed by Shekarforoush & Chellapa [10]. They proved that super resolution image can be directly constructed by a linear combination of a basis which is biorthogonal to the PSF function. The combination coefficients are the input images intensity values. They also presented an algorithm for the estimation of the camera PSF from the images.

2. Efficient Gradient-based Algorithms

2.1. Mathematical Formulation

Super resolution can be presented as a large sparse linear optimization problem, and solved using explicit iterative methods [11, 21, 7]. In

the presented framework a matrix-vector formulation is used in the analysis [11], but the implementation is by standard operations on images such as convolution, warping, sampling, etc. Altering between the two formulations, a considerable speedup in the super resolution computation is achieved, by taking advantage of the two worlds: Implementing advanced gradient based optimization techniques (such as conjugate gradient), while computing the gradient in an efficient manner, using basic image operations, instead of sparse matrices multiplications.

In the analysis part images are represented as column vectors, (with any arbitrary order of the pixels). Basic image operations such as convolution, subsampling, upsampling and warping are linear, and thus can be represented as matrices operating on these vector images.

The image formation process can be formulated in the following way [11]:

$$\underline{x}_L^{(n)} = DH_nW_n\underline{x}_H + \underline{e}_n$$

where:

- $\underline{x}_H$ is the high resolution image X_H of size $[N_1 \times N_2]$, reordered in a vector.
- $\underline{x}_L^{(n)}$ is the n-th image of size $[M_1 \times M_2]$, reordered in a vector.
- $\underline{e}_n$ is the normally distributed additive noise in the n-th image, reordered in a vector.
- W_n is the geometric warp matrix, of size $[N_1N_2 \times N_1N_2]$
- H_n is the blurring matrix, of size $[N_1N_2 \times N_1N_2]$
- D is the decimation matrix, of size $[M_1M_2 \times N_1N_2]$

Stacking the vector equations from the different images into a single matrix-vector:

$$\begin{bmatrix} \underline{x}_L^{(1)} \\ \vdots \\ \underline{x}_L^{(K)} \end{bmatrix} = \begin{bmatrix} DH_1W_1 \\ \vdots \\ DH_KW_K \end{bmatrix} \underline{x}_H + \begin{bmatrix} \underline{e}_1 \\ \vdots \\ \underline{e}_k \end{bmatrix} \Longleftrightarrow \underline{x}_L = A\underline{x}_H + \underline{e}$$

For practical reasons it is assumed the noise is uncorrelated and has uniform variance. In this case, the maximum likelihood solution is found by minimizing the functional:

$$E(\underline{x}_H) = \frac{1}{2} \| \underline{x}_L - A\underline{x}_H \|^2$$

taking the derive of E with respect to $\underline{x}_H$, and setting the gradient to zero:

$$\nabla E = 0 \Longrightarrow A^T(A\underline{x}_H - \underline{x}_L) = 0 \Longleftrightarrow \sum_{n=1}^{K} W_n^T H_n^T D^T (DH_n W_n \underline{x}_H - \underline{x}_L^{(n)}) = 0$$

Gradient-based iterative methods can be used without explicit construction of these large matrices. Instead, the multiplication with A and $A^T A$ is implemented using only image operations such as warp, blur and sampling.

The matrix $A^T A$ operates on vectors $\underline{x}_H$, corresponding to an image of the size of the super resolution solution X_H,

$$A^T A\underline{x}_H = \sum_{n=1}^{K} W_n^T H_n^T D^T DW_n T_n \underline{x}_H$$

The matrix A^T operates on vectors $\underline{x}_L$, stacking of the input images $X_L^{(1)} \ldots X_L^{(K)}$ reordered in column vectors $\underline{x}_L^{(1)} \ldots \underline{x}_L^{(K)}$

$$A^T \underline{x}_L = \sum_{n=1}^{K} W_n^T H_n^T D^T \underline{x}_L^{(n)}$$

The matrices W_n, H_n, D model the image formation process, and their implementation is simply the image warping, blurring and subsampling respectively. The implementation of the transpose matrices is also very simple:

- D^T is implemented by upsampling the image without interpolation, i.e. by zero padding.
- H_n^T - For a convolution blur, this operation is implemented by convolution with the flipped kernel, i.e.if $h(i, j)$ is the imaging blur kernel, then the flipped kernel $\hat{h}$ satisfies $\forall i, j, \hat{h}(i, j) = h(-i, -j)$. For space-variant blur H_n^T is implemented by forward projection of the intensity values, using the weights of the original blur filter.
- W_n^T - If W_n is implemented by backward warping, then W_n^T should be the forward warping of the inverse motion.

The simplest implementation of this framework is using Richardson iterations [22], a from of steepest-descent with iteration step:

$$\underline{x}_H\{m+1\} = \underline{x}_H\{m\} + \sum_{n=1}^{K} W_n^T H_n^T D^T (\underline{x}_L^{(n)} - DH_n W_n \underline{x}_H\{m\})$$

This is a version of the Iterated Back Projection [8], using a specific blur kernel and forward warping in the back projection stage.

The ability to compute the gradient of the super resolution functional by image operations opens possibilities to efficiently use advanced gradient-based optimization techniques:

- Fast non-constrained optimization. An example is the Conjugate-Gradient method, which is elaborated in the following section.
- Constrained minimization, bounding the solution to a specific set. An example for this is the POCS solution [12, 18].
- Constrained minimization, using linear regularization term. The regularization operator should be also easily implemented using image operations. An example is given in the following section.

2.2. Super Resolution By the CG method

2.2.1 The Conjugate Gradient method. To demonstrate the practical benefit of computing the gradient by image operations, a super resolution algorithm using the conjugate-gradient (CG) method was implemented. The CG method is an efficient method to solve linear systems defined by symmetric positive definite matrices.

Definition 1 *Let Q be a symmetric and positive definite matrix. A vector set* $\{V_i\}_{i=1}^{n}$ *is Q-conjugate if*

$$\underline{V}_i^T Q \underline{V}_j = 0, \forall i \neq j.$$

A Q-conjugate set of vectors is linearly independent, and thus form a basis. The solution to the linear equation is therefore a linear combination of these vectors. The coefficients α_i are very easily found:

$$Q\underline{X} = \underline{Y} \Longrightarrow Q(\sum_{i=k}^{n} \alpha_k \underline{V}_k) = \underline{Y} \Longrightarrow \alpha_k = \frac{\underline{V}_k^T \underline{Y}}{\underline{V}_k^T Q \underline{V}_k}$$

The CG algorithm iteratively creates a conjugate basis, by converting the gradients computed in each iteration to vectors which are Q-conjugate to the previous ones (e.g. Graham-Shmidt procedure). Its convergence to the solution in *n* steps is guaranteed, but this is irrelevant to the super resolution problem, since *n*, the matrix size in super resolution, is huge. Still, the convergence rate of the CG is superior to steepest descent methods. Below is an implementation of CG.

$\mathrm{CG}(X, Y, Q, \epsilon, mMax)$

solves $QX = Y$, ϵ and *mMax* limit the number of iterations

1 $r = Y - QX, \rho_0 = \| r \|^2, m = 1$

2 Do While $\sqrt{\rho_{m-1}} > \epsilon \| Y \|^2$ and $m \leq kmax$

(a) if $m = 1$ then $p = r$
else $\beta = \frac{\rho_{m-1}}{\rho_{m-2}}$ and $p = r + \beta p$

(b) $w = Qp$

(c) $\alpha = \rho_{m-1}/p^T w$

(d) $x = x + \alpha p, r = r - \alpha w, \rho_m = \| r \|^2, m = m + 1$

2.2.2 CG super resolution. In the CG implementation to super resolution, the input includes the low resolution images, ϵ,*mMax* and the estimated blur function. First the motion between the input images is computed, and an initial estimate to the solution $X_H\{0\}$ is set, for example the average of the bilinearly upsampled and aligned input images.

Then, in order to use the CG code, two functions are implemented, project and backProject. The simple vector operations, such as inner product and multiplication with a scalar are easily translated to operations on images (correlation and multiplication by scalar). The matrix operations are handled in the following way:

- Step 1 - In the super resolution case $b = A^T Y$ and $Q = A^T A$. The code to compute the residual r is therefore:
  $r = 0$
  for n=1 to K do $r = r$ + backProject($X_L^{(n)}$-project($X_H\{0\}, n$), n)
- steps 2-b is replaced by the following code:
  w=0
  for n=1 to K do $w = w$ + backProject(project(p, n), n)

and the functions backProject,project are simply:

- I_3=project(p, n)
  - I_1 = blur(p, n) $\Longrightarrow$ blur image p by the blur operator H_n (e.g. convolution filter $h(i, j)$)
  - I_2 = backwardWrp(I_1, n) $\Longrightarrow$ Warp I_1 using backward warping, i.e. for each pixel in I_2, find its sub-pixel location in I_1,

based on the motion to the n-th image, and use interpolation to set its value.

- return subsample(I_2) $\Longrightarrow$ decimate the image, to get an image of the size of the input image $X_L^{(n)}$.

- I_1 =backProject(p, n)
  - I_2 = upsample(p) enlarge p to the size of the super resolved image, by zero padding.
  - I_1 = forwrdWrp (I_2, n) $\Longrightarrow$ Warp I_1 using forward warping, i.e. for each pixel in I_2, find its sub-pixel location in I_1, based on the motion to the n-th image. Spread the intensity value of the pixel on the pixels of I_2, proportionally to the interpolation coefficients.
  - return blur(I_1, n) $\Longrightarrow$ blur image p by the transpose of the blur operator H_n (e.g. in the case H is defined by a convolution filter $h(i, j)$, its transpose is implemented by convolution with the flipped filter $\hat{h}(i, j) = h(-i, -j)$)

2.2.3 Adding regularization. In many cases the super resolution does not have a unique solution, and the matrix $A^T A$ is not invertible. This can be solved by introducing constraints on the solution, e.g. smoothness. If the constraints f are differentiable, and their derivative can be approximated from the images, then they can be easily combined with our proposed framework, by minimizing:

$$E(\underline{x}_H) = \frac{1}{2}(\| \underline{x}_L - A\underline{x}_H \|^2 + \lambda f(\underline{x}_H))$$

Where λ is the regularization coefficient. Taking the derivative of E with respect to $\underline{x}_H$, results in a set of equations:

$$\nabla E = 0 \Longrightarrow A^T(A\underline{x}_L - \underline{x}_H) + \lambda \nabla f(\underline{x}_H) = 0$$

In each iteration the image corresponding to $\lambda \nabla f(\underline{x}_H\{m\})$ is added to the image corresponding to $A^T A\underline{x}_H\{m\})$, For example, when f can be expressed by a linear operator M:

$$\nabla f(\underline{x}_H) = M^T M \underline{x}_H$$

This image can be computed from the image corresponding to $\underline{x}_H$, by applying the operator M and its transpose. (The implementation of M^T in the image domain is derived similarly to the transpose of the blur operators). The selection of the optimal f and λ is beyond the scope of this paper [7].

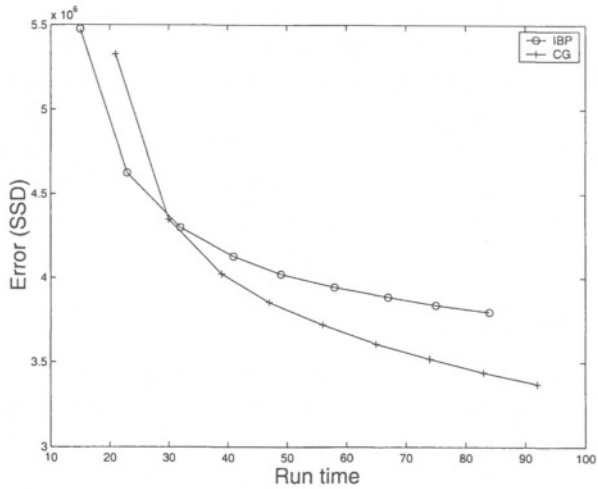


Figure 8.1. The sum of squared error in the images as a function of the running time (measured in seconds). The circles mark the iteration times of the CG algorithm, and crosses mark the iteration times of the IBP algorithm.

3. Computational Analysis and Results

To demonstrate the computational benefit of the proposed framework, the running time of the CG super resolution is compared to another image-based non-constrained algorithm, the IBP of Irani & Peleg. Images of a planar scene were captured by a hand held camera, and the projective-planar motion between them was computed. Then both methods of super resolution were applied, and the computation time and results were compared.

The graph in Figure 8.1 presents the projection error E as a function of the running time. The first iteration in the CG method is slower, since it requires additional multiplication with the matrix $A^T A$. The next iterations of both of the methods require a single multiplication with A and A^T, so the running time is similar (with small advantage to the CG method). This means that the comparison of the running time of these algorithm depends mainly on the convergence rate. It is notable in the graph that the convergence of the CG method in the first crucial iterations is much faster, yielding better results in a very short time. This can be further accelerated by using efficient image operations in the computation of the gradient.

The results of the super resolution algorithm are presented in Fig. 8.2. A set of images were captured by a hand-held camera. First the

motion between the images was computed [3]. Then, the proposed super resolution algorithm was applied (Fig. 8.2:E). For comparison, the images were enlarged and warped to a common coordinate system, and their median was computed (Fig. 8.2:C). Both the median and the SR improved the readability dramatically. The SR result is sharper than the median. After applying high-pass filter to the median results (Fig. 8.2:D), the readability is improved, but the result is not as good as the SR result.

4. Summary

Super resolution can be presented as a large sparse linear system. The presented framework allows for solving this system efficiently using image-domain operations. With the rapid advance in computing speed, applying super resolution algorithms on large video sequences becomes more and more practical. There is still work to be done in improving the noise model and the noise sensitivity of super resolution algorithms.

References

[1] M. Elad and A. Feuer, "Super-resolution reconstruction of continuous image sequence," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, September 1996.

[2] V.N. Smelyanskiy, P. Cheeseman, D. Maluf, and R. Morris, "Bayesian super-resolved surface reconstruction from images," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2000, pp. I:375–382.

[3] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *European Conf. on Computer Vision*, 1992, pp. 237–252.

[4] H.S. Sawhney and R. Kumar, "True multi-image alignment and its application to mosaicing and lens distortion correction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 3, pp. 245–243, March 1999.

[5] L. Zelnik-Manor and M. Irani, "Multi-frame estimation of planar motion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1105–1116, October 2000.

[6] A. Papoulis, "Generalized sampling expansion," *IEEE Trans. Circuits Systems*, vol. CAS-24, pp. 652–654, 1977.

[7] D. Capel and A. Zisserman, "Super-resolution enhancement of text image sequences," in *Int. Conf. Pattern Recognition*, 2000, pp. Vol I: 600–605.

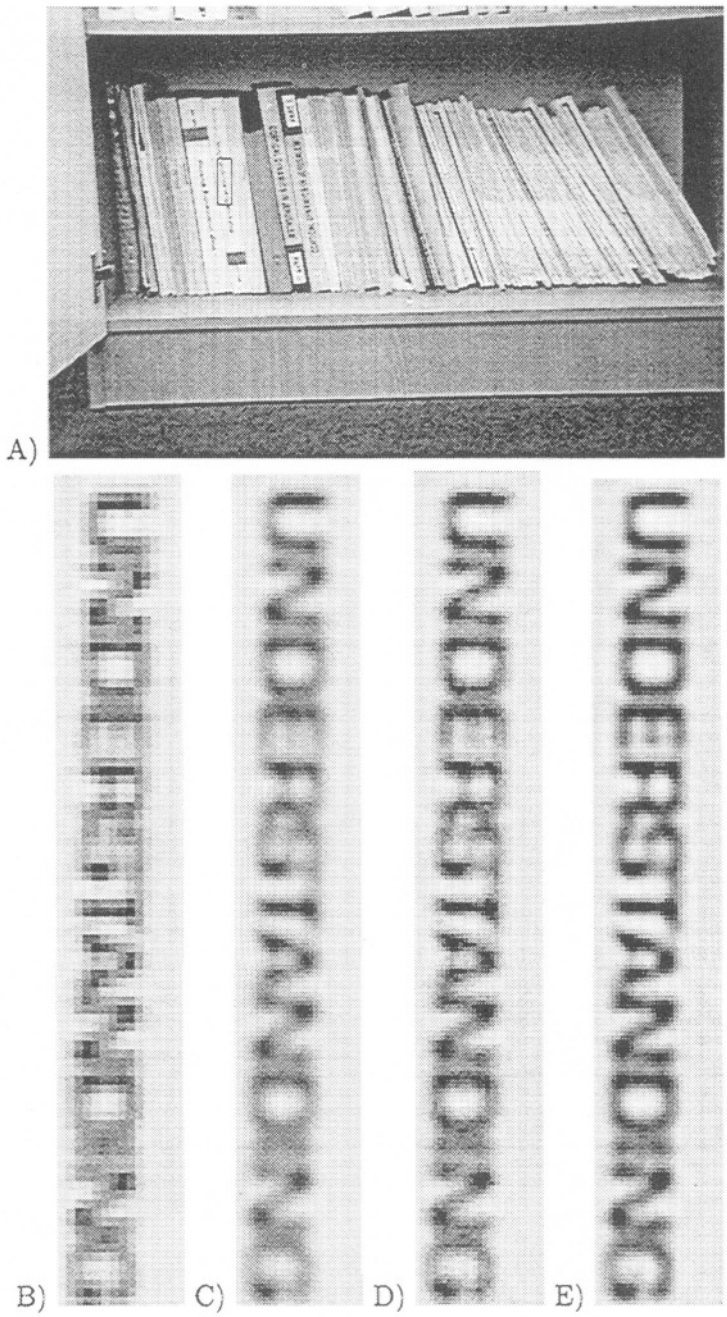


Figure 8.2. Super resolution results. A) An original frame. B) A small region from the original frame. The region is marked in Fig. A. C) The median of the aligned+enlarged images D) The results of applying high-pass filter to Fig. C. E) Super resolution.

[8] M. Irani and S. Peleg, "Improving resolution by image registration," *Graphical Models and Image Processing*, vol. 53, pp. 231–239, 1991.

[9] S. Mann and R.W. Picard, "Virtual bellows: Constructing high quality stills from video," in *Int. Conf. on Image Processing*, 1994.

[10] H. Shekarforoush and R. Chellappa, "Data-driven multichannel superresolution with application to video sequences," *Journal of the Optical Society of America*, vol. 16, no. 3, pp. 481–492, March 1999.

[11] M. Elad and A. Feuer, "Super-resolution reconstruction of image sequences," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 817–834, September 1999.

[12] A.J. Patti, M.I. Sezan, and A.M. Tekalp, "Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time," *IEEE Trans. Image Processing*, vol. 6, no. 8, pp. 1064–1076, August 1997.

[13] T.S. Huang and R.Y. Tsai, "Multi-frame image restoration and registration," in *Advances in Computer Vision and Image Processing*, T.S. Huang, Ed., vol. 1, pp. 317–339. JAI Press Inc., 1984.

[14] S. P. Kim, N. K. Bose, and H. M. Valenzuela, "Recursive reconstraction of high-resolution image from noisy undersampled frames," *IEEE Trans. Acoust., Speech and Sign. Proc.*, vol. 38, pp. 1013–1027, June 1990.

[15] H. Ur and D. Gross, "Improved resolution from subpixel shifted pictures," *Graphical Models and Image Processing*, vol. 54, no. 2, pp. 181–186, March 1992.

[16] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *Journal of Visual Communication and Image Representation*, vol. 4, pp. 324–335, 1993.

[17] D. Keren, S. Peleg, and R. Brada, "Image sequence enhancement using sub-pixel displacements," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1988, pp. 742–746.

[18] H. Stark and P. Oskoui, "High resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. Am. A*, vol. 6, no. 11, pp. 1715–1726, 1989.

[19] M. Berthod, H. Shekarforoush, M. Werman, and J. Zerubia, "Reconstruction of high resolution 3D visual information using subpixel camera displacements," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1994, pp. 654–657.

[20] R.R. Schultz and R.L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Trans. Image Processing*, vol. 5, no. 6, pp. 996–1011, June 1996.

[21] D. Capel and A. Zisserman, "Automated mosaicing with super-resolution zoom," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 1998, pp. 885–891.

[22] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, PA, 1995.

This page intentionally left blank

Chapter 9

SUPER-RESOLUTION FROM COMPRESSED VIDEO

C. Andrew Segall and Aggelos K. Katsaggelos
Department of Electrical and Computer Engineering
Northwestern University
Evanston, IL 60208-3118
{asegall,aggk}@ece.nwu.edu

Rafael Molina and Javier Mateos
Departamento de Ciencias de la Computatión e I.A.
Universidad de Granada
18071 Granada, Spain
{rms,jmd}@decsai.ugr.es

Abstract The problem of recovering a high-resolution frame from a sequence of low-resolution and compressed images is considered. The presence of the compression system complicates the recovery problem, as the operation reduces the amount of frequency aliasing in the low-resolution frames and introduces a non-linear noise process. Increasing the resolution of the decoded frames can still be addressed in a recovery framework though, but the method must also include knowledge of the underlying compression system. Furthermore, improving the spatial resolution of the decoded sequence is no longer the only goal of the recovery algorithm. Instead, the technique is also required to attenuate compression artifacts.

Key words: super-resolution, post-processing, image scaling, resolution enhancement, interpolation, spatial scalability, standards conversion, de-interlacing, video compression, image compression, motion vector constraint

1. INTRODUCTION

Compressed video is rapidly becoming the preferred method for video delivery. Applications such as Internet streaming, wireless videophones,

DVD players and HDTV devices all rely on compression techniques, and each requires a significant amount of data reduction for commercial viability. To introduce this reduction, a specific application often employs a low-resolution sensor or sub-samples the original image sequence. The reduced resolution sequence is then compressed in a lossy manner, which produces an estimate of the low-resolution data. For many tasks, the initial reconstruction of the compressed sequence is acceptable for viewing. However, when an application requires a high-resolution frame or image sequence, a super-resolution algorithm must be employed.

Super-resolution algorithms recover information about the original high-resolution image by exploiting sub-pixel shifts in the low-resolution data. These shifts are introduced by motion in the sequence and make it possible to observe samples from the high-resolution image that may not appear in a single low-resolution frame. Unfortunately, lossy encoding introduces several distortions that complicate the super-resolution problem. For example, most compression algorithms divide the original image into blocks that are processed independently. At high compression ratios, the boundaries between the blocks become visible and lead to "blocking" artifacts. If the coding errors are not removed, super-resolution techniques may produce a poor estimate of the high-resolution sequence, as coding artifacts may still appear in the high-resolution result. Additionally, the noise appearing in the decoded images may severely affect the quality of any of the motion estimation procedures required for resolution enhancement.

A straightforward solution to the problem of coding artifacts is to suppress any errors before resolution enhancement. The approach is appealing, as many methods for artifact removal are presented in the literature [1]. However, the sequential application of one of these post-processing algorithms followed by a super-resolution technique rarely provides a good result. This is caused by the fact that information removed during post-processing might be useful for resolution enhancement.

The formulation of a recovery technique that incorporates the tasks of post-processing and super-resolution is a natural approach to be followed. Several authors have considered such a framework, and a goal of this chapter is to review relevant work. Discussion begins in the next section, where background is presented on the general structure of a hybrid motion compensation and transform encoder. In Section 3, super-resolution methods are reviewed that derive fidelity constraints from the compressed bit-stream. In Section 4, work in the area of compression artifact removal is surveyed. Finally, a general framework for the super-resolution problem is proposed in Section 5. The result is a super-resolution algorithm for compressed video.

2. VIDEO COMPRESSION BASICS

The purpose of any image compression algorithm is to decrease the number of bits required to represent a signal. Loss-less techniques can always be employed. However for significant compression, information must be removed from the original image data. Many possible approaches are developed in the literature to intelligently remove perceptually unimportant content, and while every algorithm has its own nuances, most can be viewed as a three-step procedure. First, the intensities of the original images are transformed with a de-correlating operator. Then, the transform coefficients are quantized. Finally, the quantized coefficients are entropy encoded. The choice of the transform operator and quantization strategy are differentiating factors between techniques, and examples of popular operators include wavelets, Karhunen-Loeve decompositions and the Discrete Cosine Transform (DCT) [2]. Alternatively, both the transform and quantization operators can be incorporated into a single operation, which results in the technique of vector quantization [3].

The general approach for transform coding an MxN pixel image is therefore expressed as

$$\mathbf{x} = Q[\mathbf{T}\mathbf{g}], \tag{9.1}$$

where $\mathbf{g}$ is an (MN)x1 vector containing the ordered image, $\mathbf{T}$ is an (MN)x(MN) transformation matrix, Q is a quantization operator, and $\mathbf{x}$ is an (MN)x1 vector that contains the quantized coefficients. The quantized transform coefficients are then encoded with a loss-less technique and sent to the decoder.

At the standard decoder, the quantized information is extracted from any loss-less encoding. Then, an estimate of the original image is generated according to

$$\hat{\mathbf{g}} = \mathbf{T}^{-1} Q^{*}[\mathbf{x}], \tag{9.2}$$

where $\hat{\mathbf{g}}$ is the estimate of the original image, $\mathbf{T}^{-1}$ is the inverse of the transform operator, and Q^{*} represents a de-quantization operator. Note that the purpose of the de-quantization operator is to map the quantized values in $\mathbf{x}$ to transform coefficients. However, since the original quantization operator Q is a lossy procedure, this does not completely undo the information loss and $Q^{*}[Q[\mathbf{x}]] \neq \mathbf{x}$.

The compression method described in (9.1) and (9.2) forms the foundation for current transform-based compression algorithms. For example, the JPEG standard divides the original image into 8x8 blocks and

transforms each block with the DCT [4]. The transform coefficients are then quantized with a perceptually weighted method, which coarsely represents high-frequency information while maintaining low-frequency components. Next, the quantized values are entropy encoded and passed to the decoder, where multiplying the transmitted coefficients by the quantization matrix and computing the inverse-DCT reconstructs the image.

While transform coding provides a general method for two-dimensional image compression, its extension to video sequences is not always practical. As one approach, a video sequence might be encoded as a sequence of individual images. (If JPEG is utilized, this is referred to as motion-JPEG.) Each image is compressed with the transform method of (9.1), sent to a decoder, and then reassembled into a video sequence. Such a method clearly ignores the temporal redundancies between image frames. If exploited, these redundancies lead to further compression efficiencies. One way to capitalize on these redundancies is to employ a three-dimensional transform encoder [5, 6]. With such an approach, several frames of an image sequence are processed simultaneously with a three-dimensional transform operator. Then, the coefficients are quantized and sent to the decoder, where the group of frames is reconstructed. To realize significant compression efficiencies though, a large number of frames must be included in the transform. This precludes any application that is sensitive to the delay of the system.

A viable alternative to multi-dimensional transform coding is the hybrid technique of motion compensation and transform coding [7]. In this method, images are first predicted from previously decoded frames through the use of motion vectors. The motion vectors establish a mapping between the frame being encoded and previously reconstructed data. Using this mapping, the difference between the original image and its estimate can be calculated. The difference, or error residual, is then passed to a transform encoder and quantized. The entire procedure is expressed as

$$\mathbf{x} = Q\left[\mathbf{T}\left(\mathbf{g} - \hat{\mathbf{g}}^{\mathbf{MC}}\right)\right] , \tag{9.3}$$

where $\mathbf{x}$ is the quantized transform coefficients, and $\hat{\mathbf{g}}^{\mathbf{MC}}$ is the motion compensated estimate of $\mathbf{g}$ that is predicted from previously decoded data.

To decode the result, the quantized transform coefficients and motion vectors are transmitted to the decoder. At the decoder, an approximation of the original image is formed with a two-step procedure. First, the motion vectors are utilized to reconstruct the estimate. Then, the estimate is refined with the transmitted error residual. The entire procedure is express as

$$\hat{\mathbf{g}} = \mathbf{T}^{-1} Q^{*}\left[\mathbf{x}\right] + \hat{\mathbf{g}}^{\mathbf{MC}} , \tag{9.4}$$

where $\hat{\mathbf{g}}$ is the decoded image, $\hat{\mathbf{g}}^{MC}$ is uniquely defined by the motion vectors, and Q^* is the de-quantization operator.

The combination of motion compensation and transform coding provides a very practical compression algorithm. By exploiting the temporal correlation between frames, the hybrid method provides higher compression ratios than encoding every frame individually. In addition, compression gains do not have to come with an explicit introduction of delay. Instead, motion vectors can be restricted to only reference previous frames in the sequence, which allows each image to be encoded as it becomes available to the encoder. When a slight delay is acceptable though, more sophisticated motion compensation schemes can be employed that utilize future frames for a bi-directional motion estimate [8].

The utility of motion estimation and transform coding makes it the backbone of current video-coding standards. These standards include MPEG-1, MPEG-2, MPEG-4, H.261 and H.263 [9-14]. In each of the methods, the original image is first divided into blocks. The blocks are then encoded using one of two available methods. For an intra-coded block, the block is transformed by the DCT and quantized. For inter-coded blocks, motion vectors are first found to estimate the current block from previously decoded images. This estimate is then subtracted from the current block, and the residual is transformed and quantized. The quantization and motion vector data is sent to the decoder, which estimates the original image from the transmitted coefficients.

The major difference between the standards lies in the representation of the motion vectors and quantizers. For example, motion vectors are signaled at different resolutions in the standards. In H.261, a motion vector is represented with an integer number of pixels. This is different from the methods employed for MPEG-1, MPEG-2 and H.263, where the motion vectors are sent with half-pixel accuracy and an interpolation procedure is defined for the estimate. MPEG-4 utilizes more a sophisticated method for representing the motion, which facilitates the transmission of motion vectors at quarter-pixel resolution.

Other differences also exist between the standards. For example, some standards utilize multiple reference frames or multiple motion vectors for the motion compensated prediction. In addition, the structure and variability of the quantizer is also different. Nevertheless, for the purposes of developing a super-resolution algorithm, it is sufficient to remember that quantization and motion estimation data will always be provided in the bit-stream. When a portion of a sequence is intra-coded, the quantizer and transform operators will express information about the intensities of the original image. When blocks are inter-coded, motion vectors will provide an (often crude) estimate of the motion field.

3. INCORPORATING THE BIT-STREAM

With a general understanding of the video compression process, it is now possible to incorporate information from a compressed bit-stream into a super-resolution algorithm. Several methods for utilizing this information have been presented in the literature, and a survey of these techniques is presented in this section. At the high-level, these methods can be classified according to the information extracted from the bit-stream. The first class of algorithms incorporates the quantization information into the resolution enhancement procedure. This data is transmitted to the decoder as a series of indices and quantization factors. The second class of algorithms incorporates the motion vectors into the super-resolution algorithm. These vectors appear as offsets between the current image and previous reconstructions and provide a degraded observation of the original motion field.

3.1 System Model

Before incorporating parameters from the bit-stream into a super-resolution algorithm, a definition of the system model is necessary. This model is utilized in all of the proposed methods, and it relates the original high-resolution images to the decoded low-resolution image sequence. Derivation of the model begins by generating an intermediate image sequence according to

$$\mathbf{g} = \mathbf{AHf}\,, \tag{9.5}$$

where $\mathbf{f}$ is a $(PMPN)$x1 vector that represents a (PM)x(PN) high-resolution image, $\mathbf{g}$ is an (MN)x1 vector that contains the low-resolution data, $\mathbf{A}$ is an (MN)x$(PMPN)$ matrix that realizes a sub-sampling operation and $\mathbf{H}$ is a $(PMPN)$x$(PMPN)$ filtering matrix.

The low-resolution images are then encoded with a video compression algorithm. When a standards compliant encoder is assumed, the low-resolution images are processed according to (9.3) and (9.4). Incorporating the relationship between low and high-resolution data in (9.5), the compressed observation becomes

$$\hat{\mathbf{g}} = \mathbf{T}_{DCT}^{-1} Q^{*}\left[\, Q\left[\mathbf{T}_{DCT}\left(\mathbf{AHf} - \hat{\mathbf{g}}^{\mathbf{MC}}\right)\right]\right] + \hat{\mathbf{g}}^{\mathbf{MC}}\,, \tag{9.6}$$

where $\hat{\mathbf{g}}$ is the decoded low-resolution image, $\mathbf{T}_{DCT}$ and $\mathbf{T}_{DCT}^{-1}$ are the forward and inverse DCT operators, respectively, Q and Q^{*} are the quantization and

de-quantization operators, respectively, and $\hat{\mathbf{g}}^{\mathbf{MC}}$ is the temporal prediction of the current frame based on the motion vectors. If a portion of the image is encoded without motion compensation (i.e. intra-blocks), then the predicted values for that region are zero.

Equation (9.6) defines the relationship between a high-resolution frame and a compressed frame for a given time instance. Now, the high-resolution frames of a dynamic image sequence are also coupled through the motion field according to

$$\mathbf{f}_l = \mathbf{C}_{l,k}\,\mathbf{f}_k, \tag{9.7}$$

where $\mathbf{f}_l$ and $\mathbf{f}_k$ are $(PMPN)$x1 vectors that denote the high-resolution data at times l and k, respectively, and $\mathbf{C}_{l,k}$ is a $(PMPN)$x$(PMPN)$ matrix that describes the motion vectors relating the pixels at time k to the pixels at time l. These motion vectors describe the actual displacement between high-resolution frames, which should not be confused with the motion information appearing in the bit-stream. For regions of the image that are occluded or contain objects entering the scene, the motion vectors are not defined.

Combining (9.6) and (9.7) produces the relationship between a high-resolution and compressed image sequence at different time instances. This relationship is given by

$$\hat{\mathbf{g}}_l = \mathbf{T}_{DCT}^{-1} Q^{*}\left[\, Q\left[\mathbf{T}_{DCT}\left(\mathbf{AHC}_{l,k}\mathbf{f}_k - \hat{\mathbf{g}}_l^{\mathbf{MC}}\right)\right]\right] + \hat{\mathbf{g}}_l^{\mathbf{MC}}, \tag{9.8}$$

where $\hat{\mathbf{g}}_l$ is the compressed frame at time l and $\hat{\mathbf{g}}_l^{\mathbf{MC}}$ is the motion compensated prediction utilized in generating the compressed observation.

3.2 Quantizers

To explore the quantization information that is provided in the bit-stream, researchers represent the quantization procedure with an additive noise process according to

$$\mathbf{T}_{DCT}^{-1} Q^{*}\left[\, Q\left[\mathbf{T}_{DCT}\left(\mathbf{AHC}_{l,k}\mathbf{f}_k - \hat{\mathbf{g}}_l^{\mathbf{MC}}\right)\right]\right] = \mathbf{AHC}_{l,k}\mathbf{f}_k - \hat{\mathbf{g}}_l^{\mathbf{MC}} + \mathbf{n}_l^{Q}, \tag{9.9}$$

where $\mathbf{n}_l^{Q}$ represents the quantization noise at time l. The advantage of this representation is that the motion compensated estimates are eliminated from the system model, which leads to super-resolution methods that are independent of the underlying motion compensation scheme. Substituting

(9.9) into (9.8), the relationship between a high-resolution image and the low-resolution observation becomes

$$\hat{\mathbf{g}}_l = \mathbf{AHC}_{l,k}\mathbf{f}_k + \mathbf{n}_l^Q, \tag{9.10}$$

where the motion compensated estimates in (9.8) cancel out.

With the quantization procedure represented as a noise process, a single question remains: What is the structure of the noise? To understand the answers proposed in the literature, the quantization procedure must first be understood. In standards based compression algorithms, quantization is realized by dividing each transform coefficient by a quantization factor. The result is then rounded to the nearest integer. Rounding discards data from the original image sequence, and it is the sole contributor to the noise term of (9.10). After rounding, the encoder transmits the integer index and the quantization factor to the decoder. The transform coefficient is then reconstructed by multiplying the two transmitted values, that is

$$T_{DCT}(\hat{\mathbf{g}}, i) = q(i)x(i) = q(i)\cdot \mathrm{Round}\left(\frac{T_{DCT}(\mathbf{g}, i)}{q(i)}\right), \tag{9.11}$$

where $T_{DCT}(\mathbf{g}, i)$ and $T_{DCT}(\hat{\mathbf{g}}, i)$ denote the i^{th} transform coefficient of the low-resolution image $\mathbf{g}$ and the decoded estimate $\hat{\mathbf{g}}$, respectively, $q(i)$ is the quantization factor and $x(i)$ is the index transmitted by the encoder for the i^{th} transform coefficient, and Round(·) is an operator that maps each value to the nearest integer.

Equation (9.11) defines a mapping between each transform coefficient and the nearest multiple of the quantization factor. This provides a key constraint, as it limits the quantization error to half of the quantization factor. With knowledge of the quantization error bounds, a set-theoretic approach to the super-resolution problem is explored in [15]. The method restricts the DCT coefficients of the solution to be within the uncertainty range signaled by the encoder. The process begins by defining the constraint set

$$\hat{\mathbf{f}}_k \in \left\{\hat{\mathbf{f}}_k : -\frac{\mathbf{q}_l}{2} \le \mathbf{T}_{DCT}\left(\mathbf{AHC}_{l,k}\hat{\mathbf{f}}_k - \hat{\mathbf{g}}_l\right) \le \frac{\mathbf{q}_l}{2}\right\}, \tag{9.12}$$

where $\hat{\mathbf{f}}_k$ is the high-resolution estimate, $\mathbf{q}_l$ is a vector that contains the quantization factors for time l, $\hat{\mathbf{g}}_l$ is estimated by choosing transform coefficients centered on each quantization interval, and the less-than operator is defined on an element by element basis. Finding a solution that

satisfies (9.12) is then accomplished with a Projection onto Convex Sets (POCS) iteration, where the projection of $\hat{\mathbf{f}}_k$ onto the set is defined as

$$P_l\left[\hat{\mathbf{f}}_k\right]=\begin{cases}\hat{\mathbf{f}}_k-\dfrac{\mathbf{C}_{l,k}^T\mathbf{H}^T\mathbf{A}^T\mathbf{T}_{DCT}^{-1}\left\{\mathbf{T}_{DCT}\mathbf{AHC}_{l,k}\hat{\mathbf{f}}_k-\left(\mathbf{T}_{DCT}\hat{\mathbf{g}}_l+.5\mathbf{q}_l\right)\right\}}{\left\|\mathbf{T}_{DCT}\mathbf{AHC}_{l,k}\right\|^2}, & \\ & \mathbf{T}_{DCT}\left(\mathbf{AHC}_{l,k}\hat{\mathbf{f}}_k-\hat{\mathbf{g}}_l\right)>.5\mathbf{q}_l \\ \hat{\mathbf{f}}_k-\dfrac{\mathbf{C}_{l,k}^T\mathbf{H}^T\mathbf{A}^T\mathbf{T}_{DCT}^{-1}\left\{\mathbf{T}_{DCT}\mathbf{AHC}_{l,k}\hat{\mathbf{f}}_k-\left(\mathbf{T}_{DCT}\hat{\mathbf{g}}_l-.5\mathbf{q}_l\right)\right\}}{\left\|\mathbf{T}_{DCT}\mathbf{AHC}_{l,k}\right\|^2}, & \\ & \mathbf{T}_{DCT}\left(\mathbf{AHC}_{l,k}\hat{\mathbf{f}}_k-\hat{\mathbf{g}}_l\right)<-.5\mathbf{q}_l \\ \hat{\mathbf{f}}_k, & \textit{otherwise}\end{cases} \tag{9.13}$$

where $P_l[\hat{\mathbf{f}}_k]$ is the projection operator that accounts for the influence of the observation $\hat{\mathbf{g}}_l$ on the estimate of the high-resolution image $\hat{\mathbf{f}}_k$.

The set-theoretic method is well suited for limiting the magnitude of the quantization errors in a system model. However, the projection operator does not encapsulate any additional information about the shape of the noise process within the bounded range. When information about the structure of the noise is available, then an alternative description may be more appropriate. One possible method is to utilize probabilistic descriptions of the quantization noise in the transform domain and rely on maximum *a posteriori* or maximum likelihood estimates for the high-resolution image. This approach is considered in [16], where the quantization noise is represented with the density

$$p_{\mathbf{N}}\left(\mathbf{n}_k^Q\right)=\frac{p_{\tilde{\mathbf{N}}}\left(\mathbf{T}_{DCT}\mathbf{n}_k^Q\right)}{\left|\mathbf{T}_{DCT}^{-1}\right|}, \tag{9.14}$$

where $\mathbf{n}_k^Q$ is the quantization noise in the spatial domain, $|\mathbf{T}_{DCT}^{-1}|$ is the determinant of the transform operator, and $p_{\mathbf{N}}(\cdot)$ and $p_{\tilde{\mathbf{N}}}(\cdot)$ denote the probability density functions in the spatial and transform domains, respectively [17].

Finding a simple expression for the quantization noise in the spatial domain is often difficult, and numerical solutions are employed in [16]. However, an important case is considered in [18, 19], where the quantization noise is expressed with the Gaussian distribution

$$p_{\mathbf{N}}\left(\mathbf{n}_k^{Q}\right) = Z \exp\left\{-\frac{1}{2}\left(\mathbf{n}_k^{Q}\right)^{T}\left(\mathbf{T}_{DCT}\bar{\mathbf{K}}_k^{Q}\mathbf{T}_{DCT}^{-1}\right)^{-1}\left(\bar{\mathbf{n}}_k^{Q}\right)\right\}, \tag{9.15}$$

where $\bar{\mathbf{K}}_k^{Q}$ is the covariance matrix of the quantization noise in the transform domain for the k^{th} frame of the sequence, and Z is a normalizing constant.

Several observations pertaining to (9.15) are appropriate. First, notice that if the distributions for the quantization noise in the transform domain are independent and identically distributed, then $p_{\mathbf{N}}(\mathbf{n}_k^{Q})$ is spatially uncorrelated and identically distributed. This arises from the structure of the DCT and is representative of the flat quantization matrices typically used for inter-coding. As a second observation, consider the perceptually weighted quantizers that are utilized for intra-coding. In this quantization strategy, high-frequency coefficients are represented with less fidelity. Thus, the distribution of the noise in the DCT domain depends on the frequency. When the quantization noise is independent in the transform domain, then $p_{\mathbf{N}}(\mathbf{n}_k^{Q})$ will be spatially correlated.

Incorporating the quantizer information into a super-resolution algorithm should improve the results, as it equips the procedure with knowledge of the non-linear quantization process. In this section, three approaches to utilizing the quantizer data have been considered. The first method enforces bounds on the quantization noise, while the other methods employ a probabilistic description of the noise process. Now that the proposed methods have been presented, the second component of incorporating the bit-stream can be considered. In the next sub-section, methods that utilize the motion vectors are presented.

3.3 Motion Vectors

Incorporating the motion vectors into the resolution enhancement algorithm is also an important problem. Super-resolution techniques rely on sub-pixel relationships between frames in an image sequence. This requires a precise estimate of the actual motion, which has to be derived from the observed low-resolution images. When a compressed bit-stream is available though, the transmitted motion vectors provide additional information about the underlying motion. These vectors represent a degraded observation of the actual motion field and are generated by a motion estimation algorithm within the encoder.

Several traits of the transmitted motion vectors make them less than ideal for representing actual scene motion. As a primary flaw, motion vectors are not estimated at the encoder by utilizing the original low-resolution frames. Instead, motion vectors establish a correspondence between the current low-resolution frame and compressed frames at other time instances. When the

compressed frames represent the original image accurately, then the correlation between the motion vectors and actual motion field is high. As the quality of compressed frames decreases, the usefulness of the motion vectors for estimating the actual motion field is diminished.

Other flaws also degrade the compressed observation of the motion field. For example, motion estimation is a computationally demanding procedure. When operating under time or resource constraints, an encoder often employs efficient estimation techniques. These techniques reduce the complexity of the algorithm but also decrease the reliability of the motion vectors. As a second problem, motion vectors are transmitted with a relatively coarse sampling. At best, one motion vector is assigned to every 8x8 block in a standards compliant bit-stream. Super-resolution algorithms, however, require a much denser representation of motion.

Even with the inherent errors in the transmitted motion vectors, methods have been proposed that capitalize on the transmitted information. As a first approach, a super-resolution algorithm that estimates the motion field by refining the transmitted data is proposed in [18, 19]. This is realized by initializing a motion estimation algorithm with the transmitted motion vectors. Then, the best match between decoded images is found within a small region surrounding each initial value. With the technique, restricting the motion estimate adds robustness to the search procedure. More importantly, the use of a small search area greatly reduces the computational requirements of the motion estimation method.

A second proposal does not restrict the motion vector search [20, 21]. Instead, the motion field can contain a large deviation from the transmitted data. In the approach, a similarity measure between each candidate solution and the transmitted motion vector is defined. Then, motion estimation is employed to minimize a modified cost function. Using the Euclidean distance as an example of similarity, the procedure is expressed as

$$\hat{\mathbf{C}}_{l,k} = \arg\min_{\mathbf{C}_{l,k}} \left\{ \left\| \mathbf{AHC}_{l,k} \hat{\mathbf{f}}_k - \hat{\mathbf{g}}_l \right\|^2 + \lambda \sum_{i=0}^{MN-1} \left\| \mathbf{c}_{l,k}(i) - A_{MV}^T \left(\mathbf{C}_{l,k}^{Encoder}, i \right) \right\|^2 \right\}, \tag{9.16}$$

where $\hat{\mathbf{C}}_{l,k}$ is a matrix that represents the estimated motion field, $\mathbf{c}_{l,k}(i)$ is a two-dimensional vector the contains the motion vector for pixel location i, $\mathbf{C}_{l,k}^{Encoder}$ is a matrix that contains the motion vectors provided by the encoder, $A_{MV}^T(\mathbf{C}_{l,k}^{Encoder}, i)$ produces an estimate for the motion at pixel location i from the transmitted motion vectors, and λ quantifies the confidence in the transmitted information.

In either of the proposed methods, an obstacle to incorporating the transmitted motion vectors occurs when motion information is not provided

for the frames of interest. In some cases, such as intra-coded regions, the absence of motion vectors may indicate an occlusion. In most scenarios though, the motion information is simply being signaled in an indirect way. For example, an encoder may provide the motion estimates $\mathbf{C}_{l,l'}^{Encoder}$ and $\mathbf{C}_{l',k}^{Encoder}$, while not explicitly transmitting $\mathbf{C}_{l,k}^{Encoder}$. When a super-resolution algorithm needs to estimate $\hat{\mathbf{C}}_{l,k}$, the method must determine $\mathbf{C}_{l,k}^{Encoder}$ from the transmitted information. For vectors with pixel resolution, a straightforward approach is to add the horizontal and vertical motion components to find the mapping $\mathbf{C}_{l,k}^{Encoder}$. The confidence in the estimate must also be adjusted, as adding the transmitted motion vectors increases the uncertainty of the estimate. In the method of [18, 19], a lower confidence in $\mathbf{C}_{l,k}^{Encoder}$ results in a larger search area when finding the estimated motion field. In [20, 21], the decreased confidence results in smaller values for λ.

4. COMPRESSION ARTIFACTS

Exploring the influence of the quantizers and motion vectors is the first step in developing a super-resolution algorithm for compressed video. These parameters convey important information about the original image sequence, and each is well suited for restricting the solution space of a high-resolution estimate. Unfortunately, knowledge of the compressed bit-stream does not address the removal of compression artifacts. Artifacts are introduced by the structure of an encoder and must also be considered when developing a super-resolution algorithm. In this section, an overview of post-processing methods is presented. These techniques attenuate compression artifacts in the decoded image and are an important component of any super-resolution algorithm for compressed video. In the next sub-section, an introduction to various compression artifacts is presented. Then, three techniques for attenuating compression artifacts are discussed.

4.1 Artifact Types

Several artifacts are commonly identified in video coding. A first example is blocking. This artifact is objectionable and annoying at all bit-rates of practical interest, and it is most bothersome as the bit-rate decreases. In a standards based system, blocking is introduced by the structure of the encoder. Images are divided into equally sized blocks and transformed with a de-correlating operator. When the transform considers each block independently, pixels outside of the block region are ignored and the continuity across boundaries is not captured. This is perceived as a

synthetic, grid-like error at the decoder, and sharp discontinuities appear between blocks in smoothly varying regions.

Blocking errors are also introduced by poor quantization decisions. Compression standards do not define a strategy for allocating bits within a bit-stream. Instead, the system designer has complete control. This allows for the development of encoders for a wide variety of applications, but it also leads to artifacts. As an example, resource critical applications typically rely on heuristic allocation strategies. Very often different quantizers may be assigned to neighboring regions even though they have similar visual content. The result is an artificial boundary in the decoded sequence.

Other artifacts are also attributed to the improper allocation of bits. In satisfying delay constraints, encoders operate without knowledge of future sequence activity. Thus, bits are distributed on an assumption of future content. When the assumption is invalid, an encoder must quickly adjust the amount of quantization to satisfy a given rate constraint. The encoded video sequence possesses a temporally varying image quality, which manifests itself as a temporal flicker.

Edges and impulsive features introduce a final coding error. Represented in the frequency domain, these signals have high spatial frequency content. Quantization removes some of the information for encoding and introduces quantization error. However, when utilizing a perceptually weighted technique, additional errors appear. Low frequency data is preserved, while high frequency information is coarsely quantized. This removes the high-frequency components of the edge and introduces a strong *ringing artifact* at the decoder. In still images, the artifact appears as strong oscillations in the original location of the edge. Image sequences are also plagued by ringing artifacts but are usually referred to as *mosquito* errors.

4.2 Post-processing Methods

Post-processing methods are concerned with removing all types of coding errors and are directly applicable to the problem of super-resolution. As a general framework, post-processing algorithms attenuate compression artifacts by developing a model for spatial and temporal properties of the original image sequence. Then, post-processing techniques find a solution that satisfies the ideal properties while also remaining faithful to the available data.

One approach for post-processing follows a constrained least squares (CLS) methodology [22-24]. In this technique, a penalty function is assigned to each artifact type. The post-processed image is then found by minimizing the following cost functional

$$E(\mathbf{p}) = \left\| \mathbf{p} - \hat{\mathbf{g}} \right\|^2 + \lambda_1 \left\| \mathbf{B}\mathbf{p} \right\|^2 + \lambda_2 \left\| \mathbf{R}\mathbf{p} \right\|^2 + \lambda_3 \left\| \mathbf{p} - \hat{\mathbf{g}}^{\mathbf{MC}} \right\|^2 , \qquad (9.17)$$

where $\mathbf{p}$ is a vector representing the post-processed image, $\hat{\mathbf{g}}$ is the estimate decoded from the bit-stream, $\mathbf{B}$ and $\mathbf{R}$ are matrices that penalize the appearance of blocking and ringing, respectively, $\hat{\mathbf{g}}^{\mathbf{MC}}$ is the motion compensated prediction and λ_1, λ_2, and λ_3 express the relative importance of each constraint. In practice, the matrix $\mathbf{B}$ is implemented as a difference operator across the block boundaries, while the matrix $\mathbf{R}$ describes a high-pass filter within each block.

Finding the derivative of (9.17) with respect to $\mathbf{p}$ and setting it to zero represents the necessary condition for a minimum of (9.17). A solution is then found using the method of successive approximations according to

$$\mathbf{p}^{k+1} = \mathbf{p}^k - \alpha \left\{ \mathbf{p}^k - \hat{\mathbf{g}} + \lambda_1 \mathbf{B}^T \mathbf{B}\mathbf{p}^k + \lambda_2 \mathbf{R}^T \mathbf{R}\mathbf{p}^k + \lambda_3 \left(\mathbf{p}^k - \hat{\mathbf{g}}^{\mathbf{MC}} \right) \right\} , \quad (9.18)$$

where α determines the convergence and rate of convergence of the algorithm, and $\mathbf{p}^k$ and $\mathbf{p}^{k+1}$ denote the post-processed solution at iteration k and $k+1$, respectively [25]. The decoded image is commonly defined as the initial estimate, $\mathbf{p}^0$. Then, the iteration continues until a termination criterion is satisfied.

Selecting the smoothness constraints ($\mathbf{B}$ and $\mathbf{R}$) and parameters (λ_1, λ_2 and λ_3) defines the performance of the CLS technique, and many approaches have been developed for compression applications. As a first example, parameters can be calculated at the encoder from the intensity data of the original images, transmitted through a side channel and supplied to the post-processing mechanism [26]. More appealing techniques vary the parameters relative to the contents of the bit-stream, incorporating the quantizer information and coding modes into the choice of parameters [27-29].

Besides the CLS approach, other recovery techniques are also suitable for post-processing. In the framework of POCS, blocking and ringing artifacts are removed by defining images sets that do not exhibit compression artifacts [30, 31]. For example, the set of images that are smooth would not contain ringing artifacts. Similarly, blocking artifacts are absent from all images with smooth block boundaries. To define the set, the amount of smoothness must be quantified. Then, the solution is constrained by

$$\mathbf{p} \in \left\{ \mathbf{g} : \left\| \mathbf{B}\mathbf{g} \right\|^2 \le T_B \right\} , \qquad (9.19)$$

where T_B is the smoothness threshold used for the block boundaries and $\mathbf{B}$ is a difference operator between blocks.

An additional technique for post-processing relies on the Bayesian framework. In the method, a post-processed solution is computed as a maximum *a posteriori* (MAP) estimate of the image sequence presented to the encoder, conditioned on the observation [32, 33]. Thus, after applying Bayes' rule, the post-processed image is given by

$$\mathbf{p} = \arg\max_{\mathbf{g}} \frac{p(\hat{\mathbf{g}} \mid \mathbf{p}) p(\mathbf{p})}{p(\hat{\mathbf{g}})}. \tag{9.20}$$

Taking logarithms, the technique becomes

$$\mathbf{p} = \arg\max_{\mathbf{p}} \log p(\hat{\mathbf{g}} \mid \mathbf{p}) + \log(\mathbf{p}), \tag{9.21}$$

where $p(\hat{\mathbf{g}} \mid \mathbf{p})$ is often assumed constant within the bounds of the quantization error.

Compression artifacts are removed by selecting a distribution for the post-processed image with few compression errors. One example is the Gaussian distribution

$$p(\mathbf{g}) = \exp\left\{-\lambda_1 \|\mathbf{B}\mathbf{g}\|^2 - \lambda_2 \|\mathbf{R}\mathbf{g}\|^2\right\}. \tag{9.22}$$

In this expression, images that are likely to contain artifacts are assigned a lower probability of occurrence. This inhibits the coding errors from appearing in the post-processed solution.

5. SUPER-RESOLUTION

Post-processing methods provide the final component of a super-resolution approach. In the previous section, three techniques are presented for attenuating compression artifacts. Combining these methods with the work in Section 3 produces a complete formulation of the super-resolution problem. This is the topic of the current section, where a concise formulation for the resolution enhancement of compressed video is proposed. The method relies on the MAP estimation techniques to address compression artifacts as well as to incorporate the motion vectors and quantizer data from the compressed bit-stream.

5.1 MAP Framework

The goal of the proposed super-resolution algorithm is to estimate the original image sequence and motion field from the observations provided by the encoder. Within the MAP framework, this joint estimate is expressed as

$$\hat{\mathbf{f}}_k, \hat{\mathbf{D}}_{TB,TF} = \arg\max_{\mathbf{f}_k, \mathbf{D}_{TB,TF}} \left\{ p\left(\mathbf{f}_k, \mathbf{D}_{TB,TF} \mid \mathbf{G}, \mathbf{D}_{TB,TF}^{Encoder}\right) \right\}$$
$$= \arg\max_{\mathbf{f}_k, \mathbf{D}_{TB,TF}} \left\{ \frac{p\left(\mathbf{G}, \mathbf{D}_{TB,TF}^{Encoder} \mid \mathbf{f}_k, \mathbf{D}_{TB,TF}\right) p\left(\mathbf{f}_k, \mathbf{D}_{TB,TF}\right)}{p\left(\mathbf{G}, \mathbf{D}_{TB,TF}^{Encoder}\right)} \right\}, \quad (9.23)$$

where $\hat{\mathbf{f}}_k$ is the estimate for the high-resolution image, $\mathbf{G}$ is an (MN)xL matrix that contains the L compressed observations $\hat{\mathbf{g}}_{k-TB}, \ldots, \hat{\mathbf{g}}_{k+TF}$, TF and TB are the number of frames contributing to the estimate in the forward and backward direction of the temporal axis, respectively, and $\mathbf{D}_{TB,TF}$ and $\mathbf{D}_{TB,TF}^{Encoder}$ are formed by lexicographically ordering the respective motion vectors $\mathbf{C}_{k-TB,k}, \ldots, \mathbf{C}_{k+TF,k}$ and $\mathbf{C}_{k-TB,k}^{Encoder}, \ldots, \mathbf{C}_{k+TF,k}^{Encoder}$ into vectors and storing the result in a $(PMPN)$x$(TF+TB+1)$ matrix.

5.1.1 Fidelity Constraints

Definitions for the conditional distributions follow from the previous sections. As a first step, it is assumed that the decoded intensity values and transmitted motion vectors are independent. This results in the conditional density

$$p\left(\mathbf{G}, \mathbf{D}_{TB,TF}^{Encoder} \mid \mathbf{f}_k, \mathbf{D}_{TB,TF}\right) = p\left(\mathbf{G} \mid \mathbf{f}_k, \mathbf{D}_{TB,TF}\right) p\left(\mathbf{D}_{TB,TF}^{Encoder} \mid \mathbf{f}_k, \mathbf{D}_{TB,TF}\right) \quad (9.24)$$

Information from the encoder is then included in the algorithm. The density function $p(\mathbf{G} \mid \mathbf{f}_k, \mathbf{D}_{TB,TF})$ describes the noise that is introduced during quantization, and it can be derived through the mapping presented in (9.14). The corresponding conditional density is

$$p\left(\mathbf{G} \mid \mathbf{f}_k, \mathbf{D}_{TB,TF}\right) \propto \exp\left\{ -\frac{1}{2} \sum_{l=k-TB}^{k+TF} \left(\mathbf{AHC}_{l,k}\mathbf{f}_k - \hat{\mathbf{g}}_l\right)^T \mathbf{K}_l^{-1} \left(\mathbf{AHC}_{l,k}\mathbf{f}_k - \hat{\mathbf{g}}_l\right) \right\}, \quad (9.25)$$

when $\hat{\mathbf{g}}_l$ is the decoded image at time instant l and $\mathbf{K}_l$ is the noise covariance matrix in the spatial domain that is found by modeling the noise in the transform domain as Gaussian distributed and uncorrelated.

The second conditional density relates the transmitted motion vectors to the original motion field. Following the technique appearing in (9.16), an example distribution is

$$p\left(\mathbf{D}_{TB,TF}^{Encoder} \mid \mathbf{f}_k, \mathbf{D}_{TB,TF}\right) \propto \exp\left\{ -\gamma \sum_{l=k-TB}^{k+TF} \sum_{i=0}^{MN-1} \left\| \mathbf{c}_{l,k}(i) - A_{MV}^{T}\left(\mathbf{C}_{l,k}^{Encoder}, i\right) \right\|^2 \right\}, \tag{9.26}$$

where $\mathbf{c}_{l,k}(i)$ is the motion vector for pixel location i, $A_{MV}^{T}(\mathbf{C}_{l,k}^{Encoder}, i)$ estimates the motion at pixel i from the transmitted motion vectors, and γ is a positive value that expresses a confidence in the transmitted vectors.

As a final piece of information from the decoder, bounds on the quantization error should be exploited. These bounds are known in the transform domain and express the maximum difference between DCT coefficients in the original image and in the decoded data. High-resolution estimates that exceed these values are invalid solutions to the super-resolution problem, and the MAP estimate must enforce the constraint. This is accomplished by restricting the solution space so that

$$\mathbf{f}_k \in \left\{ \mathbf{f}_k : \mathbf{T}_{DCT}\left(\mathbf{A}\mathbf{H}\mathbf{C}_{l,k}\,\mathbf{f}_k - \hat{\mathbf{g}}_l\right) < \frac{\mathbf{q}_l}{2},\ l = k - TB, \ldots, k + TF \right\}, \tag{9.27}$$

where $\mathbf{q}_l$ is the vector defined in (9.12) containing the quantization factors for time l

5.1.2 Prior Models

After incorporating parameters from the compressed bit-stream into the recovery procedure, the prior model $p(\mathbf{f}_k, \mathbf{D}_{TB,TF})$ is defined. Assuming that the intensity values of the high-resolution image and the motion field are independent, the distribution for the original, high-resolution image can be utilized to attenuate compression artifacts. Borrowing from work in post-processing, the distribution

$$p(\mathbf{f}_k) \propto \exp\left\{ -\left(\lambda_1 \left\| \mathbf{B}\,\mathbf{f}_k \right\|^2 + \lambda_2 \left\| \mathbf{R}\,\mathbf{f}_k \right\|^2\right)\right\} \tag{9.28}$$

is well motivated, where **R** penalizes high frequency content within each block, **B** penalizes significant differences across the horizontal and vertical block boundaries and λ_1 and λ_2 control the influence of the different smoothing parameters. The definitions of **R** and **B** are changed slightly from

the post-processing method in (9.22), as the dimension of a block is larger in the high-resolution estimate and block boundaries may be several pixels wide. The distribution for $p(\mathbf{D}_{TB,TF})$could be defined with the methods explored in [34].

5.2 Realization

By substituting the models presented in (9.25)-(9.28) into the estimate in (9.23), a solution that simultaneously estimates the high-resolution motion field as well as the high-resolution image evolves. Taking logarithms, the super-resolution image and motion field are expressed as

$$\hat{\mathbf{f}}_k, \hat{\mathbf{D}}_{TB,TF} = \arg \min_{\mathbf{f}_k, \mathbf{D}_{TB,TF}} \left\{ \frac{1}{2} \sum_{l=k-TB}^{k+TF} \left(\mathbf{AHC}_{l,k} \mathbf{f}_k - \hat{\mathbf{g}}_l \right)^T \mathbf{K}_l^{-1} \left(\mathbf{AHC}_{l,k} \mathbf{f}_k - \hat{\mathbf{g}}_l \right) + \lambda_1 \left\| \mathbf{B} \mathbf{f}_k \right\|^2 + \lambda_2 \left\| \mathbf{R} \mathbf{f}_k \right\|^2 + \gamma \sum_{l=k-TB}^{k+TF} \sum_{i=0}^{MN-1} \left\| \mathbf{c}_{l,k}(i) - A_{MV}^T \left(\mathbf{C}_{l,k}^{Encoder}, i \right) \right\|^2 \right\}$$

$$s.t. \ \ \hat{\mathbf{f}}_k \in \left\{ \mathbf{f}_k : -\frac{\mathbf{q}_l}{2} < \mathbf{T}_{DCT} \left(\mathbf{AH}\hat{\mathbf{C}}_{l,k} \mathbf{f}_k - \hat{\mathbf{g}}_l \right) < \frac{\mathbf{q}_l}{2}, \ \ l = k - TB, \ldots, k + TF \right\}. \tag{9.29}$$

The minimization of (9.29) is accomplished with a cyclic coordinate-decent optimization procedure [35]. In the approach, an estimate for the motion field is found while the high-resolution image is assumed known. Then, the high-resolution image is predicted using the recently found motion field. The motion field is then re-estimated using the current solution for the high-resolution frame, and the process iterates by alternatively finding the motion field and high-resolution images. Treating the high-resolution image as a known parameter, the estimate for the motion field becomes

$$\hat{\mathbf{D}}_{TB,TF} = \arg \min_{\mathbf{D}_{TB,TF}} \left\{ \frac{1}{2} \sum_{l=k-TB}^{k+TF} \left(\mathbf{AHC}_{l,k} \bar{\mathbf{f}}_k - \hat{\mathbf{g}}_l \right)^T \mathbf{K}_l^{-1} \left(\mathbf{AHC}_{l,k} \bar{\mathbf{f}}_k - \hat{\mathbf{g}}_l \right) + \gamma \sum_{l=k-TB}^{k+TF} \sum_{i=0}^{MN-1} \left\| \mathbf{c}_{l,k}(i) - A_{MV}^T \left(\mathbf{C}_{TB,TF}^{Encoder}, i \right) \right\|^2 \right\}, \tag{9.30}$$

where $\bar{\mathbf{f}}_k$ is the current estimate for the high-resolution image at time *k*. Finding a solution for $\hat{\mathbf{D}}_{TB,TF}$ is accomplished with a motion estimation algorithm, and any algorithm is allowable within the framework. An example is the well-known block matching technique.

Once the estimate for the motion field is found, then the high-resolution image is computed. For the current estimate of the motion field, $\mathbf{D}_{TB,TF}$ the minimization of (9.29) is accomplished by the method of successive approximations and is expressed with the iteration

$$\bar{\mathbf{f}}_k^{n+1} = \mathbf{P}_{k-TB} \cdots \mathbf{P}_{k+TF} \left[\bar{\mathbf{f}}_k^n + \alpha \left\{ \sum_{l=k-TB}^{k+TF} \bar{\mathbf{C}}_{l,k}^T \mathbf{H}^T \mathbf{A}^T \mathbf{K}_l^{-1} \left(\mathbf{A}\mathbf{H}\bar{\mathbf{C}}_{l,k} \bar{\mathbf{f}}_k^n - \hat{\mathbf{g}}_l \right) + \lambda_1 \mathbf{B}^T \mathbf{B} \bar{\mathbf{f}}_k^n + \lambda_2 \mathbf{R}^T \mathbf{R} \bar{\mathbf{f}}_k^n \right\} \right], \tag{9.31}$$

where $\bar{\mathbf{f}}_l^n$ and $\bar{\mathbf{f}}_l^{n+1}$ are the enhanced frames at the n^{th} and $(n+1)^{th}$ iteration, respectively, α is a relaxation parameter that determines the convergence and rate of convergence of the algorithm, $\bar{\mathbf{C}}_{l,k}^T$ compensates an image backwards along the motion vectors, $\mathbf{A}^T$ defines the up-sampling operation and $\mathbf{P}_i$ is the projection operator for the quantization noise in frame *i*, as defined in (9.13).

5.3 Experimental Results

To explore the performance of the proposed super-resolution algorithm, several scenarios must be considered. In this sub-section, experimental results that illustrate the characteristics of the algorithm are presented by utilizing a combination of synthetically generated and actual image sequences. In all of the experiments, the spatial resolution of the high-resolution image sequence is 352x288 pixels, and the frame rate is 30 frames per second. The sequence is decimated by a factor of two in both the horizontal and vertical directions and compressed with an MPEG-4 compliant encoder to generate the low-resolution frames.

5.3.1 Synthetic Experiments

In the first set of experiments, a single frame is synthetically shifted by pixel increments according to

$$\mathbf{f}_k = \mathbf{C}_{o,\mathrm{mod}(k,4)} \mathbf{f}_o , \tag{9.32}$$

where $\mathbf{f}_o$ is the original frame, mod(*k*,*4*) is the modulo arithmetic operator that divides *k* by 4 and returns the remainder, and $\mathbf{C}_{o,0}$ $\mathbf{C}_{o,1}$ $\mathbf{C}_{o,2}$ and $\mathbf{C}_{o,3}$

represent the identity transform, a horizontal pixel shift, a vertical pixel shift, and a diagonal pixel shift, respectively. The original frame is shown in Figure 9.1, and the goal of the experiment is to establish an upper bound on the performance of the super-resolution algorithm. This is achieved since the experiment ensures that every pixel in the high-resolution image appears in the decimated image sequence.

The resulting image sequence is sub-sampled and compressed with an MPEG-4 compliant encoder utilizing the VM5+ rate control mechanism. No filtering is utilized, that is **H=I**. In the first experiment, a bit-rate of 1 Mbps is employed, which simulates applications with low compression ratios. In the second experiment, a bit-rate of 256 kbps is utilized to simulate high compression tasks. Both experiments maintain a frame rate of 30 frames per second.

An encoded frame from the low and high compression experiments is shown in Figure 9.2(a) and (b), respectively. Both images correspond to frame 19 of the compressed image sequence and are representative of the quality of the sequence. The original low-resolution frame 19 supplied to the encoder also appears in Figure 9.2(c). Inspecting the compressed images shows that at both compression ratios there are noticeable coding errors. Degradations in the 1 Mbps experiment are evident in the numbers at the lower right-hand corner of the image. These errors are amplified in the 256 kbps experiment, as ringing artifacts appear in the vicinity of the strong edge

Figure 9.1. Original High Resolution Frame

features throughout the image.

Visual inspection of the decoded data is consistent with the objective peak signal-to-noise ratio (PSNR) metric, which is defined as

$$PSNR = \frac{255^2}{\frac{1}{MN}\left\|\mathbf{f} - \hat{\mathbf{f}}\right\|^2}, \tag{9.33}$$

where $\mathbf{f}$ is the original image and $\hat{\mathbf{f}}$ is the high-resolution estimate. Utilizing this error criterion, the PSNR values for the low and high compression images in Figures 9.2(a) and (b) are 35.4dB and 29.3dB, respectively.

With the guarantee that every pixel in the high-resolution image appears in one of the four frames of the compressed image sequence, the super-resolution estimate of the original image and high-resolution motion field is computed with (9.29), where TB=1 and TF=2. In the experiments, the shifts in (9.32) are not assumed to be known, but a motion estimation algorithm is implemented instead. However, the motion vectors transmitted in the compressed bit-stream provide the fidelity data. The influence of these vectors is controlled by the parameter γ, which is chosen as γ=1. Other

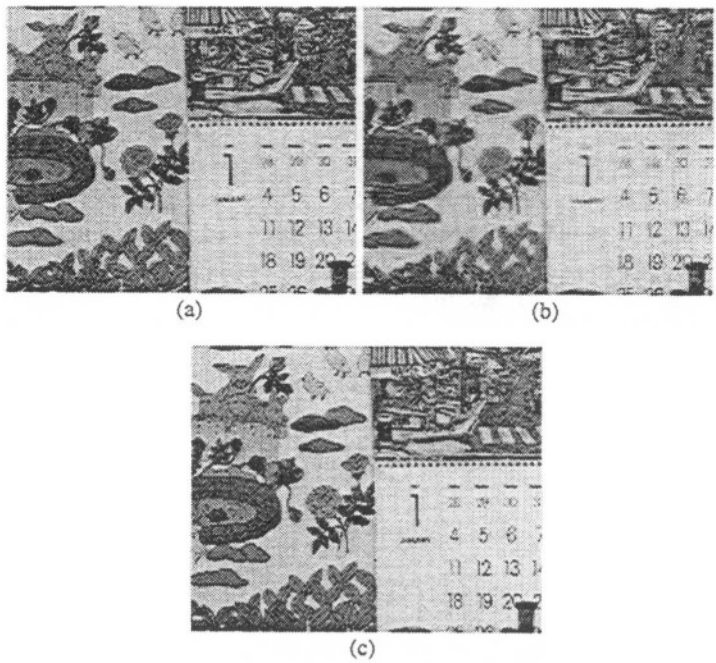


Figure 9.2. Low-Resolution Frame: (a) Compressed at 1 Mbps; (b) Compressed at 256 kbps, and (c) Uncompressed. The PSNR values for (a) and (b) are 35.4dB and 29.3dB, respectively.

parameters include the amount of smoothness in the solution, expressed as λ_1 and λ_2 in (9.31) and chosen to vary relative to the amount of quantization in the scene. For the low compression ratio experiment, $\lambda_1=\lambda_2=0.1$, while for the high compression experiments $\lambda_1=\lambda_2=0.6$. Finally, the relaxation parameter is defined as $\alpha=.125$; the iterative algorithm is terminated when $\|\tilde{\mathbf{f}}_l^n-\tilde{\mathbf{f}}_l^{n-1}\|^2<50$, and a new estimate for $\mathbf{D}_{TB,TF}$ is computed whenever $\|\tilde{\mathbf{f}}_l^n-\tilde{\mathbf{f}}_l^{n-1}\|^2<100$.

The high-resolution estimate for the 1 Mbps experiment appears in Figure 9.3(a), while the result from the 256 kbps experiment appears in Figure 9.4(a). For comparison, the decoded results are also up-sampled by bi-linear interpolation, and the interpolated images for the low and high compression ratios appear in Figure 9.3(b) and 9.4(b), respectively. As can be seen from the figure, ringing artifacts in both of the super-resolved images are attenuated, when compared to the bi-linear estimates. Also, the resolution of the image frames is increased. This is observable in many part of the image frame, and it is most evident in the numbers at the lower right portion of the image. The improvement in signal quality also appears in the PSNR metric. Comparing the super-resolved images to the original high-resolution data, the PSNR values for the low and high compression ratio experiments are 34.0dB and 29.7dB, respectively. These PSNR values are higher than the corresponding bi-linear estimates, which produce a PSNR of 31.0dB and 28.9dB, respectively.

Computing the difference between the bi-linear and super-resolution estimates provides additional insight into the problem of super-resolution from compressed video. In the 1 Mbps experiment, the PSNR of the super-resolved image is 3.0dB higher than the PSNR of the bi-linear estimate. This is a greater improvement than realized in the 256 kbps experiment, where the high-resolution estimate is only .8dB higher than the PSNR of the bi-linear estimate. The improvement realized by the super-resolution algorithm is inversely proportional to the severity of the compression. Higher compression ratios complicate the super-resolution problem in a major way, as aliased high frequency information in the low-resolution image sequence is removed by the compression process. Since relating the low and high-resolution data through a motion field is the foundation of a super-resolution algorithm, the removal of this information limits the amount of expected improvement. Moreover, the missing data often introduces errors when estimating the motion field, which further limits the procedure.

Overcoming the problem of high-resolution data that is observable at other time instances but removed during encoding is somewhat mitigated by incorporating additional frames into the high-resolution estimate. This improves the super-resolved image, as an encoder may preserve the data in one frame but not the other. In addition, the approach benefits video

(a)

(b)

Figure 9.3. Results of the Synthetic Experiment at 1 Mbps: (a) Super-Resolved Image and (b) Bi-Linear Estimate. The PSNR values for (a) and (b) are 34.0dB and 31.0dB, respectively.

(a)

(b)

Figure 9.4. Results of the Synthetic Experiment at 256 kbps: (a) Super-Resolved Image and (b) Bi-Linear Estimate. The PSNR values for (a) and (b) are 29.7dB and 28.9dB, respectively.

sequences that do not undergo a series of sub-pixel shifts. In either case, increasing the number of frames makes it more likely that information about the high-solution image appear at the decoder. The amount of improvement is however restricted by the fact that objects may only appear in a limited number of frames and motion estimates from temporally distant frames may be unreliable.

5.3.2 Non-Synthetic Experiments

Increasing the number of frames that are utilized in the super-resolution estimate is considered in the second set of experiments. In this scenario, the high-resolution image sequence is considered that contains the frame appearing in the synthetic example. The scene consists of a sequence of images that are generated by a slow panning motion. In addition, the calendar object is also moving in the horizontal direction.

Like the previous experiments, the high-resolution frames are down-sampled by a factor two and compressed with an MPEG-4 compliant encoder utilizing the VM5+ rate control. No filtering is utilized, and the sub-sampled image sequence is encoded at both 1 Mbps and 256 kbps to simulate both high and low compression environments. Encoded images from both experiments are shown in Figure 9.5(a) and (b), respectively, and correspond to frame 19 of the sequence. As in the synthetic example, some degradations appear in the low compression ratio result, which become more noticeable as the compression ratio is increased. These errors appear throughout the frame but are most noticeable around the high-frequency components of the numbers in the lower right-hand corner.

The super-resolution estimates for the 1 Mbps and 256 kbps experiments appear in Figure 9.6(a) and 9.7(a), respectively, while the decoded results after up-sampling with bi-linear interpolation appear in Figure 9.6(b) and 9.7(b), respectively. By inspecting the figure, conclusions similar to the synthetic experiments are made. Ringing artifacts in both of the super-resolution estimates are reduced, as compared to the bi-linear estimates. In addition, the resolution of the image frames is increased within the numbers appearing at the lower right of the frame. (Specifically, notice the improvement on the 6.) These improvements result in an increase in PSNR. For the super-resolved images, the low and high compression ratio experiments produce a PSNR of 31.6dB and 29.1dB, respectively. The bi-linear estimate provides lower PSNR values of 30.9dB and 28.7dB, respectively.

Comparing the improvement in PSNR between the synthetic and actual image sequences provides a quantitative measure of the difficulties introduced by processing real image sequences. For the 1Mbps experiments,

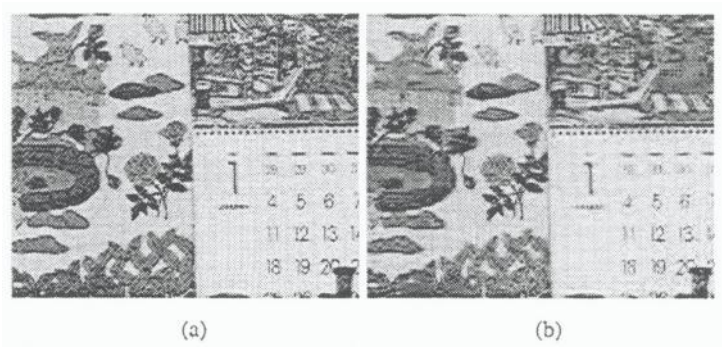


Figure 9.5. Low-Resolution Frame: (a) Compressed at 1 Mbps, and (b) Compressed at 256 kbps. The PSNR values for (a) and (b) are 35.5dB and 29.2dB, respectively.

the PSNR of the high-resolution estimate is .7dB larger than the bi-linear result. This is a smaller improvement than realized with the synthetic example, where the gain is 3.0dB. Differences between the experiments are even more noticeable at the lower bit-rate, where the PSNR of the high-resolution estimate is only .4dB greater than the bi-linear estimate. This is also a decrease in performance, as compared to the .8dB gain of the synthetic simulations.

As discussed previously, several problems with actual image sequences contribute to a decrease in performance. These problems include the removal of information by a compression system and the absence of sub-pixel shift in the image sequence. To address these problems, it is advantageous to include additional frames in the super-resolution estimate, as these frames contain additional observations of the high-resolution estimate. The impact of the additional frames is explored in the final experiment, where the super-resolution estimate for the 256 kbps actual image sequence is recomputed. Parameters for the experiment are equal to the previously defined values, except that nine frames are included in the high-resolution estimate, corresponding to TB=3 and TF=5.

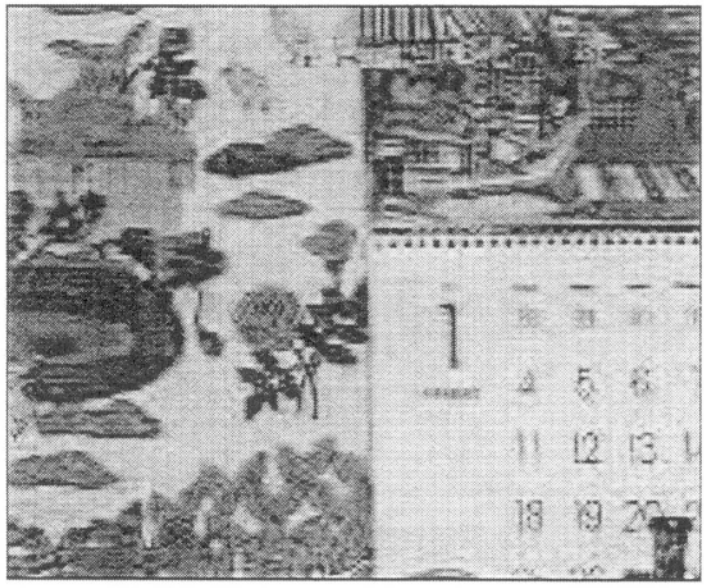
The super-resolution image for the nine frame experiment appears in Figure 9.8, and it illustrates an improvement when compared to the four frame estimate shown in Figure 9.8. As in the previous experiments, differences between the images are most noticeable in the regions surrounding the numbers, where the addition of the five frames into the super-resolution algorithm further attenuates the ringing and improves the definition of the numbers. These improvements also increase the PSNR of the high-resolution estimate, which increase from 29.1dB to 29.5dB after incorporating the extra five frames.

(a)

(b)

Figure 9.6. Results of the Non-Synthetic Experiment at 1Mbps: (a) Super-Resolved Image and (b) Bi-Linear Estimate. The PSNR values for (a) and (b) are 31.6dB and 30.8dB, respectively.

(a)



(b)

Figure 9.7. Results of the Non-Synthetic Experiment at 256 kbps: (a) Super-Resolved Image and (b) Bi-Linear Estimate. The PSNR values for (a) and (b) are 29.1dB and 28.7dB, respectively.

Figure 9.8. Result of the Non-Synthetic Experiment with Nine Frames. The compression rate is 256 kbps, and the PSNR is 29.5dB.

6. CONCLUSIONS

In this chapter, the problem of recovering a high-resolution frame from a sequence of low-resolution and compressed images is considered. Special attention is focused on the compression system and its effect on the recovery technique. In a traditional resolution recovery problem, the low-resolution images contain aliased information from the original high-resolution frames. Sub-pixel shifts within the low-resolution sequence facilitate the recovery of spatial resolution from the aliased observations. Unfortunately when the low-resolution images are compressed, the amount of aliasing is decreased. This complicates the super-resolution problem and suggests that a model of the compression system be included in the recovery technique. Several methods are explored in the chapter for incorporating the compression system into the recovery framework. These techniques exploit the parameters in the compressed bit-stream and lead to a general solution approach to the problem of super-resolution from compressed video.

REFERENCES

1. A.K. Katsaggelos and N.P. Galatsanos, eds. *Signal Recovery Techniques for Image and Video Compression and Transmission*, Kluwar Academic Publishers, 1998.
2. J.D. Gibson, *et al.*, *Digital Compression for Multimedia*. Morgan Kaufmann, 1998.
3. A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwar Academic Publishers, 1992.
4. ISO/IEC JTC1/SC29 International Standard 10918-1, *Digital Compression and Coding of Continuous-Tone Still Images*, 1991.
5. W.A. Pearlman, B.-J. Kim, and Z. Xiong, *Embedded video coding with 3D SPIHT*, in *Wavelet Image and Video Compression*, P.N. Topiwala, Editor, Kluw. 1998, Kluwer: Boston, MA.
6. C. Podilchuk, N. Jayant, and N. Farvardin, *Three-dimensional subband coding of video*. IEEE Transactions on Image Processing, 1995. **4**(2): p. 125-139.
7. B.G. Haskell and J.O. Limb, *Predictive video encoding using measured subjective velocity*, U.S. Patent 3,632,856, January 1972.
8. V. Bhaskaran and K. Konstantinides, *Image and Video Compression Standards*. 2 ed. Kluwer Academic Publishers, 1997.
9. ITU-T Recommendation H.261, *Video Codec for Audio Visual Services at px64 kbits/s*, 1993.
10. ITU-T Recommendation H.263, *Video Coding for Low Bitrate Communications*, 1998.
11. ISO/IEC JTC1/SC29 International Standard 11172-2, *Information Technology -- Generic Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5Mbps -- Part 2: Video*, 1993.
12. ISO/IEC JTC1/SC29 International Standard 13818-2, *Information Technology -- Generic Coding of Moving Pictures and Associated Audio Information: Video*, 1995.
13. ISO/IEC JTC1/SC29 International Standard 14496-2, *Information Technology -- Generic Coding of Audio-Visual Objects: Visual*, 1999.
14. ISO/IEC JTC1/SC29 International Standard 14496-2AM1, *Information Technology -- Generic Coding of Audio-Visual Objects: Visual*, 2000.
15. A.J. Patti and Y. Altunbasak. *Super-Resolution Image Estimation for Transform Coded Video with Application to MPEG*. in *IEEE International Conference on Image Processing*. 1999. Kobe, Japan.
16. Y. Altunbasak and A.J. Patti. *A Maximum a Posteriori Estimator for High Resolution Video Reconstruction from MPEG Video*. in *IEEE International Conference on Image Processing*. 2000. Vancouver, BC.

REFERENCES

17. A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering.* 2 ed. 1994, Reading, MA: Addison-Wesley Publishing Company, Inc.

18. D. Chen and R.R. Schultz. *Extraction of High-Resolution Video Stills from MPEG Image Sequences*. in *IEEE International Conference on Image Processing.* 1998. Chicago, IL.

19. K.J. Erickson and R.R. Schultz. *MPEG-I Super-Resolution Decoding for the Analysis of Video Stills*. in *Fourth IEEE Southwest Symposium on Image Analysis.* 2000. Austin, TX.

20. J. Mateos, A.K. Katsaggelos, and R. Molina. *Simultaneous Motion Estimation and Resolution Enhancement of Compressed Low Resolution Video*. in *IEEE International Conference on Image Processing.* 2000. Vancouver, BC.

21. J. Mateos, A.K. Katsaggelos, and R. Molina. *Resolution Enhancement of Compressed Low Resolution Video*. in *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2000. Istanbul, Turkey.

22. A. Kaup. *Adaptive Constrained Least Squares Restoration for Removal of Blocking Artifacts in Low Bit Rate Video Coding*. in *IEEE International Conference on Acoustics, Speech and Signal Processing.* 1997. San Jose, CA.

23. R. Rosenholtz and A. Zakhor, *Iterative Procedures for Reduction of Blocking Effects in Transform linage Coding.* IEEE Transactions on Circuits and Systems for Video Technology, 1992. **2**(1): p. 91-94.

24. Y. Yang, N.P. Galatsanos, and A.K. Katsaggelos, *Regularized Reconstruction to Reduce Blocking Artifacts of Block Discrete Cosine Transform Compressed Images.* IEEE Transactions on Circuits and Systems for Video Technology, 1993. **3**(6): p. 421-432.

25. A.K. Katsaggelos, *Iterative Image Restoration Algorithms.* Optical Engineering, 1989. **28**(7): p. 735-748.

26. M.G. Kang and A.K. Katsaggelos, *Simultaneous Multichannel Image Restoration and Estimation of the Regularization Parameters.* IEEE Transactions on Image Processing, 1992. **6**(5): p. 774-778.

27. C.-J. Tsai, *et al. A Compressed Video Enhancement Algorithms*, in *IEEE International Conference on Image Processing.* 1999. Kobe, Japan.

28. C.A. Segall and A.K. Katsaggelos. *Enhancement of Compressed Video using Visual Quality Metrics.* in *IEEE International Conference on Image Processing.* 2000. Vancouver, BC.

29. J. Mateos, A.K. Katsaggelos, and R. Molina, *A Bayesian Approach for the Estimation and Transmission of Regularization Parameters for Reducing Blocking Artifacts.* IEEE Transactions on Image Processing, 2000. **9**(7): p. 1200-1215.

30. Y. Yang, N.P. Galatsanos, and A.K. Katsaggelos, *Projection-Based Spatially Adaptive Reconstruction of Block-Transform Compressed Images.* IEEE Transactions on Image Processing, 1995. **4**(7): p. 896-908.

31.Y. Yang and N.P. Galatsanos, *Removal of Compression Artifacts Using Projections onto Convex Sets and Line Process Modeling.* IEEE Transactions on Image Processing, 1998. **6**(10): p. 1345-1357.

32.T. Ozcelik, J.C. Brailean, and A.K. Katsaggelos, *Image and Video Compression Algorithms Based on Recovery Techniques using Mean Field Annealing.* Proceedings of the IEEE, 1995. **83**(2): p. 304-316.

33.T.P. O'Rourke and R.L. Stevenson, *Improved Image Decompression for Reduced Transform Coding Artifacts.* IEEE Transactions on Circuits and Systems for Video Technology, 1995. **5**(6): p. 490-499.

34.J.C. Brailean and A.K. Katsaggelos, *Simultaneous Recursive Motion Estimation and Restoration of Noisy and Blurred Image Sequences.* IEEE Transactions on Image Processing, 1995. **4**(9): p. 1236-1251.

35.D.G. Luenberger, *Linear and Nonlinear Programming.* 1984, Reading, MA: Addison-Wesley Publishing Company, Inc.

Chapter 10

SUPER-RESOLUTION: LIMITS AND BEYOND

Simon Baker and Takeo Kanade
Robotics Institute, School of Computer Science,
Carnegie Mellon University,
Pittsburgh, PA 15213, USA.
{simonb, tk}@cs.cmu.edu

Abstract

A variety of super-resolution algorithms have been described in this book. Most of them are based on the same source of information however; that the super-resolution image should generate the lower resolution input images when appropriately warped and down-sampled to model image formation. (This information is usually incorporated into super-resolution algorithms in the form of reconstruction constraints which are frequently combined with a smoothness prior to regularize their solution.) In this final chapter, we first investigate how much extra information is actually added by having more than one image for super-resolution. In particular, we derive a sequence of analytical results which show that the reconstruction constraints provide far less useful information as the decimation ratio increases. We validate these results empirically and show that for large enough decimation ratios any smoothness prior leads to overly smooth results with very little high-frequency content however many (noiseless) low resolution input images are used. In the second half of this chapter, we propose a super-resolution algorithm which uses a completely different source of information, in addition to the reconstruction constraints. The algorithm recognizes local "features" in the low resolution images and then enhances their resolution in an appropriate manner, based on a collection of high and low-resolution training samples. We call such an algorithm a *hallucination* algorithm.

Keywords: Super-resolution, analysis of limits, learning, faces, text, hallucination.

1. Introduction

A large number of super-resolution algorithms have been described in this book. Most of them, however, are based on the same source of information; specifically, that the super-resolution image, when appropriately warped and down-sampled to model the image formation process, should yield the low resolution images. This information is typically embedded in a set of reconstruction constraints, first introduced by (Peleg et al., 1987; Irani and Peleg, 1991). These reconstruction constraints can be embedded in a Bayesian framework incorporating a prior on the super-resolution image (Schultz and Stevenson, 1996; Hardie et al., 1997; Elad and Feuer, 1997). Their solution can also be estimated either in batch mode or recursively using a Kalman filter (Elad and Feuer, 1999; Dellaert et al., 1998). Several other refinements have been proposed, including simultaneously computing 3D structure (Cheeseman et al., 1994; Shekarforoush et al., 1996; Smelyanskiy et al., 2000) and removing other degrading artifacts such as motion blur (Bascle et al., 1996).

In the first part of this chapter, we analyze the super-resolution reconstruction constraints. We derive three analytical results which show that the amount of information provided by having more than one image available for super-resolution becomes very much less as the decimation ratio q increases. Super-resolution therefore becomes inherently much more difficult as q increases. This reduction in the amount of information provided by the reconstruction constraints is traced to the fact that the pixel intensities in the input images take discrete values (typically 8-bit integers in the range 0–255). This causes a loss of information and imposes inherent limits on how well super-resolution can be performed from the reconstruction constraints (and other equivalent formulations based on the same underlying source of information.)

How, then, can high-decimation ratio super-resolution be performed? Our analytical results hold for an arbitrary number of images so using more low resolution images does not help. Suppose, however, that the input images contain printed text. Moreover, suppose that it is possible to perform optical character recognition (OCR) and recognize the text. If the font can also be determined, it would then be easy to perform super-resolution for *any decimation ratio.* The text could be reproduced at any resolution by simply rendering it from the script of the text and the definition of the font. In the second half of this chapter, we describe a super-resolution algorithm based on this idea which we call *hallucination* (Baker and Kanade, 1999; Baker and Kanade, 2000a). Our super-resolution hallucination algorithm is based, however, on the recognition of generic local "features" (rather than the characters de-

tected by OCR). It can therefore be applied to other phenomena such as images of human faces.

2. The Reconstruction Constraints

Denote the low resolution input images by $x_L^{(k)}(i,j)$ where $k = 1, ..., K$. The starting point in the derivation of the reconstruction constraints is then the continuous image formation equation (Horn, 1996):

$$x_L^{(k)}(i,j) = \left(I^{(k)} * h^{(k)}\right)(i,j) = \int_{x_L^{(k)}} I^{(k)}(x,y) \cdot h^{(k)}(x-i, y-j)\,\mathrm{dx}\,\mathrm{dy} \tag{10.1}$$

where $I^{(k)}(x,y)$ is the continuous irradiance function that would have reached the image plane of the k^{th} camera under the pinhole model, and $h^{(k)}$ is point spread function of the k^{th} camera. The (double) integration is performed over the image plane of $x_L^{(k)}$. See Figure 10.1 for an illustration.

2.1. Modeling the Point Spread Function

We decompose the point spread function into two parts (see Figure 10.1):

$$h^{(k)}(x,y) = \left(\omega^{(k)} * a^{(k)}\right)(x,y) \tag{10.2}$$

where $\omega^{(k)}(x,y)$ models the blurring caused by the optics and $a^{(k)}(x,y)$ models the spatial integration performed by the CCD sensor (Baker et al., 1998). The optical blurring $\omega^{(k)}$ is typically further split into a defocus factor that can be approximated by a pill-box function and a diffraction-limited optical transfer function that can be modeled by the square of the first-order Bessel function of the first kind (Born and Wolf, 1965). We aim to be as general as possible and so avoid making any assumptions about $\omega^{(k)}$. Instead, (most of) our analysis is performed for arbitrary optical blurring functions. We do, however, assume a parametric form for $a^{(k)}$. We assume that the the photo-sensitive areas of the CCD pixels are square and uniformly sensitive to light, as in (Baker et al., 1998; Barbe, 1980). If the length of the side of the square photosensitive area is $S^{(k)}$, the spatial integration function is then:

$$a^{(k)}(x,y) = \begin{cases} \frac{1}{S^{(k)} \times S^{(k)}} & \text{if } |x| \leq \frac{S^{(k)}}{2} \text{ and } |y| \leq \frac{S^{(k)}}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{10.3}$$

In general the photosensitive area is not the entire pixel since space is needed for the circuitry to read out the charge. Therefore the only assumption we make about $S^{(k)}$ is that it lies in [0,1]. Our analysis

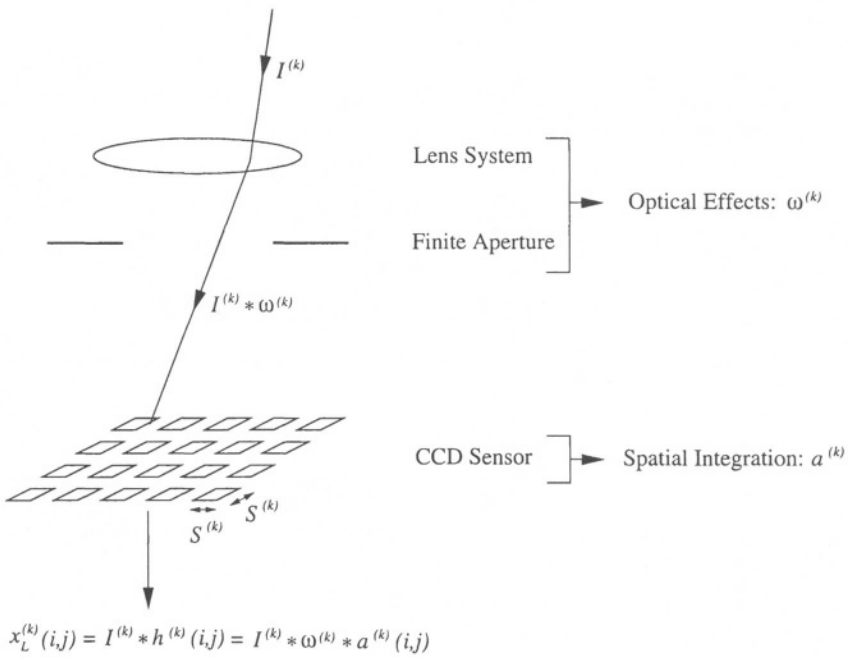


Figure 10.1. The low resolution input images $x_L^{(k)}$ are formed by the convolution of the irradiance $I^{(k)}$ with the camera point spread function $h^{(k)}$. We model the point spread function itself as the convolution of two terms: (1) $\omega^{(k)}$ models the optical effects caused by the lens and the finite aperture, and (2) $a^{(k)}$ models the spatial integration performed by the CCD sensor.

is then in terms of $S^{(k)}$ (rather than the inter-pixel distance which is assumed to define the unit distance.)

2.2. What is Super-Resolution Anyway?

We wish to estimate a super-resolution image $x_H(i', j')$. Precisely what does this mean? Let us begin with the coordinate frame of x_H. The coordinate frame of a super-resolution image is typically defined relative to that of the corresponding low resolution input image. If the decimation ratio is q, the pixels in x_H will be q times closer to each other than those in the corresponding low resolution image, $x_L^{(k')}$ say. The coordinate frame of x_H can therefore be defined in terms of that for $x_L^{(k')}$ via:

$$(i', j') = \left(\frac{i}{q}, \frac{j}{q}\right). \tag{10.4}$$

In this chapter we assume that the input images have already been registered with each other and therefore with the coordinate frame of x_H. Then, denote the point in image $x_L^{(k)}$ (where k may or may not equal k') that corresponds to (x, y) in x_H by $\mathbf{r}^{(\mathbf{k})}(\mathbf{x}, \mathbf{y})$. From now on we assume that $\mathbf{r}^{(\mathbf{k})}$ isknown.

The integration in Equation (10.1) is performed over the low resolution image plane. Transforming to the super-resolution image plane of x_H gives:

$$x_L^{(k)}(i,j) = \int_{x_H} I^{(k)}(\mathbf{r}^{(\mathbf{k})}(\mathbf{x},\mathbf{y})) \cdot \mathbf{h}^{(\mathbf{k})}(\mathbf{r}^{(\mathbf{k})}(\mathbf{x},\mathbf{y}) - (\mathbf{i},\mathbf{j})) \cdot \left|\frac{\partial \mathbf{r}^{(\mathbf{k})}}{\partial \mathbf{x},\mathbf{y}}\right| \mathrm{dx\,dy} \tag{10.5}$$

where $\left|\frac{\partial \mathbf{r}^{(\mathbf{k})}}{\partial x,y}\right|$ is the determinant of the Jacobian of the registration $\mathbf{r}^{(\mathbf{k})}$.

Now, $I^{(k)}(\mathbf{r}^{(\mathbf{k})}(\mathbf{x},\mathbf{y}))$ is the irradiance that would have reached the image plane of the k^{th} camera under the pinhole model, transformed onto the super-resolution image plane. Assuming that the registration is correct, and that the radiance of every point in the scene does change across k (a Lambertian-like assumption), $I^{(k)}(\mathbf{r}^{(\mathbf{k})}(\mathbf{x},\mathbf{y}))$ should be the same for all k. Moreover, it equals the irradiance that would have reached the super-resolution image plane of x_H under the pinhole model. Denoting this function by $I(x, y)$, we have:

$$x_L^{(k)}(i,j) = \int_{x_H} I(x,y) \cdot h^{(k)}(\mathbf{r}^{(\mathbf{k})}(\mathbf{x},\mathbf{y}) - (\mathbf{i},\mathbf{j})) \cdot \left|\frac{\partial \mathbf{r}^{(\mathbf{k})}}{\partial \mathbf{x},\mathbf{y}}\right| \mathrm{dx\,dy}. \tag{10.6}$$

The goal of super-resolution is then to recover (a representation of) $I(x, y)$. Doing this requires both increasing the resolution and "deblur-

ring" the image; i.e. removing the effects of the convolution with the point spread function $h^{(k)}$.

In order to proceed we need to specify which continuous function $I(x, y)$ is represented by the discrete image $x_H(i', j')$. For simplicity, we assume that $x_H(i', j')$ represents the piecewise constant function:

$$I(x, y) = x_H(i', j') \tag{10.7}$$

for all $x \in (i' - 0.5, j' + 0.5]$ and $y \in (j' - 0.5, j' + 0.5]$. Then, Equation (10.6) can be rearranged to give the super-resolution reconstruction constraints:

$$x_L^{(k)}(i, j) = \sum_{i', j'} W^{(k)}(i, j, i', j') \cdot x_H(i', j') \tag{10.8}$$

where $k = 1, \ldots, K$ and:

$$W^{(k)}(i, j, i', j') = \int_{i'-0.5, j'-0.5}^{i'+0.5, j'+0.5} h^{(k)}(\mathbf{r}^{(k)}(\mathbf{x}, \mathbf{y}) - (\mathbf{i}, \mathbf{j})) \cdot \left| \frac{\partial \mathbf{r}^{(k)}}{\partial \mathbf{x}, \mathbf{y}} \right| \mathrm{dx}\, \mathrm{dy}. \tag{10.9}$$

The super-resolution reconstruction constraints are therefore a set of linear constraints on the unknown super-resolution pixels $x_H(i', j')$ in terms of the known low resolution pixels $x_L^{(k)}(i, j)$ and the coefficients $W^{(k)}(i, j, i', j')$.

3. Analysis of the Constraints

The constant coefficients $W^{(k)}(i, j, i', j')$ in the reconstruction constraints depend on both the point spread function $h^{(k)}$ and the registration $\mathbf{r}^{(k)}$. Without some assumptions about these functions any analysis would be meaningless. If the point spread function is arbitrary, it can be chosen to simulate the "small pixels" of the super-resolution image. Similarly, if the registration is arbitrary, it can be chosen (in effect) to move the camera towards the scene and thereby directly capture the super-resolution image. We therefore have to make some (reasonable) assumptions about the imaging conditions.

Assumptions Made About the Point Spread Function

As mentioned above, we assume that the point spread function takes the form of Equation (10.3). Moreover, we assume that the width of the photosensitive area $S^{(k)}$ is the same for all of the images (and equals S.) In the first part of our analysis, we also assume that $\omega^{(k)}(x, y) = \delta(x) \cdot \delta(y)$, where δ is the Dirac delta function. Afterwards, in the second and third parts of our analysis, we allow $\omega^{(k)}$ to be arbitrary; i.e. our analysis holds for *any* optical blurring.

Assumptions Made About the Registration

To outlaw motions which (effectively) allow the camera to be moved towards the scene, we assume that each registration takes the form:

$$\mathbf{r}^{(\mathbf{k})}(\mathbf{x},\mathbf{y}) \;=\; \frac{\mathbf{1}}{\mathbf{q}}(\mathbf{x},\mathbf{y}) + (\mathbf{c}^{(\mathbf{k})},\mathbf{d}^{(\mathbf{k})}) \qquad (10.10)$$

where $(c^{(k)}, d^{(k)})$ is a constant translation (which in general may be different for each low resolution image k) and the $\frac{1}{q}$ accounts for the change of coordinate frame from high to low resolution images. See also Equation (10.4).

Even given these assumptions, the performance of any super-resolution algorithm will depend upon the exact number of input images K, the values of $(c^{(k)}, d^{(k)})$, and, moreover, how well the algorithm can register the low resolution images to estimate the $(c^{(k)}, d^{(k)})$. Our goal is to show that super-resolution becomes fundamentally more difficult as the decimation ratio q increases. We therefore assume that the conditions are as favorable as possible and perform the analysis for an arbitrary number of input images K, with arbitrary translations $(c^{(k)}, d^{(k)})$. We also assume that the algorithm has estimated these values perfectly. Any results derived under these conditions will only be stronger in practice, where the registrations may be degenerate or inaccurate.

3.1. Invertibility Analysis

We first analyze when the reconstruction constraints are invertible, and what the rank of the null space is when they are not. In order to get an easily interpretable result, the analysis in this section is performed under the scenario that the optical blurring can be ignored; i.e. $\omega^{(k)}(x,y) = \delta(x)\cdot\delta(y)$. (This assumption will be removed in the following two sections.) The expression for $W^{(k)}(i,j,i',j')$ in Equation (10.9) then simplifies to:

$$\frac{1}{q^2}\int_{i'-0.5}^{i'+0.5}\int_{j'-0.5}^{j'+0.5} a^{(k)}\left(\frac{1}{q}(x,y) + (c^{(k)},d^{(k)}) - (i,j)\right)\,\mathrm{dx}\,\mathrm{dy}. \qquad (10.11)$$

Using the definition of $a^{(k)}$ it can be seen that $W^{(k)}(i,j,i',j')$ is equal to $1/(q\cdot S)^2$ times the area of the intersection of the two squares in Figure 10.2 (the high resolution pixel $[i'-0.5, i'+0.5]\times[j'-0.5, j'+0.5]$ and the region where $a^{(k)}$ non-zero and equals $\frac{1}{S^2}$.) We then have:

Theorem 1 *If $q\cdot S$ is an integer greater than 1, then for all $(c^{(k)}, d^{(k)})$ the reconstruction constraints (Equations (10.8) and (10.11)) are not invertible. Moreover, the dimension of the null space is at least* $(q\cdot S -$

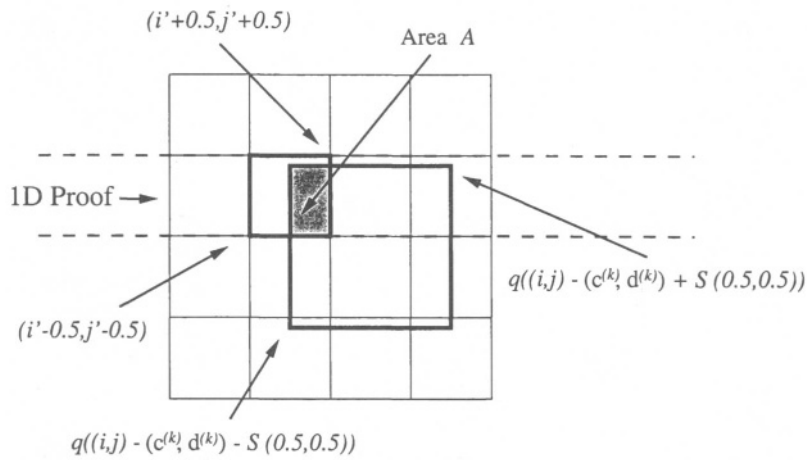


Figure 10.2. The high-resolution pixel (i', j') over which the integration is performed in Equation (10.11) is indicated by the small square at the upper middle left of the figure. The larger square towards the bottom right is the region in which $a^{(k)}$ is non-zero. Since $a^{(k)}$ takes the value $1/S^2$ in this region, the integral in Equation (10.11) equals A/S^2, where A is the area of the intersection of the two squares. This figure is used to illustrate the 1D proof of Theorem 1.

1)^2. *If $q \cdot S$ is not an integer, $c^{(k)}$ and $d^{(k)}$ always exist such that the constraints are invertible.*

Proof: We provide a proof for 1D images. (See Figure 10.2.) The extension to 2D is conceptually no more difficult and so is omitted for reasons of brevity.

The null space is defined by $\sum_{i',j'} \overline{W}^{(k)}(i, j, i', j') \cdot x_H^{(k)}(i', j') = 0$ where $\overline{W}^{(k)}(i, j, i', j') = (q \cdot S)^2 \cdot W^{(k)}(i, j, i', j')$ is the area of intersection of the 2 squares in Figure 10.2. Any element of the null space therefore corresponds to an assignment of values to the small squares such that their weighted sum (over the large square) equals zero, where the weights are the areas of intersection.

In 1D we just consider one row of the figure. Changing $c^{(k)}$ (and $d^{(k)}$) to slide the large square along the row by a small amount, we get a similar constraint on the elements in the null space. The only difference is in the left-most and right-most small squares. Subtracting these two constraints shows that the left-most square and the right-most square must have the same value.

If $q \cdot S$ is not an integer (or is 1), this proves that neighboring values of $x_H^{(k)}$ must be equal and hence 0. (Since $q \cdot S$ is not an integer, the

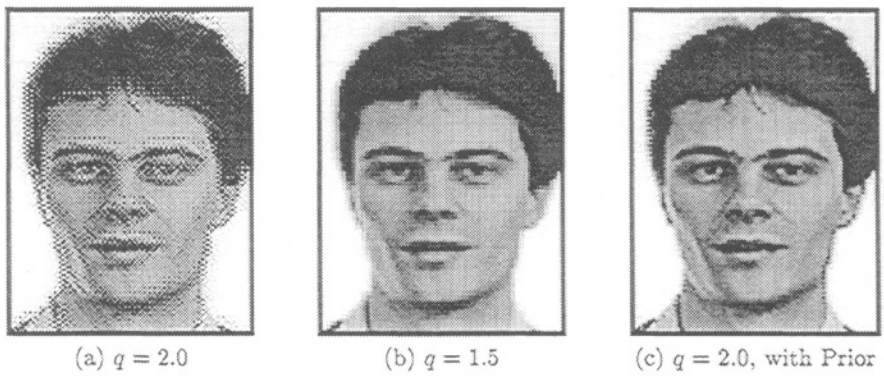


Figure 10.3. Validation of Theorem 1: The results of solving the reconstruction constraints using gradient descent for a square point spread function with $S = 1.0$. (a) When $q \cdot S$ is an integer, the equations are not invertible and so a random periodic image in the null space is added to the original image, (b) When $q \cdot S$ is not an integer, the reconstruction constraints are invertible (in general) and so a smooth solution is found, even without a prior. (The result for $q = 1.5$ was interpolated to make it the same size as that for $q = 2.0$.) (c) When a smoothness prior is added to the reconstruction constraints the difficulties seen in (a) disappear. (For larger values of q simply adding a smoothness prior does not solve this problem, as will be seen.)

big square slides out of one small square before the other and the result then follows by transitivity of equality.) Therefore, there exist values for the translations $c^{(k)}$ (and $d^{(k)}$) such that the null space only contains the zero vector; i.e. the reconstruction constraints are invertible in general if $q \cdot S$ is not an integer (or is 1).

If $q \cdot S$ is an integer greater than 1, this same constraint places an upper bound of $q \cdot S - 1$ on the maximum dimension of the null space computed over all possible translations $c^{(k)}$ (and $d^{(k)}$). The space of all assignments to $x_H^{(k)}$ that are periodic with period $q \cdot S$ and which have a zero mean can also easily be seen to always lie in the null space and so this value is also a lower bound on the dimension of the null space for any translations $c^{(k)}$ (and $d^{(k)}$). □

To validate this theorem, we solved the reconstruction constraints using gradient descent for the two cases $q = 2.0$ and $q = 1.5$, (where $S = 1.0$.) The results are presented in Figure 10.3. In this experiment, no smoothness prior is used and gradient descent is run for a sufficiently long time that the (smooth) initial image does not bias the results. The input in both cases consisted of multiple down-sampled images of the face. Specifically, 1024 randomly translated images were used as input. Exactly the same inputs are used for the two experiments. The only

difference is the decimation ratio. (The output for $q = 1.5$ is actually smaller than that for $q = 2.0$ and was interpolated to be the same size for display purposes. This is the reason it appears slightly smoother than (c).)

As can be seen in Figure 10.3, for $q = 2.0$ the (additive) error is approximately a periodic image with period 2 pixels. For $q = 1.5$ the equations are invertible and so a smooth solution is found, even though no smoothness prior was used. For $q = 2.0$, the fact that the problem is not invertible does not have any practical significance. Adequate solutions can be obtained by simply adding a smoothness prior to the reconstruction constraints, as shown in Figure 10.3(c). For $q \gg 2$ the situation is different, however. The rapid rate of increase of the dimension of null space (quadratic in $q \cdot S$) is the root cause of the problems, as will be seen in the next two sections.

3.2. Conditioning Analysis

Most linear systems that are close to being not invertible are usually ill-conditioned. It is no surprise then that changing from a square point spread function to an arbitrary blurring function $h^{(k)} = \omega^{(k)} * a^{(k)}$ results in an ill-conditioned system, as we now show in the second part of our analysis:

Theorem 2 *If $\omega^{(k)}(x, y)$ is a function for which $\omega^{(k)}(x, y) \geq 0$ for all (x, y) and $\int\int \omega^{(k)}(x, y)\, dx\, dy = 1$, then the condition number of the reconstruction constraints (Equations (10.8) and (10.9)) grows at least as fast as $(q \cdot S)^2$.*

Proof: We first prove the theorem for the square point spread function $h^{(k)} = a^{(k)}$ (i.e. for Equations (10.8) and (10.11)) and then generalize. The condition number of a linear operator A can be written as:

$$\text{Cond(A)} = \frac{\sup_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{Ax}\|_\infty}{\inf_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{Ax}\|_\infty}. \tag{10.12}$$

It follows from Equations (10.8) and (10.11) that if $x_H(i', j') = 1$ for all (i', j'), then $x_L^{(k)}(i, j) = 1$ for all (i, j). Hence the numerator in Equation (10.12) is at least 1. Setting $x_H(i', j')$ to be the checkerboard pattern (1 if $i' + j'$ is even, -1 if odd) we find that $|x_L^{(k)}(i, j)| \leq 1/(q \cdot S)^2$ since the integration of the checkerboard over any square in the real plane lies in the range [–1, 1]. (Proof omitted.) Hence the denominator is at most $1/(q \cdot S)^2$. The desired result for $h^{(k)} = a^{(k)}$ follows immediately.

For arbitrary point spread functions, note that Equations (10.8) and (10.9) can be combined and then rewritten as:

$$\begin{aligned} x_L^{(k)}(i,j) &= \int_{x_H} \frac{x_H(x,y)}{q^2} \cdot h^{(k)}\left(\frac{1}{q}(x,y) + (c^{(k)}, d^{(k)}) - (i,j)\right) \mathrm{dx}\,\mathrm{dy} \\ &= \left(h^{(k)} * \overline{x}_H\right)\left((c^{(k)}, d^{(k)}) - (i,j)\right) \\ &= \left[\omega^{(k)} * \left(a^{(k)} * \overline{x}_H\right)\right]\left((c^{(k)}, d^{(k)}) - (i,j)\right) \end{aligned} \tag{10.13}$$

where we have set $\overline{x}_H(x,y) = x_H(-qx,-qy)$ and changed variables $(x,y) \Rightarrow -\frac{1}{q}(x,y)$. Both of the properties of $x_L^{(k)}$ that we used to prove the result for square point spread functions therefore also hold with $a^{(k)}$ replaced by $h^{(k)} = \omega^{(k)} * a^{(k)}$ using standard properties of the convolution operator. Hence, the desired, more general, result follows immediately from Equation (10.13). □

This theorem is more general than the previous one because it applies to arbitrary optical blurring functions. On the other hand, it is a weaker result (in some situations) because it only predicts that super-resolution is ill-conditioned (rather than not invertible.) This theorem on its own, therefore, does not entirely explain the poor performance of super-resolution. As we showed in Figure 10.3, problems that are ill-conditioned (or even not invertible, where the condition number is infinite) can often be solved by simply adding a smoothness prior. (The not invertible super-resolution problem in Figure 10.3(a) is solved in Figure 10.3(c) in this way.) Several researchers have performed conditioning analysis of various forms of super-resolution, including (Elad and Feuer, 1997; Shekarforoush, 1999; Qi and Snyder, 2000). Although useful, none of these results fully explain the drop-off in performance with the decimation ratio q. The weakness of conditioning analysis is that an ill-conditioned system may be ill-conditioned because of a single "almost singular value." As indicated by the rapid growth in the dimension of the null space in Theorem 1, super-resolution has a large number of "almost singular values" for large q. This is the real cause of the difficulties seen in Figure 10.4, as we now show.

3.3. Analysis of the Volume of Solutions

If we could work with noiseless, real-valued quantities and perform arbitrary precision arithmetic then the fact that the reconstruction constraints are ill-conditioned might not be a problem. In reality, however, images are always intensity discretized (typically to 8-bit values in the range 0–255 grey levels.) There will therefore always be noise in the measurements, even if it is only plus-or-minus half a grey-level. Suppose

that int[·] denotes the operator which takes a real-valued irradiance measurement and turns it into an integer-valued intensity. If we incorporate this quantization into our image formation model, the reconstruction constraints in Equation (10.13) become:

$$x_L^{(k)}(i,j) = \text{int}\left[\int_{x_H} \frac{x_H(x,y)}{q^2} h^{(k)}\left(\frac{1}{q}(x,y) + (c^{(k)}, d^{(k)}) - (i,j)\right) dx\, dy\right]. \tag{10.14}$$

Suppose also that x_H is a finite size image with n pixels. We then have:

Theorem 3 *The volume of solutions of the intensity discretized reconstruction constraints in Equation (10.14) grows asymptotically at least as fast as* $(q \cdot S)^{2 \cdot n}$.

Proof: First note that the space of solutions is convex since integration is linear. Next note that one solution of Equation (10.14) is the solution of:

$$x_L^{(k)}(i,j) - 0.5 = \int_{x_H} \frac{x_H(x,y)}{q^2} h^{(k)}\left(\frac{1}{q}(x,y) + (c^{(k)}, d^{(k)}) - (i,j)\right) dx\, dy. \tag{10.15}$$

The definition of the point spread function as $h^{(k)} = \omega^{(k)} * a^{(k)}$ and the properties of the convolution give $0 \leq h^{(k)} \leq 1/S^2$. Therefore, adding $(q \cdot S)^2$ to any pixel in x_H is still a solution since the right hand side of Equation (10.15) increases by at most 1. (The integrand is increased by less than 1 grey-level in the pixel, which only has an area of 1 unit.) The volume of solutions of Equation (10.14) therefore contains an n-dimensional simplex, where the angles at one vertex are all right-angles, and the sides are all $(q \cdot S)^2$ units long. The volume of such a simplex grows asymptotically like $(q \cdot S)^{2n}$ (treating n as a constant and M and S as variables). The desired result follows. □

In Figures 10.4 and 10.5 we present results to illustrate Theorems 2 and 3. We took a high resolution image of a face and translated it by random sub-pixel amounts, blurred it with a Gaussian, and then down-sampled it. We repeated this procedure for several decimation ratios; q = 2, 4, 8, and 16. In each case, we generated multiple down-sampled images, each with a different translation. We generated enough images so that there were as many low resolution pixels in total as pixels in the original high resolution image. For example, we generated 4 half size images, 16 quarter size images, and so on. We then applied the algorithms of (Hardie et al., 1997) and (Schultz and Stevenson, 1996).

The results for (Hardie et al., 1997) are shown in the figure. The results for (Schultz and Stevenson, 1996) were very similar and are omitted. We provided the algorithms with exact knowledge of both the

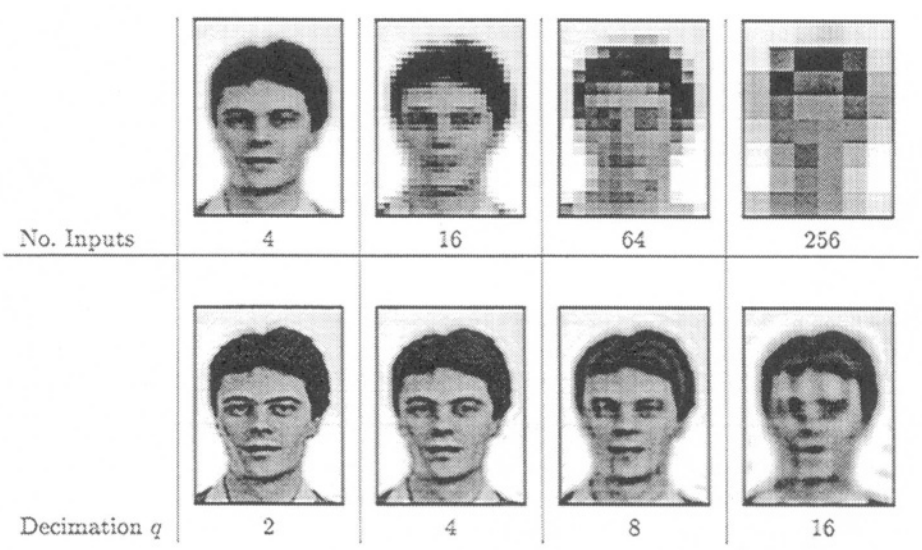


Figure 10.4. Results of the reconstruction-based super-resolution algorithm (Hardie et al., 1997) for various decimation ratios. A high high-resolution image of a face is translated multiple times by random sub-pixel amounts, blurred with a Gaussian, and then down-sampled. (The algorithm is provided with exact knowledge of the point spread function and the sub-pixel translations.) Comparing the images in the right-most column, we see that the algorithm does quite well given the very low resolution of the input. The degradation in performance as the decimation ratio increases from left to right is very dramatic, however.

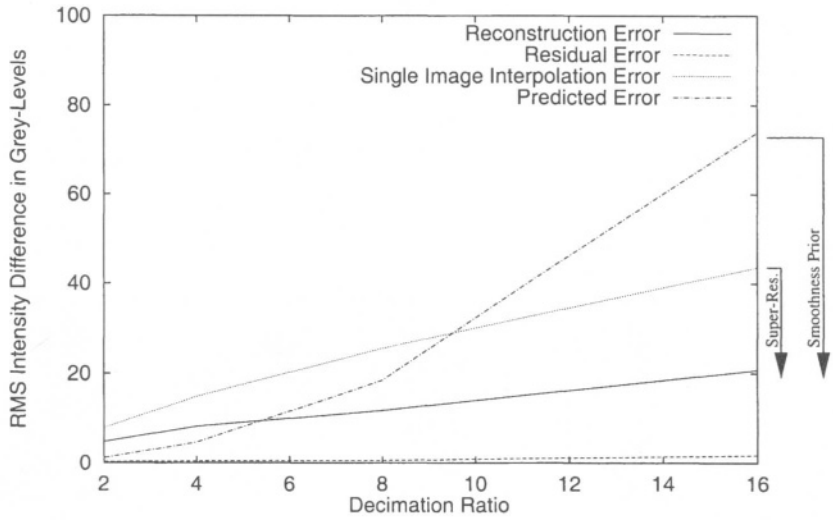


Figure 10.5. An illustration of Theorems 2 and 3 using the same inputs as in Figure 10.4. The reconstruction error is much higher than the residual, as would be expected for an ill-conditioned system. For low decimation ratios, the prior is unnecessary and so the results are worse than predicted. For high decimation ratios, the prior does help, but at the price of smooth results. (See Figure 10.4.) An estimate of the amount of information provided by the reconstruction constraints is given by the improvement of the reconstruction error over the interpolation error. Similarly, the improvement from the predicted error to the reconstruction error is an estimate of the amount of information provided by the smoothness prior. By this measure, the smoothness prior provides more information than the reconstruction constraints for $q = 16$.

point spread function used in the down-sampling and the random subpixel translations. Restricting attention to the right-most column of Figure 10.4, the results look very good. The algorithm is able to do a decent job of reconstructing the face from input images which barely resemble faces. On the other hand, the performance gets much worse as the decimation ratio increases (from left to right.)

Our third and final theorem provides the best explanation of these results. For large decimation ratios q = 8 and 16, there is a huge volume of solutions to the discretized reconstruction constraints in Equation (10.14). The smoothness prior which is added to resolve this ambiguity simply ensures that it is one of the overly smooth solutions that is chosen. (Of course, without the prior any solution might be chosen which would generally be even worse.)

Using the same inputs as Figure 10.4, we plot the reconstruction error against the decimation ratio in Figure 10.5; i.e. the difference between the reconstructed high resolution image and the original. We compare this error with the residual error; i.e. the difference between the low resolution inputs and their predictions from the reconstructed high resolution image. As expected for an ill-conditioned system, the reconstruction error is much higher than the residual. We also compare with a rough prediction of the reconstruction error obtained by multiplying the lower bound on the condition number $(q \cdot S)^2$ by an estimate of the expected residual assuming that the grey-levels are discretized from a uniform distribution. For low decimation ratios, this estimate is an under-estimate because the prior is unnecessary for noise free data; i.e. better results would be obtained without the prior. For high decimation ratios the prediction is an over-estimate because the local smoothness assumption does help the reconstruction (albeit at the expense of overly smooth results.)

We also plot interpolation results in Figure 10.5; i.e. just using the reconstruction constraints for one image (as was proposed, for example, in (Schultz and Stevenson, 1994).) The difference between this curve and the reconstruction error curve is a measure of how much information the reconstruction constraints provide. Similarly, the difference between the predicted error and the reconstruction error is a measure of how much information the smoothness prior provides. For a decimation ratio of 16, we see that the prior provides more information than the super-resolution reconstruction constraints.

4. Super-Resolution by Hallucination

How then is it possible to perform super-resolution with a high decimation ratio without the results looking overly smooth? As we have just shown, the required high-frequency information was lost from the reconstruction constraints when the input images were discretized to 8-bit values. Smoothness priors may help regularize the problem, but cannot replace the missing information.

Our goal in this section is to develop a super-resolution algorithm which uses the information contained in a collection of recognition decisions (in addition to the reconstruction constraints.) Our approach (which we call *hallucination*) is to embed the results of the recognition decisions in a *recognition-based prior* on the solution of the reconstruction constraints, thereby hopefully resolving the inherent ambiguity in their solution.

Our approach is somewhat related to that of (Freeman and Pasztor, 1999) who recently, and independently, proposed a learning framework for low-level vision, one application of which is image interpolation. Besides being applicable to an arbitrary number of images, the other major advantage of our approach is that it uses a prior which is specific to the class of object (in the "class-based" sense of (Riklin-Raviv and Shashua, 1999)) and a set of local recognition decisions. Our algorithm is also related to (Edwards et al., 1998), in which active-appearance model are used for model-based super-resolution.

4.1. Bayesian MAP Formulation

We use a Bayesian formulation of super-resolution (Cheeseman et al., 1994; Schultz and Stevenson, 1996; Hardie et al., 1997; Elad and Feuer, 1997). In this approach, super-resolution is posed as finding the maximum *a posteriori* (or MAP) super-resolution image x_H: i.e. estimating $\arg\max_{\mathbf{x}_H} \mathrm{P}[\mathbf{x}_H \mid \mathbf{x}_L^{(k)}]$. Bayes law for this estimation problem is:

$$P[x_H \mid x_L^{(k)}] = \frac{P[x_L^{(k)} \mid x_H] \cdot P[x_H]}{P[x_L^{(k)}]}. \tag{10.16}$$

Since $P[x_L^{(k)}]$ is a constant because the images $x_L^{(k)}$ are (known) inputs, and since the logarithm function is a monotonically increasing function, we have:

$$\arg\max_{\mathbf{x}_H} \mathrm{P}[\mathbf{x}_H \mid \mathbf{x}_L^{(k)}] \;=\; \arg\min_{\mathbf{x}_H} \left(-\ln \mathrm{P}[\mathbf{x}_L^{(k)} \mid \mathbf{x}_H] - \ln \mathrm{P}[\mathbf{x}_H] \right). \tag{10.17}$$

The first term in this expression $-\ln P[x_L^{(k)} \mid x_H]$ is the (negative log) probability of reconstructing the low resolution images $x_L^{(k)}$ given that the super-resolution image is x_H. It is therefore normally set to be a quadratic (i.e. energy) function of the error in the reconstruction constraints:

$$-\ln P[x_L^{(k)} \mid x_H] = \frac{1}{2\sigma_\eta^2} \sum_{i,j,k} \left[x_L^{(k)}(i,j) - \sum_{i',j'} W^{(k)}(i,j,i',j') \cdot x_H(i',j') \right]^2 \tag{10.18}$$

where $W^{(k)}(i,j,i',j')$ is defined in Equation (10.9). In this expression, we are implicitly assuming that the noise is independently and identically distributed (across both the images and the pixels) and is Gaussian with covariance σ_η^2.

4.2. Recognition-Based Priors

The second term on the right-hand side of Equation (10.17) is (the negative logarithm of) the prior $-\ln P[x_H]$. Usually the prior is chosen to be a simple smoothness prior (Cheeseman et al., 1994; Schultz and Stevenson, 1996; Hardie et al., 1997; Elad and Feuer, 1997). Instead, we would like it to depend upon the results of a set of recognition decisions. Suppose the outputs of the recognition decisions partition the inputs (i.e. the low resolution input images $x_L^{(k)}$) into a set of subclasses $\{C_{k,l} \mid l = 1, 2, \ldots\}$. We then define a *recognition-based prior* as one that can be written in the following form:

$$P[x_H] = \sum_l P[x_H \mid x_L^{(k)} \in C_{k,l}] \cdot P[x_L^{(k)} \in C_{k,l}]. \tag{10.19}$$

Essentially there is a separate prior $P[x_H \mid x_L^{(k)} \in C_{k,l}]$ for each possible partition $C_{k,l}$ of the input space. Once the low resolution input images $x_L^{(k)}$ are available, the various recognition algorithms can be applied, and it can be determined which partition the inputs lie in. The recognition-based prior $P[x_H]$ then reduces to the more specific prior $P[x_H \mid x_L^{(k)} \in C_{k,l}]$. This prior can be made more powerful than the overall prior because it can be tailored to $C_{k,l}$.

4.3. Multi-Scale Derivative Features

We decided to try to recognize generic local image features (rather than higher level concepts such as ASCII characters) because we want to apply our algorithm to a variety of phenomena. Motivated by (De Bonet, 1997), we also decided to use multi-scale features. In particular, given an image x, we first form its Gaussian pyramid $G_0(x), \ldots, G_N(x)$ (Burt, 1980). Afterwards, we also form its Laplacian pyramid $L_0(x), \ldots, L_N(x)$ (Burt and Adelson, 1983), the horizontal $H_0(x), \ldots, H_N(x)$ and vertical $V_0(x), \ldots, V_N(x)$ first derivatives of the Gaussian pyramid, and the horizontal $H_0^2(x), \ldots, H_N^2(x)$ and vertical $V_0^2(x), \ldots, V_N^2(x)$ second derivatives of the Gaussian pyramid. (See Figure 10.6 for examples of these pyramids.) Finally, we form a feature pyramid:

$$\mathbf{F_l(x)} = \left(\mathbf{L_l(x)}, \mathbf{H_l(x)}, \mathbf{V_l(x)}, \mathbf{H_l^2(x)}, \mathbf{V_l^2(x)}\right) \quad \text{for } \mathrm{l} = 0, \ldots, \mathrm{N}. \tag{10.20}$$

The pyramid $\mathbf{F_0(x)}, \ldots, \mathbf{F_N(x)}$ is a pyramid where there are 5 values stored at each pixel, the Laplacian and the 4 derivatives.

Then, given a pixel in the low resolution image that we are performing super-resolution on, we want to find (i.e. recognize) a pixel in a collection

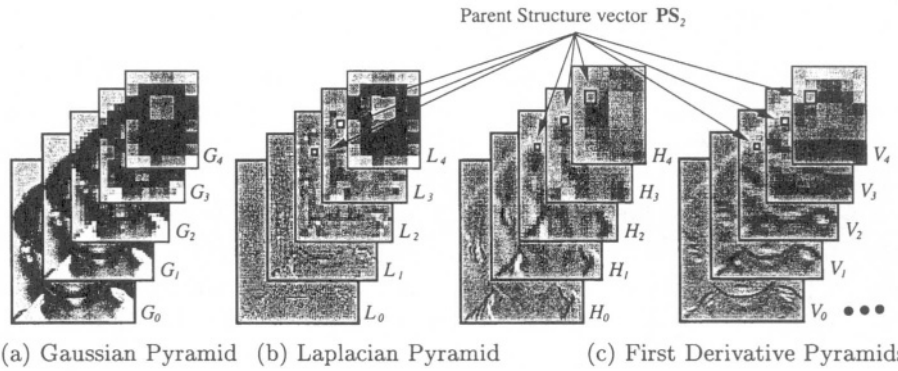


Figure 10.6. The Gaussian, Laplacian, and first derivative pyramids of an image of a face. (We also use two second derivatives but omit them from the figure.) We combine these pyramids into a single multi-valued pyramid, where we store a vector of the Laplacian and the derivatives at each pixel. The Parent Structure vector $\mathbf{PS}_l(i,j)$ of a pixel (i,j) in the l^{th} level of the pyramid consists of the vector of values for that pixel, the vector for its parent in the $l+1^{th}$ level, the vector for its parent's parent, etc (De Bonet, 1997). The Parent Structure vector is therefore a high-dimensional vector of derivatives computed at various scales. In our algorithms, recognition means finding the training sample with the most similar Parent Structure vector.

of training data that is locally "similar." By similar, we mean that both the Laplacian and the image derivatives are approximately the same, at all scales. To capture this notion, we define the Parent Structure vector (De Bonet, 1997) of a pixel (i,j) in the l^{th} level of the feature pyramid $\mathbf{F_0(x)},\ldots,\mathbf{F_N(x)}$ to be: $\mathbf{PS_l(x)(i,j)} =$

$$\left(\mathbf{F_l(x)(i,j)},\mathbf{F_{l+1}(x)}\left(\left\lfloor\frac{\mathbf{i}}{\mathbf{2}}\right\rfloor,\left\lfloor\frac{\mathbf{j}}{\mathbf{2}}\right\rfloor\right),\ldots,\mathbf{F_N(x)}\left(\left\lfloor\frac{\mathbf{i}}{\mathbf{2^{N-l}}}\right\rfloor,\left\lfloor\frac{\mathbf{j}}{\mathbf{2^{N-l}}}\right\rfloor\right).\right) \tag{10.21}$$

As illustrated in Figure 10.6, the Parent Structure vector at a pixel in the pyramid consists of the feature vector at that pixel, the feature vector of the parent of that pixel, the feature vector of its parent, and so on. Exactly as in (De Bonet, 1997), our notion of two pixels being similar is then that their Parent Structure vectors are approximately the same (measured by some norm.)

4.4. Finding the Closest Parent Structure

Suppose we have a set of high resolution training images $x_T^{(m)}$ where $m=1,\ldots,M$. We first form feature pyramids $\mathbf{F_0}(\mathbf{x_T^{(m)}}),\ldots,\mathbf{F_N}(\mathbf{x_T^{(m)}})$. Also suppose that the input image $x_L^{(k)}$ is at a resolution that is $q=2^l$ times smaller than the training samples. (The image may have to be interpolated to make this ratio exactly a power of 2.) We can then com-

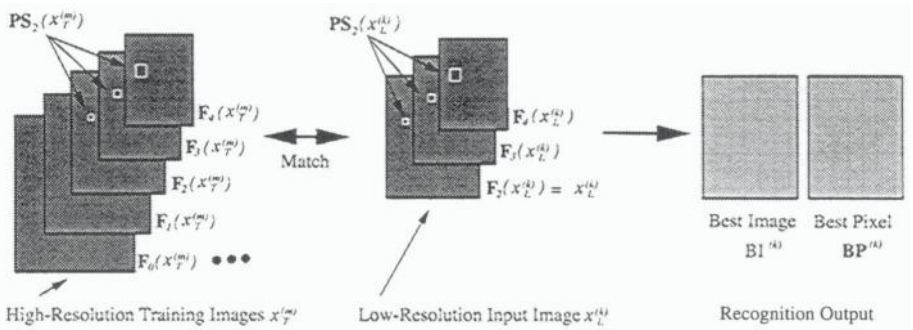


Figure 10.7. We compute the feature pyramids $\mathbf{F_0}(\mathbf{x_T^{(m)}}), \ldots, \mathbf{F_N}(\mathbf{x_T^{(m)}})$ for the training images $x_T^{(m)}$ and the feature pyramids $\mathbf{F_l}(\mathbf{x_L^{(k)}}), \ldots, \mathbf{F_N}(\mathbf{x_L^{(k)}})$ for the low resolution input images $x_L^{(k)}$. For each pixel in the low resolution images, we find (i.e. recognize) the closest matching Parent Structure in the high resolution data. We record and output the best matching image $\mathbf{BI^{(k)}}$ and the pixel location of the best matching Parent Structure $\mathbf{BP^{(k)}}$. Note that these data structures are both defined independently for each pixel (i, j) in each image $x_L^{(k)}$.

pute the feature pyramid for the input image from level l and upwards $\mathbf{F_k}(\mathbf{x_L^{(k)}}), \ldots, \mathbf{F_N}(\mathbf{x_L^{(k)}})$. Figure 10.7 shows an illustration of this scenario for $l = 2$.

Independently for each pixel (i, j) in the input $x_L^{(k)}$, we compare its Parent Structure vector $\mathbf{PS_l}(\mathbf{x_L^{(k)}})(\mathbf{i}, \mathbf{j})$ against all of the training Parent Structure vectors at the same level l; i.e. we compare against $\mathbf{PS_l}(\mathbf{x_T^{(m)}})(\mathbf{i'}, \mathbf{j'})$ for all m and for all (i', j'). The best matching image $\mathbf{BI^{(k)}}(\mathbf{i}, \mathbf{j}) = \mathbf{m}$ and the best matching pixel $\mathbf{BP^{(k)}}(\mathbf{i}, \mathbf{j}) = (\mathbf{i'}, \mathbf{j'})$ are stored as the output of the recognition decision, independently for each pixel (i, j) in $x_L^{(k)}$. (We found the performance to be largely independent of the distance function used to determine the best matching Parent Structure vector. We actually used a weighted $\mathbf{L^2}$-norm, giving the derivative components half as much weight as the Laplacian values and reducing the weight by a factor of 2 for each increase in the pyramid level.)

Recognition in our hallucination algorithm therefore means finding the closest matching pixel in the training data in the sense that the Parent Structure vectors of the the two pixels are the most similar. This search is, in general, performed over all pixels in all of the images in the training data. If we have frontal images of faces, however, we restrict this search to consider only the corresponding pixels in the training data. In this way, we treat each pixel in the input image differently, depending on its spatial location, similarly to the "class-based" approach of (Riklin-Raviv and Shashua, 1999).

4.5. The Recognition-based Gradient Prior

For each pixel (i, j) in the input image $x_L^{(k)}$, we have recognized the pixel that is the most similar in the training data, specifically, the pixel $\mathbf{BP}^{(\mathbf{k})}(\mathbf{i},\mathbf{j})$ in the l^{th} level of the pyramid for training image $x_T^{(\mathrm{BI}^{(\mathbf{k})}(\mathbf{i},\mathbf{j}))}$. These recognition decisions partition the inputs into a collection of subclasses, as required by the recognition-based prior described in Section 4.2. If we denote the subclasses by $C_{k,\mathbf{BP}^{(\mathbf{k})},\mathrm{BI}^{(\mathbf{k})}}$ (i.e. using a multi-dimensional index rather than l) Equation (10.19) can be rewritten as:

$$P[x_H] = \sum_{\mathbf{BP}^{(\mathbf{k})},\mathrm{BI}^{(\mathbf{k})}} P[x_H \mid x_L^{(k)} \in C_{k,\mathbf{BP}^{(\mathbf{k})},\mathrm{BI}^{(\mathbf{k})}}] \cdot P[x_L^{(k)} \in C_{k,\mathbf{BP}^{(\mathbf{k})},\mathrm{BI}^{(\mathbf{k})}}] \tag{10.22}$$

where $P[x_H \mid x_L^{(k)} \in C_{k,\mathbf{BP}^{(\mathbf{k})},\mathrm{BI}^{(\mathbf{k})}}]$ is the probability that the super-resolution image is x_H, given that the inputs $x_L^{(k)}$ lie in the subclass that will be recognized to have $\mathbf{BP}^{(\mathbf{k})}$ as the best matching pixel in training image $x_T^{(\mathrm{BI}^{(\mathbf{k})}(\mathbf{i},\mathbf{j}))}$.

We now need to define $P[x_H \mid x_L^{(k)} \in C_{k,\mathbf{BP}^{(\mathbf{k})},\mathrm{BI}^{(\mathbf{k})}}]$. We decided to make this recognition-based prior a function of the gradient because the base, or average, intensities in the super-resolution image are defined by the reconstruction constraints. It is the high-frequency gradient information that is missing. So, we want to define the prior to encourage the gradient of the super-resolution image to be close to the gradient of the closest matching training samples.

Each low resolution input image $x_L^{(k)}$ has a (different) closest matching (Parent Structure) training sample for each pixel. Moreover, each such Parent Structure corresponds to a number of different pixels in the 0^{th} level of the pyramid, (2^l of them to be precise. See also Figure 10.7.) We therefore impose a separate gradient constraint for each pixel (i, j) in the 0^{th} level of the pyramid (and for each input image $x_L^{(k)}$.) The best matching pixel $\mathbf{BP}^{(\mathbf{k})}$ is only defined on the l^{th} level of the pyramid. For notational convenience, therefore, given a pixel (i, j) on the 0^{th} level of the pyramid, define the best matching pixel on the 0^{th} level of the pyramid to be:

$$\overline{\mathbf{BP}}^{(\mathbf{k})}(\mathbf{i},\mathbf{j}) \equiv \mathbf{2^l} * \mathbf{BP}^{(\mathbf{k})}\left(\left\lfloor \frac{\mathbf{i}}{\mathbf{2^l}} \right\rfloor, \left\lfloor \frac{\mathbf{j}}{\mathbf{2^l}} \right\rfloor\right) + (\mathbf{i},\mathbf{j}) - \mathbf{2^l} * \left(\left\lfloor \frac{\mathbf{i}}{\mathbf{2^l}} \right\rfloor, \left\lfloor \frac{\mathbf{j}}{\mathbf{2^l}} \right\rfloor\right). \tag{10.23}$$

Also define the best matching image as $\overline{\mathrm{BI}}^{(\mathbf{k})}(\mathbf{i},\mathbf{j}) \equiv \mathrm{BI}^{(\mathbf{k})}\left(\left\lfloor \frac{\mathbf{i}}{\mathbf{2^l}} \right\rfloor, \left\lfloor \frac{\mathbf{j}}{\mathbf{2^l}} \right\rfloor\right)$.

If (i, j) is a pixel in the 0^{th} level of the pyramid for image $x_L^{(k)}$, the corresponding pixel in the super-resolution image x_H is $\left(\mathbf{r}^{(\mathbf{k})}\right)^{-1}\left(\frac{i}{2^l}, \frac{j}{2^l}\right)$.

We therefore want to impose the constraint that the first derivatives of x_H at this point should equal the derivatives of the closest matching pixel (Parent Structure) in the training data. Parametric expressions for $H_0(x_H)$ and $V_0(x_H)$ at $\left(\mathbf{r}^{(\mathbf{k})}\right)^{-1}(\frac{i}{2^l}, \frac{j}{2^l})$ can easily be derived as linear functions of the unknown pixels in the high resolution image x_H. We assume that the errors in the gradient values between the recognized training samples and the super-resolution image are independently and identically distributed and moreover that they are Gaussian with covariance σ_∇^2. Therefore: $P[x_H \mid x_L^{(k)} \in C_{k,\mathbf{BP}^{(\mathbf{k})},\mathrm{BI}^{(\mathbf{k})}}] =$

$$\frac{1}{2\sigma_\nabla^2}\left(\sum_{i,j,k}\left[H_0(x_H)(\left(\mathbf{r}^{(\mathbf{k})}\right)^{-1}(\frac{i}{2^l},\frac{j}{2^l})) - H_0(x_T^{\overline{\mathrm{BI}}^{(\mathbf{k})}(\mathrm{i,j})})(\overline{\mathbf{BP}}^{(\mathbf{k})}(\mathbf{i},\mathbf{j}))\right]^2\right.$$
$$\left.\sum_{i,j,k}\left[V_0(x_H)(\left(\mathbf{r}^{(\mathbf{k})}\right)^{-1}(\frac{i}{2^l},\frac{j}{2^l})) - V_0(x_T^{\overline{\mathrm{BI}}^{(\mathbf{k})}(\mathrm{i,j})})(\overline{\mathbf{BP}}^{(\mathbf{k})}(\mathbf{i},\mathbf{j}))\right]^2\right). \tag{10.24}$$

This prior enforces the constraints that the gradient of the super resolution image x_H should equal to the gradient of the best matching training image.

4.6. Algorithm Practicalities

Equations (10.17), (10.18), (10.22), and (10.24) form a high dimensional linear least squares problem. The constraints in Equation (10.18) are the standard super-resolution reconstruction constraints. Those in Equation (10.24) are the recognition-based prior. The relative weights of these constraints are defined by the noise covariances σ_η^2 and σ_∇^2. We assume that the reconstruction constraints are the more reliable ones and so set $\sigma_\eta^2 \ll \sigma_\nabla^2$.

The number of unknowns is equal to the number of pixels in x_H. Inverting a linear system of such a size can prove problematic. We therefore implemented a gradient descent algorithm using the standard diagonal approximation to the Hessian (Press et al., 1992) to set the step size in a similar way to (Szeliski and Golland, 1998). Since the error function is quadratic, the algorithm converges to the (single) global minimum without any problem.

4.7. Experimental Results on Human Faces

Our experiments for human face images were conducted with a subset of the FERET dataset (Philips et al., 1997) consisting of 596 images of

278 individuals (92 women and 186 men). Most people appear twice, with the images taken on the same day under similar illumination conditions, but with different expressions (one expression is neutral, the other typically a smile.) A small number of people appear 4 times, with the images separated by several months.

The images in the FERET dataset are 256 × 384 pixels, however the area occupied by the face varies considerably, but most of the faces are around 96 × 128 pixels or larger. In the class-based approach (Riklin-Raviv and Shashua, 1999), the input images (which are all frontal) need to be aligned so that we can assume that the same part of the face appears in roughly the same part of the image every time. This alignment was performed by hand marking the location of 3 points, the centers of the two eyes and the lower tip of the nose. These 3 points define an affine warp (Bergen et al., 1992), which was used to warp the images into a canonical form. These canonical 96 × 128 pixel images were then used as the training samples $x_T^{(m)}$ where $m = 1, \ldots 596$.

We used a "leave-one-out" methodology to test our algorithm. To test on any particular person, we removed all occurrences of that individual from the training set. We then trained the algorithm on the reduced training set, and tested on the images of the individual that had been removed. Because this process is quite time consuming, we used a test set of 100 randomly selected images of 100 different individuals rather than the entire training set.

Comparison with Existing Super-Resolution Algorithms

We initially restrict attention to the case of enhancing 24 × 32 pixel images four times to give 96 × 128 pixel images. Later we will consider the variation in performance with the decimation ratio. We simulate the multiple slightly translated images required for super-resolution using the FERET database by randomly translating the original FERET images multiple times by sub-pixel amounts before down-sampling them to form the low resolution input images.

In our first set of experiments we compare our algorithm with those of (Hardie et al., 1997) and (Schultz and Stevenson, 1996). In Figure 10.8(a) we plot the RMS pixel error against the number of low resolution inputs, computed over the 100 image test set. (We compute the RMS error using the original high resolution image used to synthesize the inputs from.) We also plot results for cubic B-spline interpolation (which only uses one image) for comparison.

In Figure 10.8(a) we see that our hallucination algorithm does outperform the reconstruction-based super-resolution algorithms, from one input image to 25. The improvement is consistent across the number

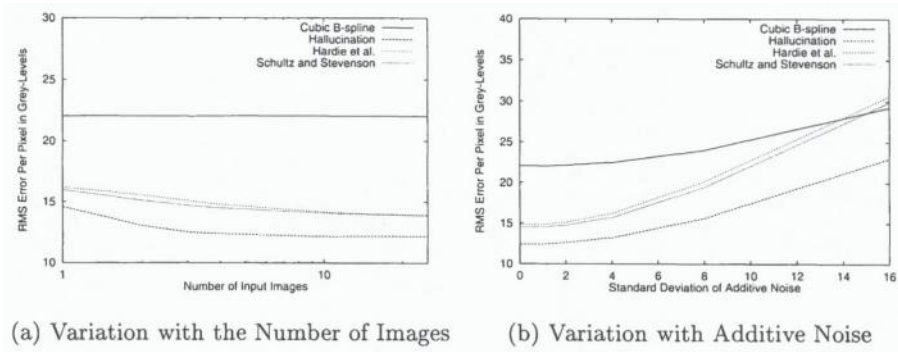


Figure 10.8. A comparison of our hallucination algorithm with the reconstruction-based super-resolution algorithms of (Schultz and Stevenson, 1996) and (Hardie et al., 1997). In (a) we plot the RMS pixel intensity error computed across the 100 image test set against the number of low resolution input images. Our algorithm outperforms the the traditional super-resolution algorithms across the entire range. In (b) we vary the amount of additive noise. Again we find that our algorithm does better than the traditional super-resolution algorithms.

of input images and is around 20%. The improvement is also largely independent of the actual input. In particular, Figure 10.9 contains the best and worst results obtained across the entire test set in terms of the RMS error of the hallucination algorithm for 9 low resolution inputs. As can be seen, there is little difference between the best results in Figure 10.9(a)–(d) and the worst ones in (e)–(g). Notice, also, how the hallucinated results are a dramatic improvement over the low resolution input, and moreover are visibly sharper than the results for Hardie *et al.*.

Robustness to Additive Intensity Noise

Figure 10.8(b) contains the results of an experiment investigating the robustness of the 3 super-resolution algorithms to additive noise. In this experiment, we added zero-mean, white Gaussian noise to the low resolution images before passing them as inputs to the algorithms. In the figure, the RMS pixel intensity error is plotted against the standard deviation of the additive noise. The results shown are for 4 low resolution input images, and again, the results are an average over the 100 image test set. As might be expected, the performance of all 4 algorithms gets much worse as the standard deviation of the noise increases. The hallucination algorithm and cubic B-spline interpolation, however, seem somewhat more robust than the reconstruction-based super-resolution algorithms. The reason for this increased robustness is probably that

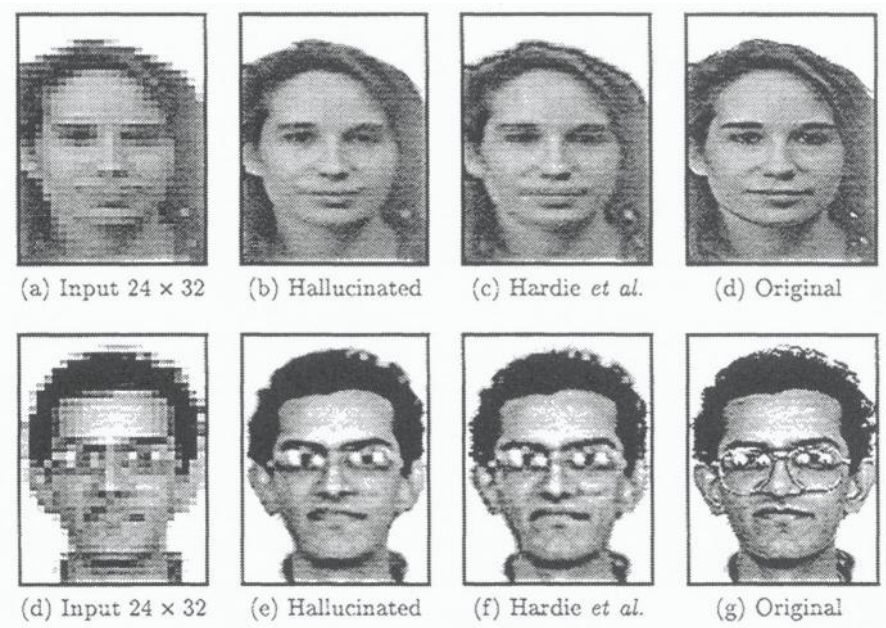


Figure 10.9. The best and worst results in Figure 10.8(a) in terms of the RMS error of the hallucination algorithm for 9 input images. In (a)–(d) we display the results for the best performing image in the 100 image test set. The results for the worst image are presented in (e)–(g). (The results for Schultz and Stevenson are similar to those for Hardie *et al.* and are omitted.) There is little difference in image quality between the best and worst hallucinated results.

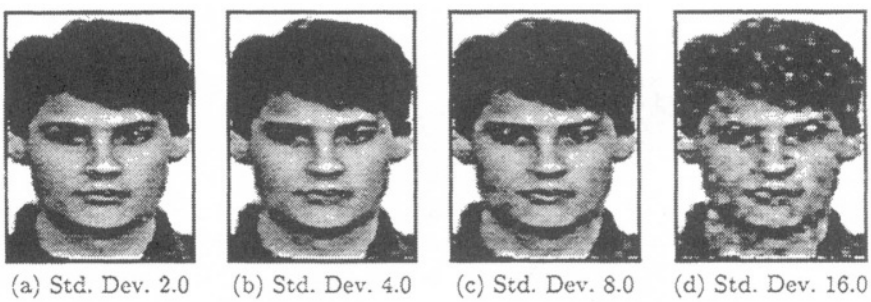


Figure 10.10. An example from Figure 10.8(b) of the variation in the performance of the hallucination algorithm with additive zero-mean, white Gaussian noise. As can be seen, the output is hardly affected until around 4-bits of intensity noise have been added to the inputs. The reason the hallucination algorithm is so robust to noise it that it uses the strong recognition-based face prior to generate smooth, face-like images however noisy the inputs are.

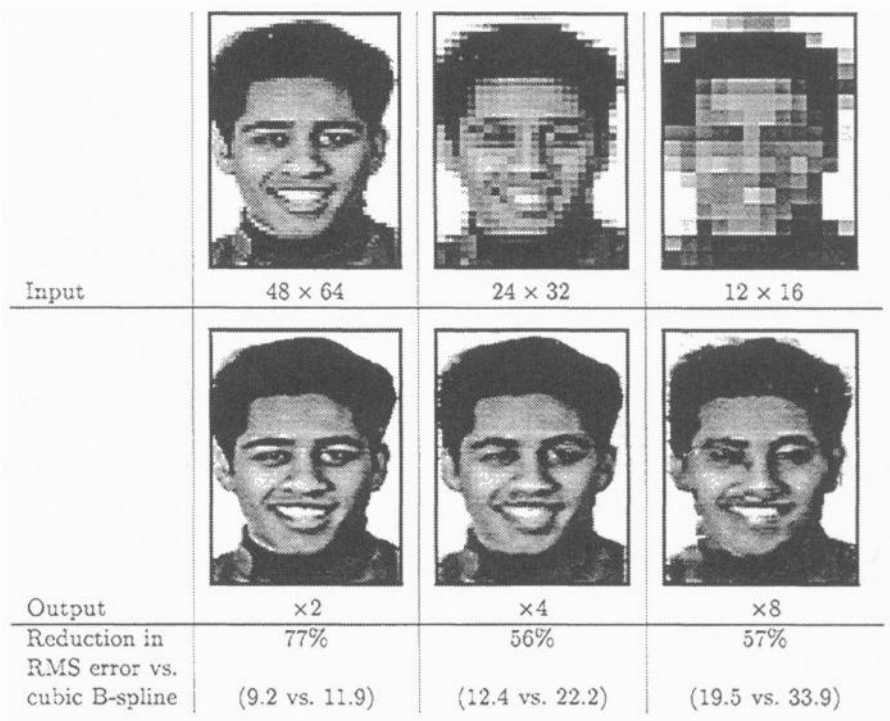


Figure 10.11. The variation in the performance of our hallucination algorithm with the input image size. We see that the algorithm works well down to 12 × 16 pixel images. It begins to break down for 6 × 8 pixel images. See (Baker and Kanade, 1999) for examples.

the hallucination algorithm always tends to generate smooth, face-like images (because of the strong recognition-based prior) however noisy the inputs are. One example of how the hallucination algorithm degrades with the amount of additive noise is presented in Figure 10.10.

Variation in Performance with the Input Image Size

We do not expect our hallucination algorithm to work for all sizes of input. Once the input gets too small, the recognition decisions will be based on essentially no information. In the limit that the input image is just a single pixel, the algorithm will always generate the same face (for a single input image), but with different average grey levels. We therefore investigated the lowest resolution at which our hallucination algorithm works reasonable well.

In Figure 10.11 we show example results for one face in the test set for 3 different input sizes. (All of the results use just 4 input images.)

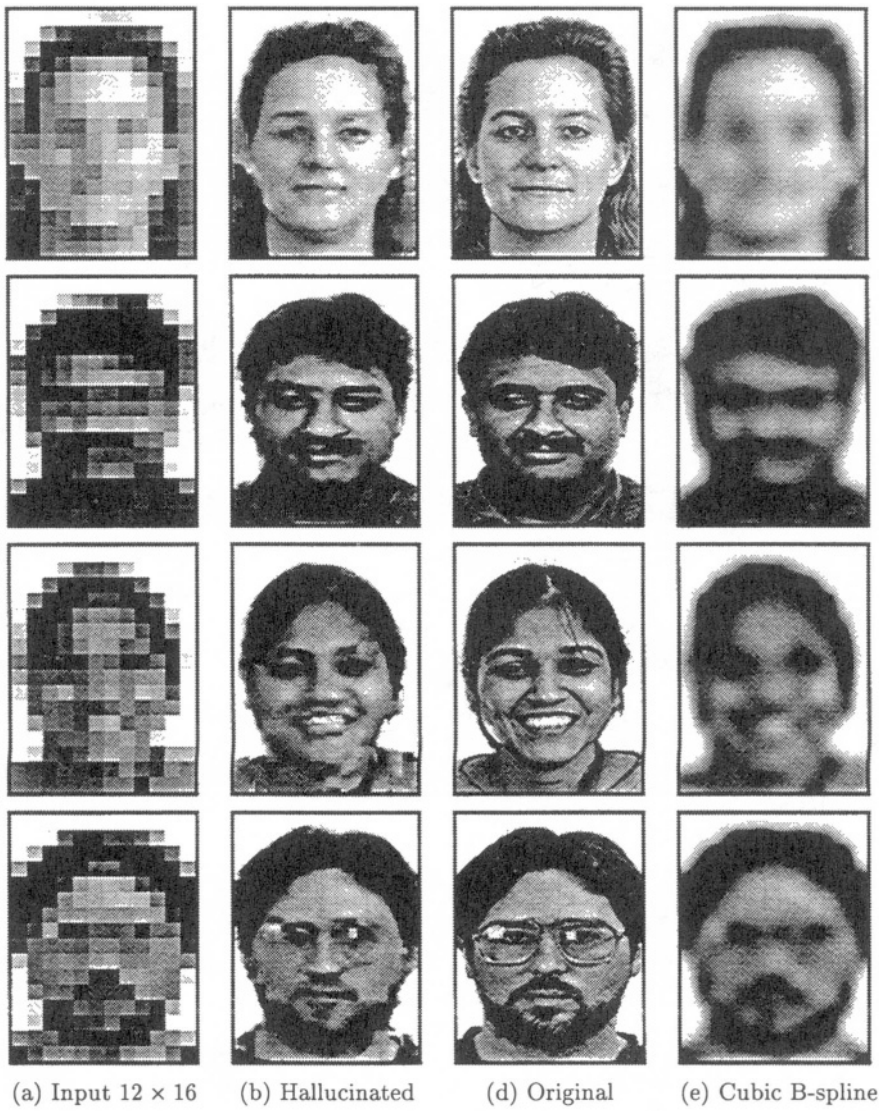


Figure 10.12. Selected results for 12 × 16 pixel images, the smallest input size for which our hallucination algorithm works reliably. (The input consists of only 4 low resolution input images.) Notice how sharp the hallucinated results are. See (Baker and Kanade, 1999) for the results of (Hardie et al., 1997) which are omitted due to lack of space.

We see that the algorithm works reasonably well down to 12 × 16 pixels. (For 6 × 8 pixel images it produces a face that appears to be a pieced-together combination of a variety of faces. See (Baker and Kanade, 1999) for examples.)

In the last row of Figure 10.11, we give numerical results of the average improvement in the RMS error over cubic B-spline interpolation (computed over the 100 image test set.) We see that for 24 × 32 and 12 × 16 pixel images, the reduction in the error is very dramatic. It is roughly halved. For 48 × 64 pixel images, the RMS is only cut by about 25% because cubic B-spline does so well it is hard to do much better.

The results for the 12 × 16 pixel image are excellent, however. (Also see Figure 10.12 which contains several more examples.) The input images are barely recognizable as faces and the facial features such as the eyes, eye-brows, and mouths only consist of a handful of pixels. The outputs, albeit slightly noisy, are clearly recognizable to the human eye. The facial features are also clearly discernible. The hallucinated results are also a huge improvement over (Hardie et al., 1997) and (Schultz and Stevenson, 1996). See (Baker and Kanade, 1999) for these results which are omitted due to a lack of space.

Results on Non-FERET Test Images

In our final experiment for human faces, we tried our algorithm on an image not in the FERET dataset. The results in Figure 10.13 give a big improvement over the cubic B-spline interpolation algorithm. The facial features, such as the eyes, nose, and mouth are all enhanced and appear much sharper in the hallucinated result that either in the input or in the interpolated image.

Results on Images Not Containing Faces

In Figure 10.14 we briefly present a few results on images that do not contain faces, even though the algorithm has been trained on the FERET dataset. (Figure 10.14(a) is a miscellaneous image and Figure 10.14(c) is a constant image.) As might be expected, our algorithm hallucinates an outline of a face in both cases, even though there is no face in the input. This is the reason we called our algorithm a "hallucination algorithm."

4.8. Experimental Results on Text Data

We also applied our algorithm to text data. In particular, we grabbed an image of an window displaying one page of a letter and used the bit-map as the input. The image was split into disjoint training and test samples. The results are presented in Figures 10.15. The input in Figure 10.15(a) is half the resolution of the original in Figure 10.15(f).

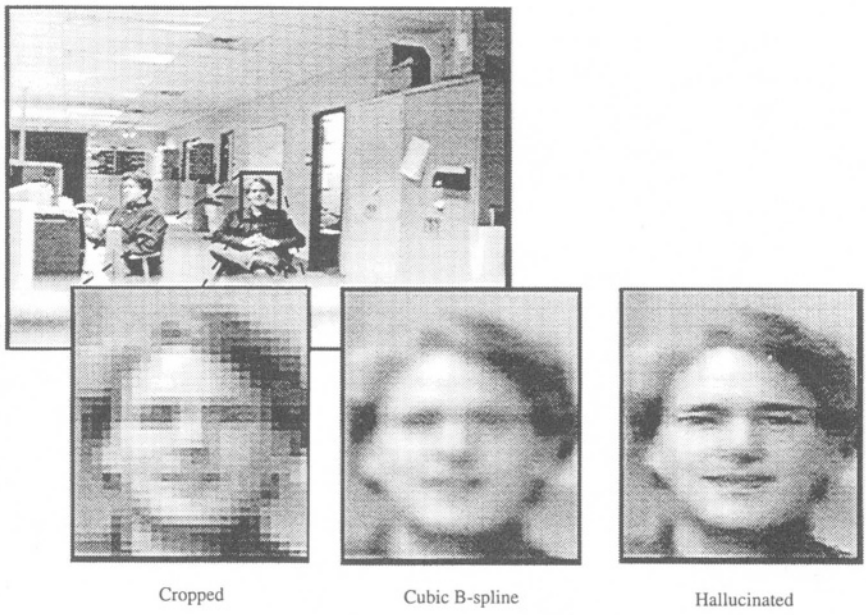


Figure 10.13. Example results on a face not in the FERET dataset. The facial features, such as eyes, nose, and mouth, which are blurred and unclear in the original cropped face, are enhanced and appear much sharper in the hallucinated image. The cubic B-spline result is overly smooth.

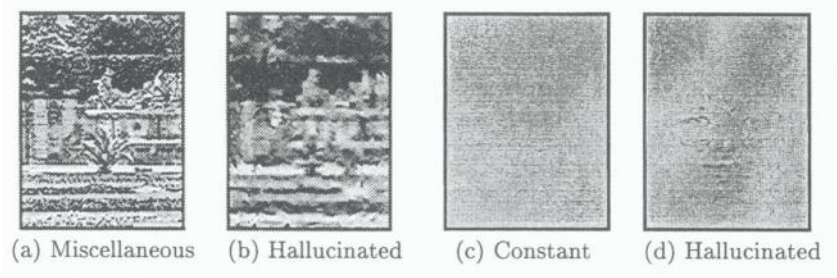


Figure 10.14. The results of applying our algorithm to images not containing faces. (We have omitted the low resolution input and just display the high resolution one.) A face is hallucinated by our algorithm even when none is present, hence the term "hallucination."

The hallucinated result in Figure 10.15 (c) is the best reconstruction of the text, both visually and in terms of the RMS intensity error. For example, compare the appearance of the word "was" in the second sentence in Figures 10.15(b)–(f). The hallucination algorithm also has an RMS error of only 24.5 grey levels, compared to over 48.0 for the other algorithms.

5. Summary

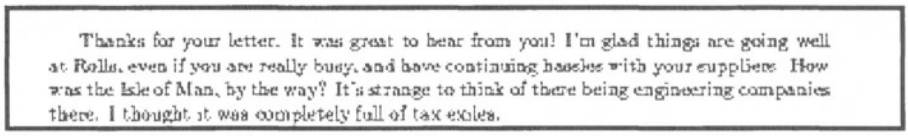
In the first half of this chapter we showed that the super-resolution reconstruction constraints provide less and less useful information as the decimation ratio increases. The major cause of this phenomenon is the spatial averaging over the photosensitive area; i.e. the fact that S is non-zero. The underlying reason that there are limits on reconstruction-based super-resolution is therefore the simple fact that CCD sensors must have a non-zero photosensitive area in order to be able to capture a non-zero number of light photons.

Our analysis assumes quantized noiseless images; i.e. the intensities are 8-bit values, created by rounding noiseless real-valued numbers. (It is this quantization that causes the loss of information, which when combined with spatial averaging, means that high decimation ratio super-resolution is not possible from the reconstruction constraints.) Without this assumption, however, it might be possible to increase the number of bits per pixel by averaging a collection of quantized noisy images (in an intelligent way). In practice, taking advantage of such information is very difficult. This point also does not affect another outcome of our analysis which was to show that reconstruction-based super-resolution inherently trades-off intensity resolution for spatial resolution.

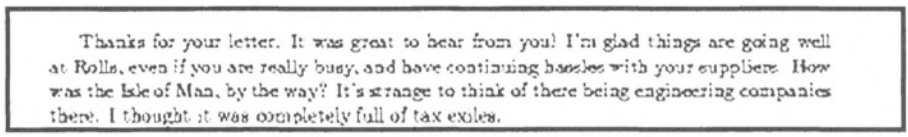
In the second half of this chapter we showed that recognition processes may provide an additional source of information for super-resolution algorithms. In particular, we developed a "hallucination" algorithm and demonstrated that this algorithm can obtain far better results than existing reconstruction-based super-resolution algorithms, both visually and quantitatively.

6. Discussion

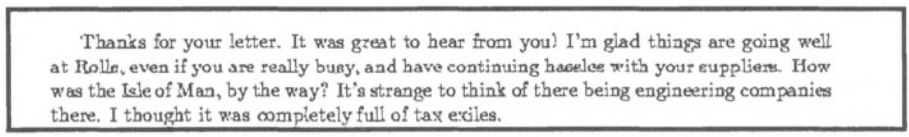
In the past 10-15 years or so much of the research on super-resolution has focused on the reconstruction constraints, and various way of incorporating simple smoothness priors to allow the constraints to be solved. It is a major accomplishment that most of this area is now fairly well understood. This does not mean that super-resolution is now a "solved" problem. As we have shown in this chapter, simply writing down the



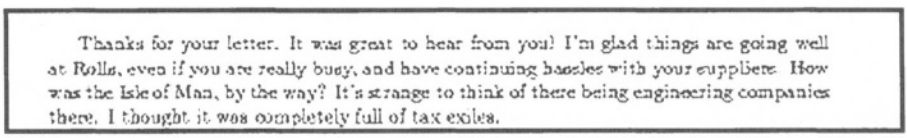
(a) Input Image. (Just one image is used.)



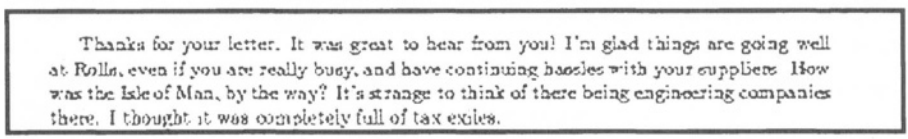
(b) Cubic B-spline, RMS Error 51.3



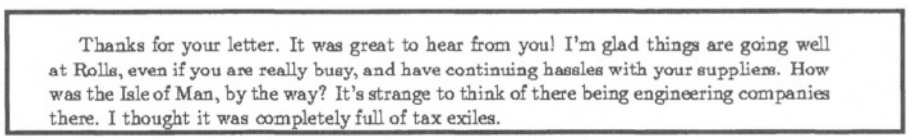
(c) Hallucinated, RMS Error 24.5x



(d) Schultz and Stevenson, RMS Error 48.4



(e) Hardie *et al.*, RMS Error 48.5



(f) Original High Resolution Image

Figure 10.15. The results of enhancing the resolution of a piece of text by a factor of 2. Our hallucination algorithm produces a clear, crisp image using no explicit knowledge that the input contains text. In particular, look at the word "was" in the second sentence. The RMS pixel intensity error is also almost a factor of 2 improvement over the other algorithms.

reconstruction constraints, adding a smoothness prior, and solving the resulting linear system does not necessarily mean that a good solution will be found. There are therefore a number of wide open areas for future super-resolution research:

- One such area involves conducting detailed analysis of the reconstruction constraints, when they provide additional information, how much additional information they provide, and how sensitive the information is to the signal to noise ratio of the input images. Some preliminary work has been done in this area, including (Elad and Feuer, 1997; Shekarforoush, 1999; Qi and Snyder, 2000; Baker and Kanade, 2000b). However, many issues are still a long way from being fully understood.

- Much of the work on super-resolution assumes a fairly simple image formation model. For example, there is almost no modeling of the effect of non-Lambertian surfaces and varying illumination. As a result, many algorithms (including the one described in this chapter) are very sensitive to illumination effects such as shadowing. Although some illumination invariant super-resolution algorithms have been proposed (Chiang and Boult, 1997), much more work remains to be done.

- In the second half of this chapter we proposed a hallucination algorithm. This algorithm is an instance of a model-based algorithm. Other examples include (Edwards et al., 1998; Freeman and Pasztor, 1999; Baker and Kanade, 2000a). These approaches appear very promising, however the area of model-based super-resolution is in its infancy and a great deal of work remains to be done for completely exploit the idea.

- Other areas which have been largely overlooked include the investigation of applications of super-resolution and the evaluation of the utility of super-resolution algorithms for those applications. There are two types of applications: (1) those where the enhanced image will be shown to a human, and (2) those where the enhanced image will be further processed by a machine. The evaluation of these two types of applications will be very different. The first will need to be done using rigorous subjective studies of how humans can make use of the super-resolution images. The second use of super-resolution is best evaluated in terms of the performance of the algorithms that will actually use the enhanced images. Both of these areas have barely been touched, even though they are vital for proving the utility of super-resolution as a whole.

Acknowledgements

We wish to thank Harry Shum for pointing out the reference (Freeman and Pasztor, 1999), Iain Matthews for pointing out (Edwards et al., 1998), and Henry Schneiderman for suggesting the conditioning analysis in Section 3.2. We would also like to thank a number of people for comments and suggestions, including Terry Boult, Peter Cheeseman, Michal Irani, Shree Nayar, Steve Seitz, Sundar Vedula, and everyone in the Face Group at CMU. The research described in this chapter was supported by US DOD Grant MDA-904-98-C-A915. A preliminary version of this chapter appeared in the IEEE Conference on Computer Vision and Pattern Recognition (Baker and Kanade, 2000b). More experimental results can be found in (Baker and Kanade, 1999).

References

Baker, S. and Kanade, T. (1999). Hallucinating faces. Technical Report CMU-RI-TR-99-32, The Robotics Institute, Carnegie Mellon University.

Baker, S. and Kanade, T. (2000.a). Hallucinating faces. In *Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition*, Grenoble, France.

Baker, S. and Kanade, T. (2000b). Limits on super-resolution and how to break them. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head, South Carolina.

Baker, S., Nayar, S., and Murase, H. (1998). Parametric feature detection. *International Journal of Computer Vision*, 27(1):27–50.

Barbe, D. (1980). *Charge-Coupled Devices.* Springer-Verlag.

Bascle, B., Blake, A., and Zisserman, A. (1996). Motion deblurring and super-resolution from an image sequence. In *Proceedings of the Fourth European Conference on Computer Vision*, pages 573–581, Cambridge, England.

Bergen, J. R., Anandan, P., Hanna, K. J., and Hingorani, R. (1992). Hierarchical model-based motion estimation. In *Proceedings of the Second European Conference on Computer Vision*, pages 237–252, Santa Margherita Liguere, Italy.

Born, M. and Wolf, E. (1965). *Principles of Optics.* Permagon Press.

Burt, P. (1980). Fast filter transforms for image processing. *Computer Graphics and Image Processing*, 16:20–51.

Burt, P. and Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540.

Cheeseman, P., Kanefsky, B., Kraft, R., Stutz, J., and Hanson, R. (1994). Super-resolved surface reconstruction from multiple images. Technical Report FIA-94-12, NASA Ames Research Center, Moffet Field, CA.

Chiang, M.-C. and Boult, T. (1997). Local blur estimation and super-resolution. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, pages 821–826, San Juan, Puerto Rico.

De Bonet, J. (1997). Multiresolution sampling procedure for analysis and synthesis of texture images. In *Computer Graphics Proceedings, Annual Conference Series, (SIGGRAPH '97)*, pages 361–368.

Dellaert, F., Thrun, S., and Thorpe, C. (1998). Jacobian images of super-resolved texture maps for model-based motion estimation and tracking. In *Proceedings of the Fourth Workshop on Applications of Computer Vision*, pages 2–7, Princeton, NJ.

Edwards, G., Taylor, C., and Cootes, T. (1998). Learning to identify and track faces in image sequences. In *Proceedings of the Third International Conference on Automatic Face and Gesture Recognition*, pages 260–265, Nara, Japan.

Elad, M. and Feuer, A. (1997). Restoration of single super-resolution image from several blurred, noisy and down-sampled measured images. *IEEE Transactions on Image Processing*, 6(12):1646–58.

Elad, M. and Feuer, A. (1999). Super-resolution reconstruction of image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):817–834.

Freeman, W. and Pasztor, E. (1999). Learning low-level vision. In *Proceedings of the Seventh International Conference on Computer Vision*, Corfu, Greece.

Hardie, R., Barnard, K., and Armstrong, E. (1997). Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6(12):1621–1633.

Horn, B. (1996). *Robot Vision*. McGraw Hill.

Irani, M. and Peleg, S. (1991). Improving resolution by image restoration. *Computer Vision, Graphics, and Image Processing*, 53:231–239.

Peleg, S., Keren, D., and Schweitzer, L. (1987). Improving image resolution using subpixel motion. *Pattern Recognition Letters*, pages 223–226.

Philips, P., Moon, H., Rauss, P., and Rizvi, S. (1997). The FERET evaluation methodology for face-recognition algorithms. In *CVPR '97*.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C*. Cambridge University Press, second edition.

Qi, H. and Snyder, Q. (2000). Conditioning analysis of missing data estimation for large sensor arrays. In *Proceedings of the 2000 IEEE Con-*

ference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina.

Riklin-Raviv, T. and Shashua, A. (1999). The Quotient image: Class based recognition and synthesis under varying illumination. In *Proceedings of the 1999 Conference on Computer Vision and Pattern Recognition*, pages 566–571, Fort Collins, CO.

Schultz, R. and Stevenson, R. (1994). A Bayseian approach to image expansion for improved definition. *IEEE Transactions on Image Processing*, 3(3):233–242.

Schultz, R. and Stevenson, R. (1996). Extraction of high resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5(6): 996–1011.

Shekarforoush, H. (1999). Conditioning bounds for multi-frame super-resolution algorithms. Technical Report CAR-TR-912, Computer Vision Laboratory, Center for Automation Research, University of Maryland.

Shekarforoush, H., Berthod, M., Zerubia, J., and Werman, M. (1996). Sub-pixel bayesian estimation of albedo and height. *International Journal of Computer Vision*, 19(3):289–300.

Smelyanskiy V., Cheeseman P., Maluf D. and Morris, R. (2000). Bayesian super-resolved surface reconstruction from images. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head, South Carolina.

Szeliski, R. and Golland, P. (1998). Stereo matching with transparency and matting. In *Sixth International Conference on Computer Vision (ICCV'98)*, pages 517–524, Bombay.

Index