High Quality Image Reconstruction by Optimal Capture and Accurate Registration

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (by Research)
in
Computer Science

by

Himanshu Kumar Arora
200607018
himanshu@research.iiit.ac.in

International Institute of Information Technology
Hyderabad, India
April, 2009

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "High Quality Image Reconstruction by Optimal Capture and Accurate Registration" by Himanshu Kumar Arora, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Advisor: Dr. Anoop M. Namboodiri

To all my teachers right from my childhood

Acknowledgments

A memorable journey which saw sizzling spells of summers, numbness of winters, visuals of autumns, passions of monsoons and rewards of springs is almost an inch away towards completion. Memorable because the lessons learned during this time is useful for whole life. Interactions with many different people on various topics contributed a lot to me or to my thesis.

First, I thank General Electric for considering me worthy enough to award a scholarship. My research assistantship for a over period of two years has come up from the scholarship.

A brilliant summer internship at General Electric in 2006 had a significant impact on several aspects. Sifting through the junk present all over the internet and discovering the right source of information for various purposes was the key learning. Various opportunities were provided during internship which were unique. For all of it I thank KS Shriram who was my advisor during the internship. I also thank Mitali More, Yogish Mallya, Srikanth Suryanarayanan and other team members of the lab.

I volunteered at several conferences. IJCAI, in 2007, is so far the best that I attended. I acknowledge the invited talk by Peter Stone. It was a general talk which had a key message that for good research to happen good research problems must be defined. It motivated me to come up with good research problems.

I also thank IIIT staff for various different reasons particularly Appaji, Jaya Prakash, Y Kishore, Somyajulu. Several things happened quickly because they knew that I'm a master student. I particularly thank R. S. Satyanarayana for maintaining CVIT, assisting our professors and keeping us updated on various fronts.

Labmates at CVIT helped a lot. Technical and non-technical discussions and analysis on wide range of topics helped directly or in-directly to my thesis. For technical discussions on topics related to my thesis, on which they are working on or any general topic I thank Anil, Avinash, Dileep, Jagmohan, Jyotirmoy, Paresh, Pawan Harish, Pradhee, Pramod Shankar, Ranjeeth, Shiben, Vibhav and Vishesh. I thank Shiben for cell-phone camera, Keerthi for light source and Avinash during various setups. Apart from these people I thank, which include non-lab mates as well, Anand, Avinash Kumar, Chhaya, Gopal, Hafez, Jinesh, Mihir, Naveen B, Naveen Tiwari, Neeba, Nirnimesh, Pavan Kumar, Pawan Harish, Pooja, Prachi, Pranav Vasishta, Pratyush, Sachin, Sanjeev, Santosh, Suhail, Suman Karthik, Sunil Mohan, Suryakant and Tarun.

I thank Prof. Jayanti, Prof. Narayanan and Prof. Jawahar for offering various relevant courses. My advisors have been very nice to me. I thank Prof. C.V. Jawahar for being very helpful. His continuous feedbacks on various issues were very important. He has been very tolerant on all my mistakes. My advisor Dr. Anoop M. Namboodiri provided ample freedom. This is perhaps the best thing happened during my masters. With encouragements from his side, I myself learned the art of coming up with new problem statements at master level. I came up with two different problems in my thesis. The first one did not go well. I came up with some other problem statements which were not related to my thesis. I also learned that seemingly impossible problems should not be given up right away. Sometime it may take good number of years before an extraordinary solution strikes in mind. I also thank him for listening passionately on various topics. Optimism is the best thing I found in him.

Abstract

Generating high-resolution and high quality images from degraded images has a variety of applications in space imaging, medical imaging, commercial videography and recognition algorithms. Recently, camera-equipped cell-phones are widely being used for day-to-day photography. Limitations of the image capturing hardware and environmental attenuations introduce degradations into the captured image. The quality of reconstructed image obtained from most of the algorithms are highly sensitive to accurate computation of different underlying parameters such as blur, noise, geometric deformation, etc. Variations in blur, illumination and noisy conditions together with occlusions further make the computation of these underlying parameters difficult.

One of the ways of generating a high-quality image is by fusing multiple images, which are displaced at sub-pixel levels. This method is also popularly known as multi-frame Super-resolution (SR). Most multi-image SR algorithms assume that the exact registration and blur parameters between the constituent frames are known. Traditional approaches for image registration are either sensitive to image degradations such as variations in blur, illumination and noise, or are limited in the class of image transformations that can be estimated. These conditions are frequently violated in real-world imaging, where specular surfaces, close light sources, small sensors and lenses create large variations in illumination, noise, and blur within the scene. Interestingly, these are the exact situations, where one would like to employ SR algorithms. We explore an alternate solution to the problem of robustness in the registration step of a SR algorithm. We present an accurate registration algorithm that uses the local phase information, which is robust to the above degradations. The registration step is formulated as optimization of the local phase alignment at various spatial frequencies. We derive the theoretical error rate of the estimates in presence of non-ideal band-pass behavior of the filter and show that the error converges to zero over iterations. We also show the invariance of local phase to a class of blur kernels. Experimental results on images taken under varying conditions are demonstrated.

Recently, Lin and Shum has shown an upper limit on multi-frame SR techniques. For practical purposes this limit is very small. Another class of SR algorithms formulate the high-quality image generation as an inference problem. High-resolution image is inferred from a set of learned training patches. This class of algorithm works well for natural structures but for many man-made structures this technique does not produce accurate results. We propose to capture the images at optimal zoom from the perspective of image super-resolution. The images captured at this zoom has sufficient amount of information so that it can be magnified further by using any SR algorithm which promotes step edges and certain features. This can have a variety of applications in consumer cameras, large-scale automated image mosaicing, robotics and improving the recognition accuracy of many computer vision algorithms. Existing efforts are limited to image a pre-determined object at the right zoom. In the proposed approach, we learn the patch structures at various down-sampling factors. To further enhance the output we impose the local context around the patch in a MRF framework. Several constraints are introduced to minimize the extent of zoom-in.

Projector-Camera systems are used for various applications in computer vision, immersive envi-

ronments, visual servoing, etc. Due to gaps between neighboring pixels on the projector's image plane and variations in scene depth, the image projected onto a scene shows pixelation and blurring artifacts. In certain vision and graphics applications, it is required that high quality composition of the scene and the projected image is captured, excluding the artifacts, while retaining the scene characteristics. The localization and restoration of projected pixels is difficult due to various factors such as spatially varying blur, background texture, noise, shapes of scene objects, and color transformations of projector and camera. We extend the usage of local phase, computed using the Gabor filter, to isolate each of the projected pixels distinctively. The local phase is also invariant to a class of blur kernels. For restoration of the captured images, we reproject the projected pixels such that these artifacts are absent. To improve the quality further we propose a mechanism to virtualize a high-resolution projector.

Contents

List of Figures

List of Tables

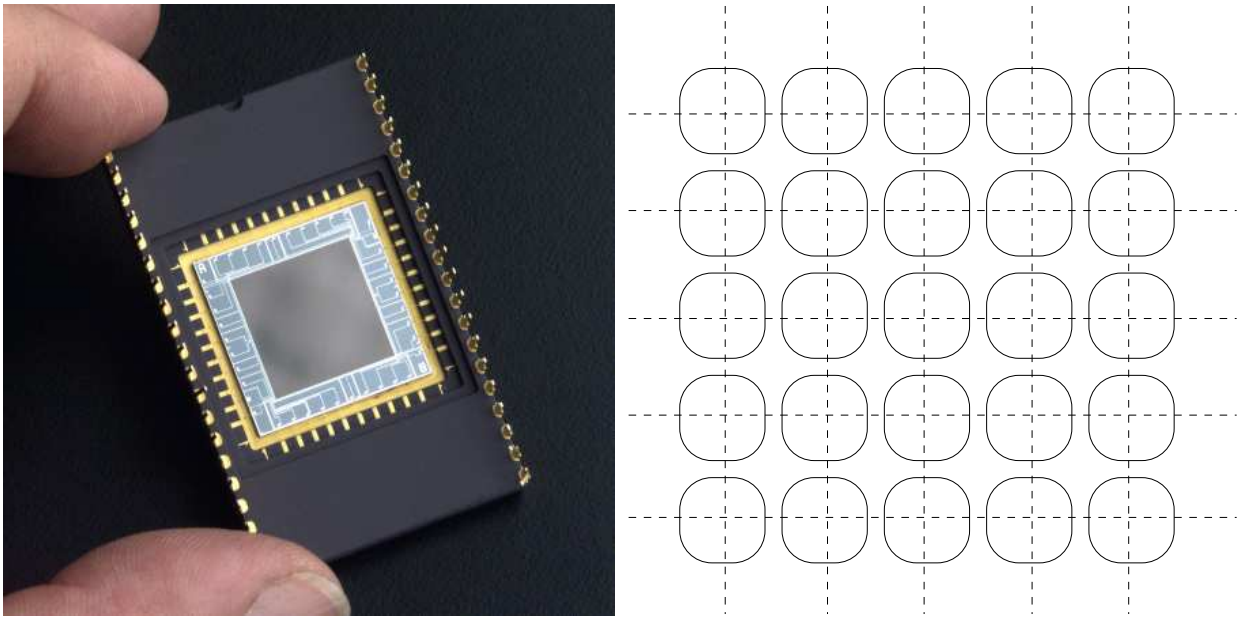
Chapter 1

Introduction

The history photography or the act of capturing scene irradiance field permanently on a surface goes back to 1826 [1] when Joseph Nicèphore Nièpce captured the first photograph. Quality was low and the technology was very poor. It took eight hours of exposure to capture the image. Over decades the technology improved. The camera became compact for practical usage, the exposure time was low and the quality of imaging plates improved. The real breakthrough in imaging came with the use of photographic films in 1885. However, it was limited to a single exposure per loading. With the invention of roll film cameras multiple exposures per loading became possible. Several enhancements were added in cameras later which includes single and twin reflex lenses. Polaroid cameras, which appeared between 1947 and 1983, was marked with the faster processing of the films. In 1978, first auto-focus camera was introduced. For the first time in 1951 an image was converted into digital format for saving the information on a digital tape. But it was in 1975 when the first image was captured through a digital camera. The camera was heavy and took 23 seconds to capture the image. Subsequently many improvements were made in digital cameras and sensors. Nowadays very high quality cameras are within the reach of consumers. Digital cameras have several advantages over their analog counterparts. A digital camera does not require a slow and expensive chemical processes and the captured image can be enhanced in later stages. It can be transmitted easily and reliably over data channels.

To sample the irradiance field into digital signals, a CCD or CMOS chip is used. The imaging chip or sensor is an array of rectangular pixels of non-zero area arranged on a rectangular grid. Sampling on these chips is not equivalent to point sampling. The irradiance field falling on each of the rectangular pixel is averaged out. Figure 1.1 (a) shows a typical CCD chip. Figure 1.1 (b) shows the geometry and arrangement of pixels on the chip. Camera manufactures refer to resolution of a camera as the number of pixels it has on a chip. But in computer vision literature, camera resolution refers to how finely and compactly irradiance field can be sampled on a chip. Mathematically, it translates to the number of pixels per unit area. To capture finer scene details high-resolution cameras are used. One of the fundamental limitation of imaging device is the number of pixels a digital camera can have for sampling. The amount of noise increases significantly with smaller pixel size. The captured images are highly degraded. To circumvent this problem, irradiance field is made to converge through a complex lens assembly on a larger CCD chip in high-end cameras, so that a high quality image is captured even with high shutter speed. Cell-phone cameras, on the other hand, have compact and small sensors. The cost of compact and small sensors are low. This might seem counter-intuitive, but one should note that the cost of a chip depends on its overall size. Compactness is irrelevant once a chip goes into manufacturing stage.

Modern digital cameras aims at capturing high quality and high resolution images using low-cost



(a) (b)

Figure 1.1: (a) Image of a typical CCD chip. Courtesy Wikipedia. CCD chip is an array of Metal-Oxide-Semiconductor capacitors (MOS capacitors). Each of the capacitors represent a pixel. (b) shows the typical geometry and arrangement of these pixels.

and highly compact sensors. Recently, a lot of research effort has gone into capturing even more scene information like depth in a single shot on the imaging sensors [2]. Various functionalities like zoom, focus, high speed photography, color imaging etc. come at the cost of various imaging artifacts. Rectifying or avoiding them at the hardware level is desirable but current hardware implementations are either constrained with physical limitations or are inappropriate for this task. As discussed before, another limitation of digital cameras is the number of pixels a chip can have for capturing finer scene details.

With all the advancements, capturing high-quality images from digital sensors is still a challenging problem. The captured degraded images are processed later to recover high-quality images. High-quality image reconstruction aims at rectifying the degraded images to the level of irradiance field of the scene as perceived by human eyes. High-quality image reconstruction measures could be objective or subjective. In objective reconstruction, exact modeling and restoration is desired, whereas in subjective reconstruction, only limited artifacts are removed so that either the image is perceived equivalent to the actual high-quality image or the image has sufficient desirable information.

Reconstruction is not a trivial problem. Various underlying factors and highly accurate measurements of each of them is necessary for any reconstruction to happen. Captured images suffer from various artifacts such as chromatic aberration, noise, blurring, aliasing etc. Various electronic limitation and geometric artifacts also contributes towards degradation of images. General forward degradation process is mathematically modeled as,

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x} + \eta_k, \tag{1.1}$$

where $\mathbf{y}_k$ is one of the low quality images captured by the camera, $\mathbf{H}_k$ models various degradation process including sampling of the irradiance field, η_k is the noise and $\mathbf{x}$ is the high quality irradiance

field falling on the sensor plate. The goal is to recover $\mathbf{x}$ precisely. However, due to storage and various other mathematical limitations we recover an image that has more detail than any of the captured image has. Multiple images of the scene are captured to account for missing data. The problem of estimating accurate value of $\mathbf{H}_k$ is not trivial as the estimation of various sub-parameters are difficult in presence of varying degrees of other degradations. After estimating various sub-parameters the inverse should be computed robustly even-if some of the data is missing.

Apart from perceptual reasons, high quality images are useful at various places in Computer Vision. Video surveillance applications require the identification of subjects and objects, which can not be seen properly in any of the captured frames. In satellite imaging, clarity and details are very important. Clear identification of targets in military applications is an important task. High-resolution image reconstruction has many commercial applications too. Restoration of old videos is a challenging task. In medical image analysis, high quality images are required for detailed analysis without over-exposing the patient to radiations. Recognition algorithms fail to perform in the absence of good quality images, and image analysis tasks such as image based rendering require very high quality images.

In this thesis, we look into various factors that affect high quality image reconstruction. We address three problems towards high quality image generation. One of the problems towards high-quality image generation is to merge data captured at different instances, with different imaging parameters. Image super-resolution is one of the way for generating high quality images. The process simulates high-quality and high-resolution camera from multiple images captured using a low quality and low-resolution camera. We address the problem of image registration for image super-resolution in presence of noise, non-uniform illumination and blur. We also look into various perspectives of zooming towards capturing sufficient information of the scene. Sufficiency is defined from the perspective of high-resolution image reconstruction. The third problem is to enhance the captured images in projector-camera domain. The remainder of this chapter provides an overview of high quality image capture, brief background on various artifacts and related work. Various research challenges in the field are listed.

1.1 High Quality Images

Quality is a highly subjective term. In imaging literature, there are two meanings of image quality. An image is of high quality objectively if the imaging sensors capture the exact image as seen by placing human eye in place of sensors. It should not have other imaging artifacts like defocus irrespective of depth, noise, chromatic aberration, vignetting etc. On the other hand, captured images would be subjectively of high quality if after capturing they can be processed or interpreted the same way as human mind does. Certain artifacts could be present in the image but their presence is indistinguishable to human being, or their presence does not affect any recognition task.

Objective Enhancements : Majority of the research work is towards restoring image artifacts introduced by the camera. Accurate modeling of these artifacts is a very important task. Denoising, deblurring, super-resolution, high-dynamic range imaging, and correction of vignetting, chromatic aberration, etc. are common techniques used to enhance the images. Equation 1.1 describes a forward process, where the original image undergoes degradation before being recorded in practice, multiple images are captured to account for missing data. Usually the images are captured from different view points or with varying imaging parameters to solve the equation.

Subjective Enhancements : There are many subjective parameters that define high quality images. Criteria is highly dependent on kind of applications. Janssen [8], in his PhD thesis defines

Figure 1.2: Several examples of high-quality image reconstruction; (a) single-image deblurring, (i) blurred image, (ii) restored image. Courtesy [3]; (b) super-resolution; (i) low-resolution image, (ii) high-resolution image. characters in the center are clearly visible in the super-resolved image; (c) selecting the right zoom of camera for meaningful scene information (see chapter4 for details), (i) original image, (ii) image captured at right zoom; (d) image denoising, (i) noisy image, (ii) restored image, (iii) original image. Courtesy [4]; (e) color constancy, (i) captured image, (ii) correction using [5], (iii) ideal correction. Courtesy [5]; (f) image inpainting, (i) image with text, (ii) after image inpainting. Courtesy [6]; (g) single image dehazing single. (i) captured image , (ii) restored image. Courtesy [7].

image quality on a four-point philosophy. a) Image is regarded as a carrier of visual information; b) Visual cognitive process is regarded as information processing; c) Visual-cognitive processing is considered as an essential stage in human interaction with environment instead of an isolated process; and d) Quality is not described in terms of visibility of distortions. Instead it is defined as the suitability of the image as an input to the vision stage of the interaction process.

There has been a significant amount of work on assessing the quality of images or videos in manner consistent to human perception. Recent work by Sheikh *et al.* [9] and the references therein provides a brief overview on assessing image quality. Various quality assessment techniques are divided into two major groups *viz.* a) based on Human Visual Systems (HVS); and b) based on arbitrary signal fidelity criteria. Though there have been significant advancements in assessing visual image quality, most of them are not used to restore the degraded images.

Various subjective image enhancement techniques include object removal or inpainting, edge

enhancements, contrast enhancements, etc. In single image high quality image reconstruction, various subjective measures are used to restore the images. Recently, certain artifacts are used to convey the scene information in the right manner (e.g. image refocusing). Understanding subjective image enhancements not only help to reconstruct high quality images but also in degrading the high quality images towards image compression and transmission. Figure 1.2 shows certain examples of high quality image reconstruction. Following section provides details on related work on various high-quality image reconstruction methods.

1.2 Methods of Obtaining High Quality Images

In this section, we provide an overview of different imaging degradations and methods towards reconstructing high-quality images. We provide an overview of different categories of solutions. Sufficient recent or key work is cited corresponding to each of them. In addition to the methods mentioned in table 1.1 other high quality enhancements include edge sharpening, contrast enhancement, geometrical distortions, veiling glare removal etc.

(a) Denoising: Digital sensors and image compression are the main sources of noise in images. A large amount of literature exist on modeling and removing noise from images. Most of the algorithms assume certain characteristics to be important like edges, and the image is denoised while preserving them.	Comprehensive literature review can be found in [4, 10]. Various denoising techniques are broadly divided into two categories: • *Spatial Domain Methods:* Spatial filtering for image denoising works only for additive noise. Median filters, mean filters, max and min filter and various spatially adaptive versions [11] are commonly used. • *Transform Domain Methods:* Transform domain methods have computational advantages. Various frequency bands, which may contain noise, can be processed specifically. Wavelet-based procedures [12, 13, 14] have received considerable attention because of the localization achieved in both spatial and transformed domain. Bandreject filters, bandpass filters, notch filters, Wiener deconvolution, Gaussian low pass filters [11], bilateral filter [15] and their variants are very popular. Recently, Yuan *et al.* [16] proposed a method to deblur and denoise the blurred/noisy image pair simultaneously.

(b) Deblurring: Image is blurred when irradiance field corresponding to a single pixel is smeared to more than one neighboring CCD pixels. Among various reasons most common include defocus due to varying scene depth, long-aperture time and camera motion. Usually multiple images are captured to model the blur kernel robustly.	Characterizing the blur type and modeling the blur kernel are very important step in image deblurring. Deblurring algorithms are different for different categories of blur. • *Lens Blur:* Blur kernel is symmetric except in depth varying scenes. Commonly used deblurring techniques include Wiener filtering, constrained least squares filtering [11], Lucy-Richardson [17]. Overview on blind image deconvolution methods is provided in [18]. In [19], a method is proposed to capture omni-focus image from multiple captured images. • *Motion Blur:* Single image methods [20, 21], multiple image method [22]. • *Camera Hand-Shake Blur:* Single image method [23], multiple image method [16]. To enhance the image deblurring task specialized cameras are also proposed [19, 24, 25, 26].
(c) Super-Resolution (SR): Super-resolution is the process of simulating a high-resolution, high-quality camera from blurred, noisy images captured using a low-resolution camera. SR algorithms are divided into two categories *viz.* multi-frame and learning based single-image super-resolution.	• *Multi-frame SR:* Multiple images of a scene are captured at sub-pixel displacements. These images are registered and high-resolution information is computed [27, 28, 29, 30, 31, 32, 33, 34]. [35] and [36] provide a comprehensive literature survey. • *Learning based single-frame SR:* In learning-based SR algorithms high quality image computation is modeled as an inference problem in a Markov Random Field framework. Key works include [37, 38, 39, 40, 41].
(d) High Dynamic Range Imaging: Typical image sensors capture smaller range of intensity levels (256 levels in most cameras). High dynamic range imaging aims at capturing a wide range of intensity levels or detailed brightness variations in a scene.	• *HDR from multiple images:* Many images are captured at different exposures and a single high dynamic range image is calculated from them [42, 43, 44, 45]. • *Hardware Enhancements:* To facilitate imaging for dynamic scenes several hardware enhancements such as [46, 47] are proposed. Comprehensive references can be found in tutorial by Goesele *et al.* [48].

(e) Vignetting: Because of optical properties of multiple element lenses which block lights on peripherals of rear elements and $\cos^4$ law, there is gradual fall-off in image brightness towards the periphery, referred to as vignetting.	• *Single image method:* [49] • *Multiple image methods:* [50, 51]
(f) Chromatic Aberration: Chromatic aberration is caused by the lenses having different refractive index for different wavelengths of light. As a result, different color bands are defocussed to different degrees.	This problem is prevalent with low quality lenses. Several lens designs exists which reduce this artifacts. Boult and Wolberg [52], Kang [53] and the references therein provide various removal mechanisms.
(g) Environmental Attenuation: Various environmental attenuations such as fog, mist, rain, smog, hail etc. affect the quality of image and performance of various computer vision algorithms. Often multiple images are used to remove these artifacts. Single image methods have also been proposed.	Removal of artifacts like snow, dense fog, hail storm, smog and heavy rain from the captured images or videos is still very challenging. Various attempts to dehaze images include capturing multiple images over time [54], single image methods [7] and the references therein. Garg *et al.* [55] proposed a solution for removal of rain from videos. The use of various optical accessories like polarization filters are used to capture images without haze [56] but it has its own limitations.
(h) Image Inpainting and Completion: Image inpainting is a technique for restoring damaged paintings and photographs, filling in the holes to the removal or replacement of complete selected objects. Usually the process is semi-user-assisted.	First proposed by Bertalmio *et al.* [6], image inpainting has come a long way since then. Most of the approaches include selection of desired regions by the user. Those regions are replaced with the texture information in the neighborhood or from multiple images. Major papers on this topic include [6, 57, 58, 59, 60, 61, 62].

(i) Color Rectification and Enhancements: Factors involving response of camera sensors, chromatic aberration, presence of unusual light sources in the scene that make the estimation of true reflectance of the object difficult. In a different scenario, one might want old monochrome images to be restored.	• *Colorization of Monochrome Images:* [63, 64]. • *Color Correction:* Also known as white balancing or color balance. The goal is to rectify the color so as to the match the sensors in human eye [65]. • *Color Constancy:* Various object recognition algorithms require the true reflectance of the object to be measured. Color constancy ensures that the camera captures the right reflectance of the object even in presence of factors extrinsic to the object. Classical algorithms include white patch algorithm [66] according to which the maximum response in RGB channel is caused by a white channel and grey world algorithm [67], which assumes that the average reflectance of a scene is achromatic. [5] defines a universal approach using image statistics. Barnard *et al.* [68] provide an overview of key algorithms.
(j) Estimating Camera Response Function(CRF): CRF maps the image irradiance field falling on the sensors to the measured pixel intensities.	• *Single-image methods* : [69, 70]. • *Multiple-image methods* : [43, 71]. Multiple image methods offer accuracy and robustness advantages.
(k) Using Artifacts to Convey Right Information: Various image artifacts like focus, noise etc. are used to highlight the desired region. After the image is captured, the artifacts in unwanted regions are removed and reintroduced at various regions to highlight a subject in images.	*Image Re-focusing:* Image is re-focused on to a different depth to highlight the subject in attention. Usually the depth field of the scene is captured or multiple images are captured at different focus settings so that the image can be refocused at different locations off-line. Ng *et al.* [2] presented a sensor design that captured 4D light field on a sensor in a single exposure but the captured image has smaller number of pixels on the sensor. Noguer *et al.* [72] presented an idea to capture the depth using the defocus of sparse set of dots projected on the screen.

Table 1.1: High quality image reconstruction methods: An overview.

1.3 Challenges in High Quality Image Reconstruction

We now outline the current research challenges in high quality image reconstruction. Some of them are already well addressed in research community. However, accurate, efficient and highly robust solutions are still desirable. High-quality image reconstruction is usually modeled as an inverse

problem. Inadequate observations and the presence of various other artifacts requires various relaxations to be incorporated making the solution space very large. Highly accurate computation of various underlying factors become extremely important in this case. Major challenges in high-quality image reconstruction revolve around this particular factor. Following are some of the important problems that are not properly addressed in the literature or require further research.

- **Registration of multiple degraded images in presence of various other artifacts** : Image registration is a process of geometrically aligning two or more images obtained from different views. In high quality image reconstruction process, various images are usually captured to calculate the inverse robustly and reduce the size of solution space. Inaccurate alignment of various images can adversely affect the high-resolution image computation. There are various image registration algorithms [73], which can compute the registration parameters with error limited within one-fifth of a pixel. But in the presence of degradations like noise, blur, non-uniform illuminations, environmental attenuations and various other camera artifacts current registration algorithms are hardly accurate and robust. These artifacts changes the effective intensity of images and degrades the key primitive features. One of the way to circumvent this problem is to compute registration and degradation parameters in cyclic fashion. This process is very slow and sometimes the process might diverge severely. Registration parameters computed in transform domain are more robust to these artifacts but the class of registration parameters that can be solved are very limited.

- **Acquiring sufficient information for meaningful image reconstruction** : There are two sub problems. Firstly, how much restoration is meaningful, and secondly, how much minimum information should be acquired to reconstruct high quality images meaningfully. Some typical scenarios include restoring images suitably for recognition tasks e.g. text from document images or restoring the images suitable for human perception. All algorithms reconstruct the high-quality image either from multiple images or from a single image using learning based methods. As image reconstruction is an ill-posed inverse problem. Exponentially higher number of images are required for complete restoration. If the acquired images have sufficient amount of information or if the number of images to be captured is known in advance, the reconstruction task can be done efficiently. Lin *et al.* [74] addressed this issue from the perspective of multi-frame super-resolution i.e., how many images are sufficient to super-resolve an image at a given magnification factor. However, magnification selection for meaningful super-resolution is still an unaddressed problem.

- **Simultaneously removing multiple artifacts** : Restoration problem is often simplified in literature. It is rare to see papers where multiple degradations are modeled and multiple artifacts are successfully removed. Problems like removing chromatic aberration or computing high dynamic range image in presence of blurs like motion blur, lens blur, hand-shake blur etc. are never addressed before. Problems are difficult and would require totally different numerical solution or in-camera enhancements. Solutions to these problems are essential for high quality imaging using low cost cell-phone cameras.

- **Computational efficiency** : Computational efficiency is essential for in-camera processing and for different real-time computer vision and robotics applications. Current reconstruction algorithms are very slow. It is difficult to sacrifice accurate computation of various underlying factors for computational speed-up. Parallel graphics hardwares like CUDA tremendously increase the computational speed but at very high cost. For in-camera computational efficiency

mathematical and camera hardware enhancements should be addressed. Scene specific processing can also increase the efficiency. Transform domain techniques are usually faster but complicated spatially varying degradations can not be processed in this framework.

- **Robust inverse strategies** : High quality image reconstruction is modeled as an inverse problem. To compute the inverse, parameters of forward process are computed from multiple images. The solution space remains large because of corrupted and insufficient observations. To stabilize the solution, various numerical constraints and regularization are introduced. Improper constraints introduce various other artifacts like ringing, degradations of details, over-smoothening of edges etc. As the complexity and multiplicity of degradations increase existing inverse computations are rarely helpful. More problem specific or scene specific regularization and efficient adaptive models for restoration need to be developed.

- **Single-image reconstruction** : High quality reconstruction from a single image has always been a challenging problem with tremendous applications. Usually prior information [37, 20, 5] or scene specific cues [66, 67] are used to rectify the images. Most of the existing single-image reconstruction algorithms restore the images perceptually. Efficient modeling and use of visual cues, robust modeling of the degradation process, hardware level enhancements would improve accuracies in single image reconstructions.

- **Limits on restoration** : Lin and Shum [74] and Lin *et al.* [75] established the theoretical and practical limits on reconstruction based and learning based super-resolution algorithms respectively for limited scenarios. Such research works save computational time, effort and time of other researchers who are finding new ways to improve the performance. However, similar limits are missing for other high-quality reconstruction algorithms.

- **Perceptual image reconstruction** : Most of the restoration algorithms compare the performance with images captured using a high quality camera. In some other cases, degraded images are simulated and after restoration the image is compared with the actual image. Sometimes, depending on the application, we can restore an image equivalent to perceptual level which is different from the actual high quality image. Advantages could include low computational effort and generalizability. Quantitative methods are straight-forward to model but difficult to solve. The case is opposite while restoring the image perceptually. Abdou and Dusaussoy [76] provide a survey on image quality measurements both qualitatively and quantitatively. Only a few of the degradations mentioned in table 1.1 has been evaluated subjectively. Earlier related research was mostly from the perspective of image compression. Further study and research in this direction would reduce considerable amount of effort and computational time.

1.4 Motivation, Problem Statement and Contributions

In this thesis, we explore various aspects of high quality image reconstruction. Three different problems on high quality image generation has been addressed. We specifically address the problem of accurate registration of multiple images for image super-resolution. We have also addressed the problem how much information content is sufficient enough for any general scene so that any further resolution enhancement can be obtained using any off the shelf super-resolution algorithm. Problem of capturing high quality image in projector-camera system domain is also addressed.

1.4.1 Accurate Registration of Images using Local Phase Information for Super-Resolution

Many existing super-resolution reconstruction algorithms assumes the availability of accurate blur and registration parameters. The primary factor that controls the quality of the super-resolved image is the accuracy of registration of the low resolution frames. Most of the existing registration algorithms perform well in presence of uniform illumination across frames as well as limited and uniform blur and noise. However, these conditions are frequently violated in real-world imaging, where specular reflections and strobe lights create large variations in illumination of the scene. Moreover, non-uniform blur often results from depth variations in the scene, while high noise levels are seen in images generated from compact sensors in mobile devices. Traditional approaches for image registration are either sensitive to image degradations such as variations in blur, illumination and noise, or are limited in the class of image transformations that can be estimated. In this thesis, we propose the use and address suitability of local phase for accurate image registration in presence of noise, non-uniform blur and illumination.

Contributions: As the primary factor which controls the quality of the super-resolved image is the accuracy of registration of the low resolution frames. So, we explore at an alternate solution to the problem of robustness in the registration step of a SR algorithm. We formulate the registration as optimization of the local phase alignment at various spatial frequencies and directions. The local phase in an image has been used for problems such as estimation of stereo disparity [77], and optical flow field estimation [78]. We extend its scope to estimate accurate registration parameters and use it for computing super-resolved images. In this thesis, we: 1) propose a registration framework using local-phase, which is known to be robust to noise and illumination parameters, 2) derive the theoretical error rate of the approach introduced by limitations of Finite Impulse Response (FIR) filters and show that the algorithm converges to the actual registration parameters, 3) show that the algorithm is not sensitive to a large class of blur kernel functions; and 4) present experimental results of SR reconstruction, that demonstrates the advantages of this approach as compared to other popular techniques.

1.4.2 Selecting the Right Zoom of Camera from the Perspective of Super-Resolution:

Various super-resolution algorithms are commonly divided into two categories, *viz.* multi-frame super-resolution [30] and learning based super-resolution [37]. Lin and Shum [74] showed that the theoretical limit on magnification for multi-frame super-resolution is 5.7, and in practical scenarios this limit is only 2.5. For higher magnification factors, the number of images required increases exponentially, making the computational cost beyond practical limits for most applications. Multi-frame SR also requires accurate registration and blur parameters, which are very difficult to obtain in many scenarios. These drawbacks limit the applicability of multi-frame SR, and it is used primarily for revealing the exact underlying details at a limited magnification. In contrast, learning based single image SR, in theory, can achieve magnification factors up to 10, as shown by Lin *et al.* [75]. The HR image generation is formulated as an inference problem. Correspondences between LR and HR patches are stored during the learning phase, and the HR image is inferred in a MRF based framework with contextual constraints. This category of algorithms perform very well for natural objects, where the perceptual quality is more important than accurate reconstruction of reality. They also work well if the training set is optimized for specific object/scene classes, such as faces [39]. However, the performance drops significantly on man-made structures. In this thesis, we address the solution to this problem in an alternate way.

Contributions: In this thesis, we propose a new problem of high-resolution generation by capturing sufficient information at the image capturing stage itself. The image is decomposed into patches and zoom level prediction is modeled as an inference problem in a MAP-MRF framework. We use Bayesian belief propagation rules to solve the network. As the optimization function contains numerous local minima, a robust technique is proposed to initialize the solution. Various practical constraints are proposed to minimize the extent of zoom-in. The results are validated on synthetic data and experiments are performed on real scenarios to show the robustness of the proposed approach.

1.4.3 Capturing Projected Image Excluding Projector Artifacts

Projector-camera systems are extensively used in computer vision, HCI, immersive environments, and improving projection quality and versatility. Image projected on a scene suffers from pixelation artifacts due to gaps between the neighboring pixels on the projector's image plane, defocus artifacts due to varying scene depth and color transformation artifacts. If the camera is sufficiently close to the scene the pixelation artifacts are clearly visible. However, the pixelation artifact is also useful in certain applications. In this thesis, we address two sub-problems in this domain,

- The first problem, we address is to accurately localize each of the projected pixels. Detection of the projected pixels in the captured image can facilitate applications such as recomposition of a projected image and a scene, which is useful in the post production stage. Procams are also useful for capturing surface properties. Considering the pixelation and blurring artifacts improves the accuracy of such estimations. The relative spatial configuration of the localized pixels also help in computing a dense shape for dynamic scenes.

- The second problem, we address is the restoration of the captured image having pixelation and defocus artifacts. Public capturing of images of various projector scene composition such as presentation slides, immersive environments [79] etc. requires the restoration. Projector-scene composition is useful in movies for special effects. Images are rendered on real objects and the video is captured [80].

Contributions: We identify the problem and propose solution for localizing projected pixels accurately and enhancing captured images. We first analyze the structure of the projected pixels on the textured scene and propose a systematic approach to localize the projected pixels and remove the projector artifacts in the captured image. As our algorithm requires only one image, our system can work for dynamic scenes as well. No camera-projector calibration or co-axial camera-projector system is required. Specifically we proposed: 1) an image re-formation model that describes the relationship between the display image, the projected scene, and the captured image with pixelation and defocus artifacts; 2) a robust algorithm for identification of the projected pixels seen in the captured image; 3) a method to remove the pixelation and blurring artifacts of the projector in the captured image; and 4) a mechanism to improve the quality of the captured image further by virtualizing a high-resolution projector, so that the captured image sees a larger number of projected pixels.

1.5 Organization of the Thesis

In Chapter 2, we present brief tutorials on the preliminary concepts *viz.*Frequency domain, local phase, Markov Random Field (MRF) and Super-resolution. Understanding of these concepts are important to understand the other chapters in this thesis. In Chapter 3, we discuss the problem

of accurate image registration using local phase information for super-resolution. The algorithm is robust in presence of noise, non-uniform blur and illumination. Theoretical and practical analysis on the robustness of local phase is provided . In Chapter 4, a learning based algorithm to select the right zoom of the camera from the perspective of super-resolution is proposed and analyzed. The problem is broken into a patch based network and analyzed in MRF framework. In Chapter 5, we discuss a new problem of capturing projected image excluding projector artifacts. Also, we discuss how to use local phase towards locating projected pixels accurately in presence of blur and intensity variations. In Chapter 6, we conclude and summarize the thesis. We also provide an overview of future problems.

Chapter 2

Theoretical Background

This chapter aims at providing comprehensive understanding of various existing techniques, theoretical concepts and frameworks that are used in this thesis. We provide a quick overview of Fourier transformation in section 2.1. In section 2.2, basic and advanced concepts of Gabor filter and local phase are discussed. Local phase information is used in chapter 3 to register images accurately. In chapter 5, local phase is used to separate out high pixels. In section 2.3, we provide a brief overview of Markov Random Field in the context of low-level vision. MRF framework is used in chapter 4 for optimal zoom imaging. In section 2.4, we discuss image super-resolution and the basic imaging model. Both multi-frame and single image learning based techniques are described. Key references for further understanding of the topics are provided in each of the section.

2.1 Frequency Domain

A signal is any quantity that is measurable over time or space. Signal processing is the analysis, interpretation or manipulation of signals. Signals can be processed in *time domain* or in *frequency domain*. Both of them carry the same information in different forms. Time domain processing is intuitive. The signal is analyzed as we perceive it, is interpreted as we understand it and manipulated as we conceive it. Frequency domain graph shows how much of a signal lies within each frequency band. A signal is analyzed based on the frequencies it contains. For several tasks, frequency domain processing has advantages over spatial domain processing. Frequency domain algorithms are computationally faster and provide good processing control over various characteristics of an image e.g. edge only processing. Frequency domain representation is also useful for efficient compression of images. High frequencies in images represent edges whereas low frequency components represent smooth regions. In image processing, a signal is a function of spatial coordinates instead of time, so Frequency is usually referred as *spatial frequency*.

Mathematical models that transform a signal from time-domain to frequency domain include Fourier transform, Discrete cosine transform, Mellin transform, Hadamard transform, Hilbert transform, and Laplace transform. Fourier transform is one of the widely used methods to convert and analyze a signal in frequency domain.

2.1.1 Fourier Analysis

Fourier transform [81, 11] convert a spatial domain signal into a summation of a series of sine and cosine terms in increasing frequencies, and back to spatial domain. Let $\psi(x, y)$ be a continuous

signal. The Fourier transform, $\Psi(u,v)$ of this signal is given by,

$$\Psi(u,v) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \psi(x,y)e^{-j2\pi(ux+vy)}dxdy, \tag{2.1}$$

where $j = \sqrt{-1}$. The original signal $\psi(x)$ is obtained by means of inverse Fourier transform as,

$$\psi(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \Psi(u,v)e^{j2\pi(ux+vy)}dudv. \tag{2.2}$$

These two equations comprise the Fourier transform pair. Images are discrete functions defined over a finite range. Fourier transform of a discrete function of two variables $\psi(x,y), x = 0,1,...,M-1, y = 0,1,...,N-1$, is given by the equation,

$$\Psi(u,v) = \frac{1}{MN}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} \psi(x,y)e^{-j2\pi(ux/M+vy/N)}, \tag{2.3}$$

for $u = 0,1,...,M-1, v = 0,1,...,N-1$. Similarly, we obtain the original image back using the inverse discrete Fourier transform as,

$$\psi(x,y) = \sum_{u=0}^{M-1}\sum_{v=0}^{N-1} \Psi(u,v)e^{j2\pi(ux/M+vy/N)} \tag{2.4}$$

for $x = 0,1,...,M-1, y = 0,1,...,N-1$. The components of Fourier transform are complex quantities. Let $\mathcal{R}(u,v)$ and $\mathcal{I}(u,v)$ denote the real and imaginary part of $\Psi(u,v)$ respectively. Then,

$$|\Psi(u,v)| = \sqrt{\mathcal{R}^2(u,v) + \mathcal{I}^2(u,v)}, \tag{2.5}$$

is the magnitude or spectrum of the Fourier transform, and

$$\phi(u,v) = \tan^{-1}\left[\frac{\mathcal{I}(u,v)}{\mathcal{R}(u,v)}\right], \tag{2.6}$$

is the phase angle or phase spectrum of the transform. Power spectrum is defined as the square of the Fourier spectrum. Both Fourier magnitude and Fourier phase have different roles in a signal.

Importance of Fourier Magnitude

The magnitude of the Fourier transform represents the contrast of the corresponding sinusoid in spatial domain i.e. the difference in intensity values in the darkest and the brightest peaks at that frequency. Phase shifted sinusoids at different frequencies are scaled by Fourier magnitude values and combined to construct the original time-domain signal. Roughly speaking Fourier magnitude information assigns abundance factors to each of the sinusoidal signal. Higher magnitude information implies more contribution of the corresponding frequency in a signal.

Fig. 2.1 shows the phase swapping experiment highlighting the importance of Fourier magnitude. In phase swapping experiment, the Fourier magnitude of one image is combined with the Fourier phase of the other image and vice-versa and the image is transformed back into spatial domain. Image ψ_2 is the aliased version of image ψ_1 generated by down-sampling the original image by a factor 4 and then up-sampling it. After swapping the phase information between images we can see that high quality image is preserved corresponding to the Fourier magnitude of the image ψ_1. Magnitude information basically suppressed the unwanted Fourier phase information by assigning low values at these locations. Sinusoids of other frequencies are no longer seen.

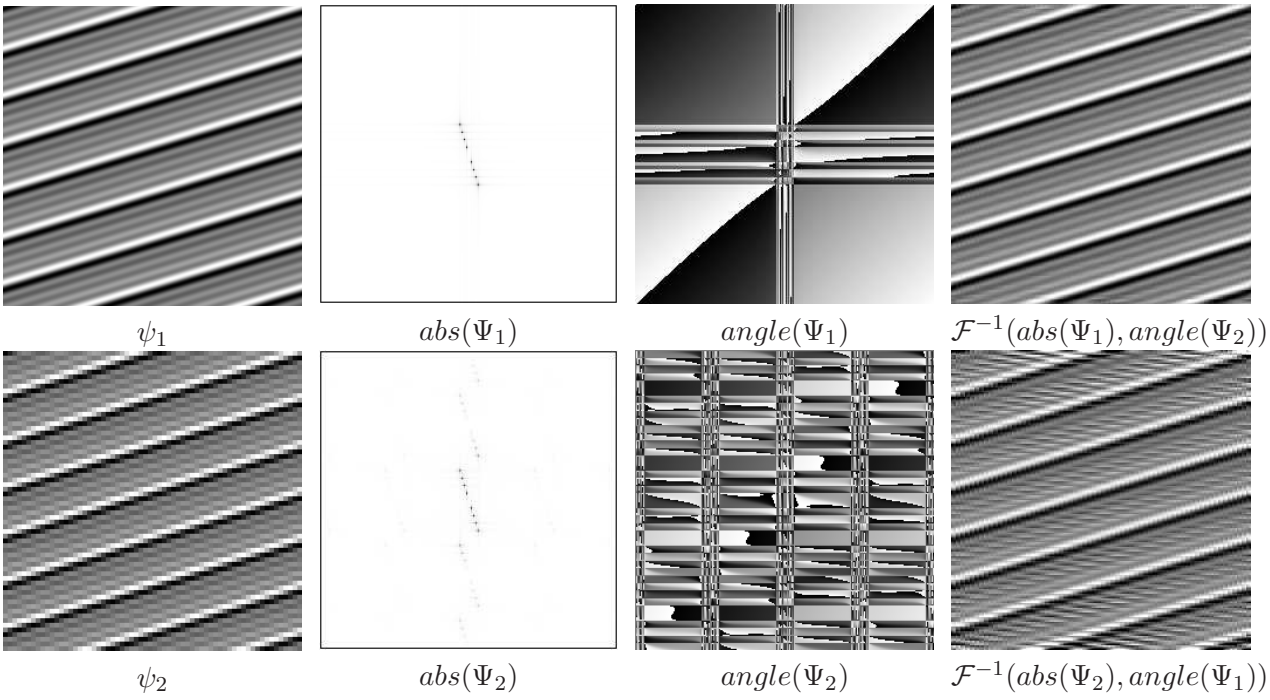


Figure 2.1: Phase swapping experiment showing the importance of Fourier magnitude; ψ_1 and ψ_2 are two images where ψ_2 is the aliased version of ψ_1. Ψ_i denote the Fourier transform of ψ_i. $abs(\ )$ and $angle(\ )$ denote magnitude and phase information respectively of a signal. The intensities in magnitude profile are inverted and center pixel value is set to 1 for better visualization.

Importance of Fourier Phase

Fourier phase has a very significant role in signals [82]. Phase value at a given frequency represents how much the corresponding sinusoid is shifted from the origin (also called as initial phase). A translation in image position has no effect on magnitude of the Fourier transform whereas the phase is shifted proportionally. Phase information preserves much of the correlation between signals. Under the assumption that a signal is of finite time duration, it is also possible to reconstruct the whole signal up to a scale factor using only the phase information. Fourier magnitude is affected by change in contrast whereas phase information remains independent of non-uniform illumination. Measurements based on Fourier phase is also known to be robust to band-limited noise.

Fig. 2.2 shows the phase swapping experiment highlighting the role of Fourier phase. Fourier phase is swapped between images while Fourier magnitude information is retained. We see that image structures are approximately preserved corresponding to the phase spectra of original images. The quality of images can be improved significantly by incorporating their own Fourier magnitude or the magnitude information from similar images.

2.2 Local Phase

2.2.1 Signals in time-frequency domain

Usually signals are represented either completely in the time domain or in the frequency domain. Information conveyed is the same, but in different forms. Both the domains have complementary advantages and disadvantages. Fourier transform is a popular method to transform signals from

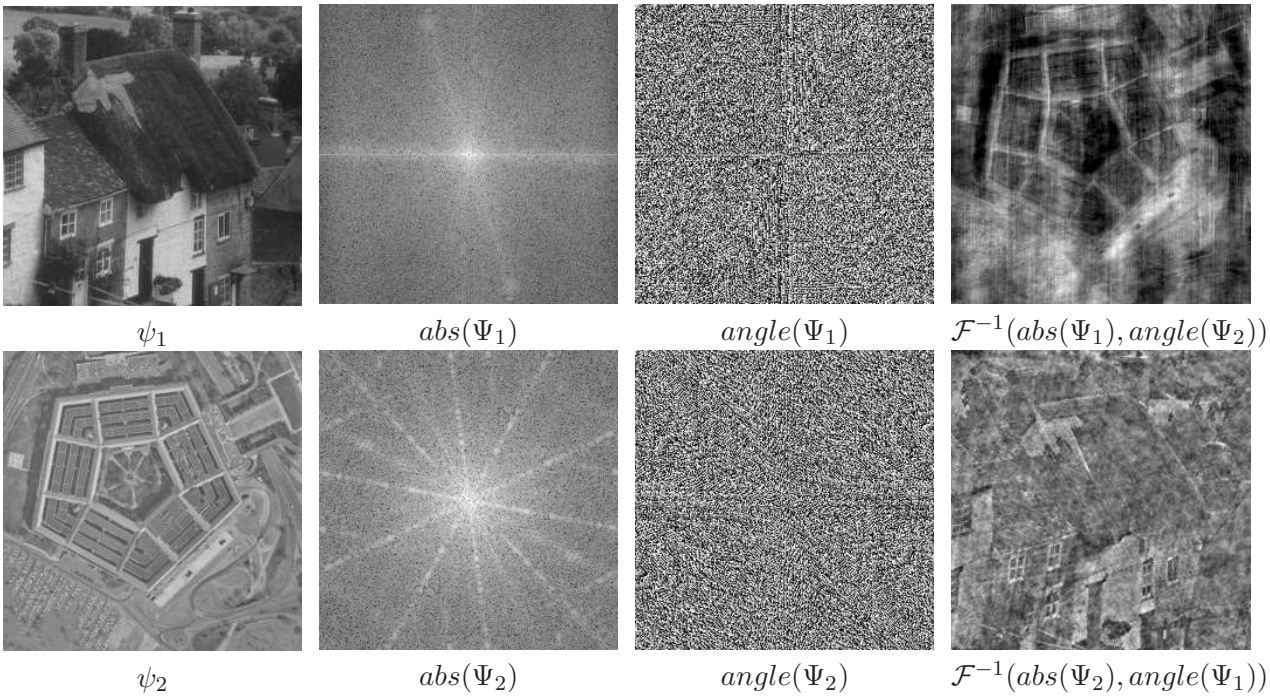


Figure 2.2: Phase swapping experiment showing the importance of Fourier phase; ψ_1 and ψ_2 are two images. Ψ_i denote the Fourier transform of ψ_i. $abs(\ )$ and $angle(\ )$ denote magnitude and phase information respectively of a signal.

one domain to another and vice-versa. Some of the image processing tasks deal with analyzing and characterizing various features or building blocks of an image. Further the goal is to locate them accurately in spatial coordinates. e.g., one may want to process only edges in an image and to extract their locations. High frequency content represent edges. And simultaneously edge location need to be known in spatial domain. The relative configuration of image features in an image is important for object analysis, detection and recognition. Another popular example from music composition is to determine what type of events or tones it contain and at what instances they are played. Classes of events or tones are well represented in frequency domain. In order to address these requirements following two questions should be addressed,

- Is there a way to represent an image content so as to preserve the advantages of both time and frequency domains ?
- What is the smallest quanta of information or building block of images. How should they be represented ?

Gabor in 1946 proposed elementary functions to represent signals simultaneously in time and frequency domain [83]. Gabor elementary functions was based on Heisenberg's uncertainty principle. These functions represents the minimum quantum of simultaneous information that occupies a minimal area in the time-frequency domain. This led to the development of windowed Fourier transformation and wavelets. In this section, we describe Heisenberg's uncertainty principle as applied to the time-frequency analysis of signals. We also describe the Gabor function and local phase computed from Gabor filters. Detailed and comprehensive treatment on these topics can be found in [83, 77, 84, 85].

2.2.2 Uncertainty in Localization

To derive the benefits of both the domains, the information should be localized in frequency and time domains simultaneously. The goal is to derive an operator that can analyze a signal simultaneously and optimally. The minimal amount of information which can be analyzed in both the domains is bounded by the uncertainty principle. To give an intuitive idea, let's consider a case when the characteristic of a signal has to be analyzed at a precise time instance. Going by Fourier transform equations, a complete frequency spectra is required. The same arguments go for analyzing the signal at a given frequency. If the accuracy regarding the exact time location and frequency location can be sacrificed, we can analyze a signal locally both in time and frequency domain.

Let Δt, also known as *spatial-width*, is the uncertainty of measurement in time duration and Δf, also known as *bandwidth*, is the uncertainty in frequency measurement. Let $\psi(t)$ and $\Psi(f)$ are operators which can analyze signals simultaneously in both domains. Dennis Gabor used root mean square bandwidth as the square root of second centralized moment of a properly normalized form of the squared spectrum about a suitably chosen point. It represents the deviation from a mean value and it is accepted as a measure of uncertainty. A similar measure is defined to specify uncertainty in time. For simplification, all equations are discussed for 1-D case only. The uncertainties are mathematically formulated as,

$$\Delta t = \sqrt{\frac{\int_{-\infty}^{\infty}(t-\mu_t)^2\psi(t)\psi^*(t)dt}{\int_{-\infty}^{\infty}\psi(t)\psi^*(t)dt}} \tag{2.7}$$

$$\Delta f = \sqrt{\frac{\int_{-\infty}^{\infty}(f-\mu_f)^2\Psi(f)\Psi^*(f)df}{\int_{-\infty}^{\infty}\Psi(f)\Psi^*(f)df}} \tag{2.8}$$

where,

$$\mu_t = \frac{\int_{-\infty}^{\infty} t\psi(t)\psi^*(t)dt}{\int_{-\infty}^{\infty}\psi(t)\psi^*(t)dt}, \tag{2.9}$$

$$\mu_f = \frac{\int_{-\infty}^{\infty} f\Psi(f)\Psi^*(f)df}{\int_{-\infty}^{\infty}\Psi(f)\Psi^*(f)df}. \tag{2.10}$$

where μ_t and μ_f can be interpreted as the mass centroids or means of function $\psi(t)$ in time and $\Psi(f)$ in frequency. Above two definitions of uncertainties are connected via Heisenberg's uncertainty principle as,

$$\Delta t \Delta f \geq \frac{1}{4\pi}, \tag{2.11}$$

i.e. for any function that analyzes a signal simultaneously in both the domains, the product of their spatial width and bandwidth assumes a value greater than or equal to a constant.

2.2.3 Band-pass Filters and Gabor Filters

Bandpass filters allow frequencies in a range to pass through and reject frequencies outside the range. Bandwidth of a filter is defined as the effective difference between the upper and lower cut-off frequencies. As discussed earlier, there is trade-off between the selection of spatial width and bandwidth. Smaller spatial width is required for accurate localization and smaller bandwidth is required for measurement of local frequencies and accurate computation of local phase.

Gabor derived a function [83] for which the product $\Delta t \Delta f$ assumes the smallest possible value. The inequality in equation 2.11 turns into an equality.

The signal which occupies the minimum area , $\Delta t \Delta f = \frac{1}{4\pi}$, is the modulation product of the harmonic oscillation of any frequency with pulse of the form of a probability function (e.g., Gaussian envelope).

$$\psi(t) = g(t)s(t) \tag{2.12}$$

$$\psi(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-t_0)^2}{2\sigma^2}} e^{j2\pi f_0 t + \phi}, \tag{2.13}$$

where σ is the sharpness of the Gaussian, t_0 denotes the centroid of the Gaussian, f_0 is the frequency of the harmonic oscillations, and ϕ denotes the phase shift of the oscillation. $g(t)$, the Gaussian shaped function, is also known as *envelope* and $s(t)$, the complex sinusoidal function, is also known as *carrier*. The function has a Fourier function of analytical form,

$$\Psi(f) = e^{-2\pi^2\sigma^2(f-f_0)^2} e^{-j2\pi t_0(f-f_0)+\phi}, \tag{2.14}$$

It is easy to show from equation 2.13 and 2.14 that $\mu_t = t_0$, $\mu_f = f_0$, $\Delta t = \frac{\sigma}{\sqrt{2}}$, $\Delta f = \frac{1}{2\sqrt{2}\pi\sigma}$ and $\Delta t \Delta f = \frac{1}{4\pi}$. Gabor functions may form an expansion space, where the distinct advantage is a representation by optimally localized time-frequency kernels. A signal can be represented as a sum of finite number of Gabor elementary functions multiplied with specific expansion coefficients.

Gabor filters in 2D

Similarly, 2D normalized formulation of a Gabor filter has an analytical form,

$$\psi(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-(\frac{x'^2}{2\sigma_x^2} + \frac{y'^2}{2\sigma_y^2})} e^{j2\pi(f_x x' + f_y y')}, \tag{2.15}$$

where, $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, (f_x, f_y) is the frequency of the filter, σ_x and σ_y controls the spatial width of the filter, θ is the orientation of the filter, and j is $\sqrt{-1}$. Fig. 2.3 illustrate the filter structure. To extract local frequencies, an image is convolved with a bank Gabor filter. If an image has local frequencies almost same as that of a Gabor filter, at central locations, it responds higher at all these pixels. The band-pass nature of the filter is clear by Fourier representation of a Gabor filter Fig. 2.3(d). Convolution in spatial domain is multiplication in frequency domain.

Bandwidth

The bandwidth (in octaves) of a bandpass filter is the base 2 logarithm of the upper and lower cut-off frequencies of the filter. Let f_0 denote the central frequency, and $\Delta\tau$ denote the half-bandwidth of the filter. Full bandwidth of the filter, Δf, is given by the equation 2.8. After simplifying the equation for Gabor filter we get $\Delta f = \frac{1}{2\sqrt{2}\pi\sigma}$. The relative full bandwidth in octave is obtained as,

$$b = \log_2\left(\frac{f_0 + \Delta\tau}{f_0 - \Delta\tau}\right). \tag{2.16}$$

2.2.4 Local Phase from a Bandpass Filter

Any bandpass filter, with a finite support, can be used for extracting the local phase information from an image. Gabor filters are commonly used as band pass filters as they achieve the theoretical

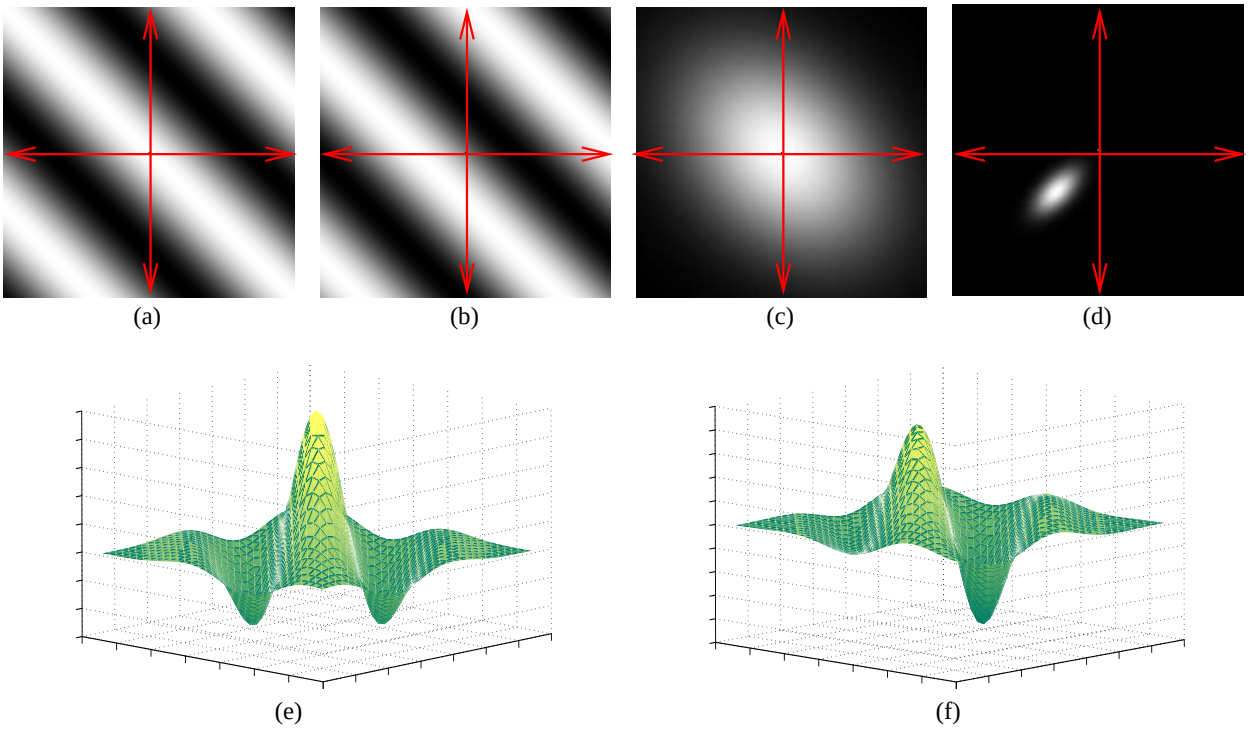


Figure 2.3: Gabor filter is a multiplication of a complex sinusoid with a Gaussian kernel. (a) real part of the complex sinusoid; (b) imaginary part of the complex sinusoid; (c) Gaussian kernel; (d) Fourier transform of the Gabor filter; (e) 3D graph showing multiplication of real part of complex sinusoid with Gaussian kernel; (f) multiplication of imaginary part with Gaussian kernel. Parameters are $\sigma_x = 50$, $\sigma_y = 70$, $f_x = 1/125$, $f_y = 0$, $\theta = \pi/4$ for (a), (b), (c), (e) and (f). $\sigma_x = 2.5$, $\sigma_y = 5$, $f_x = 1/5$, $f_y = 0$, $\theta = \pi/4$ for (d).

minimum of product of spatial width and bandwidth for any complex valued linear filter. A smaller bandwidth allows accurate computation of local phase and smaller width is desirable for localization.

Let $i(x, y)$ be the image. Local phase is computed at each image location by convolving the image with the Gabor filter $\psi(x, y)$ (equation 2.15) as,

$$s_m(x, y, f_x, f_y) = i(x, y) * \psi(x, y, f_x, f_y). \tag{2.17}$$

The local phase in image i at (x, y) is computed as,

$$\phi_m(x, y, f_x, f_y) = \arg[s_m(x, y, f_x, f_y)], \tag{2.18}$$

where arg[ ] is the complex argument in $(-\pi, \pi]$. Fig. 2.4 shows a hypothetical example for extracting signal components. The signal is convolved with a Gabor filter or a similar band-pass filter at a given frequency. Sinusoid extracted has horizontal axis as local phase, which is a function of spatial location and a vertical axis as amplitude.

2.2.5 Local Phase Difference Computation

Instead of using intensity values for comparing two images, local phase information has been found to be more robust against noise and contrast variations. Image matching tasks involve finding

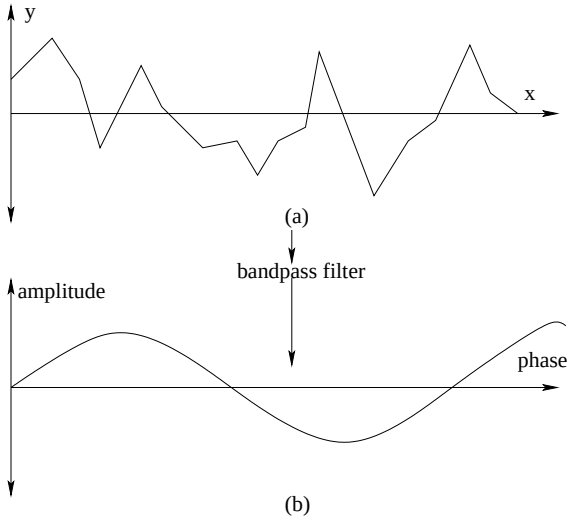


Figure 2.4: (a) segment of a hypothetical signal (b) segment after applying bandpass filter, horizontal axis denote the local phase which is a function of spatial location and vertical axis is the amplitude.

the corresponding points in two images. Assuming that the corresponding point is not very far away, the local phase difference values directly predict the location of corresponding points without explicit matching or signal reconstruction. Let $i_1(x, y)$ and $i_2(x, y)$ be the two images. These images are convolved with Gabor wavelet and local phase is computed at each spatial location at a given frequency. The phase difference is computed as,

$$\Delta\phi(x, y, f_x, f_y) = [\phi_2 - \phi_1]_{2\pi}, \tag{2.19}$$

where, $\phi_1(x, y, f_x, f_y)$ and $\phi_2(x, y, f_x, f_y)$ are the local phase map of images $i_1(x, y)$ and $i_2(x, y)$ respectively at (f_x, f_y). It is assumed that the corresponding points in two images lie within a cycle of the sinusoid.

For robust stereo disparity computation [77] local phase difference values computed at different frequencies are useful. Stereo correspondence problem often involves computing and matching the features in two images. Due to noise and illumination variations computation of these features is a challenging problem. Computing correspondences are challenging and could be misleading in the absence of sufficient number of features. Many stereo images are captured at close distance on camera baseline. Disparity has to be computed at sub-pixel level to estimate the correct relative depth for such stereo pairs.

If two images are shifted relative to each other by an amount Δx, according to Fourier shift theorem the difference in phase is proportional to the total shift between two images. Similar theoretical formulation is used to calculate disparity in a stereo pair. i.e., the local disparity is approximately equivalent to the phase difference at a particular frequency divided by the underlying spatial frequency of that signal. The computation has a certain degree of uncertainty but the correspondence computation does not involve explicit feature matching or signal reconstruction. This approach is totally correspondence-less. In practice, the estimate is computed by convolving with multiple filters, which guarantee that the error due to band-limited noise is minimum. Phase difference values are combined in an appropriate way to compute highly accurate disparity map.

2.2.6 Advantages of using Local Phase

Like global phase, local phase holds similar properties which make it useful for various computer vision problems.

- *Robustness towards Noise:* Fleet and Jepson [84] showed that the phase is more robust for image matching than the amplitude of the filter response in presence of noise. For band-limited noise, the error in the estimation is reduced by considering the phase output of those filters that do not allow those frequencies to pass through. This is done by assigning low scores to those phase difference estimates, where there is a significant amplitude mismatch in both the signals detected.
- *Contrast Invariance:* Illumination change or contrast, in the image space, is the multiplication of pixel value by another value. Smooth illumination can be modeled by the multiplication of a constant in a window. The phase information computed at these two locations will remain unchanged as compared to the magnitude of the signal, which will be scaled by the illumination constant.
- *Correspondence-less Matching:* As discussed before, the local disparity can be computed by using only the phase difference value and the underlying frequency of the signal. No explicit correspondence computation, feature matching or signal reconstruction is actually done.
- *Parallel Implementation:* Local phase computation at each spatial location depends only on the nearby pixel values. Also, the phase difference computation can be achieved without signal reconstruction. So, local phase difference computation can be easily parallelized.

2.2.7 Biological Motivation for using Gabor filters and Local Phase

Much of the usage of Gabor filter for extracting local frequencies and local phase for image matching is motivated by similar biological operations in human eyes and primate visual cortex. Sanger [77] and references therein lists several biological inspirations for using them.

- Certain experiments shows that the visual cortex encodes information using band-pass spatial frequency filters. It motivates the usage of band-pass filters for image analysis tasks. Errors due to band-limited image noise can be reduced by combining information from different bands.
- The shape profile of a Gabor filter is similar to the receptive field profile of simple cells in primate visual cortex. It motivates to decompose image information into small quantas using Gabor filters.
- It has been discovered that simple cells occur in pairs with quadrature relative phase in primate visual cortex. Phase quadrature means 90 degrees out of phase. It could represent the phase of a complex filter.

In summation, local phase has been found to be more useful than the amplitude of the filter. Local phase characterize structural information whereas amplitude information characterize how various small image quantas should be combined. Biological findings further motivates the use of local phase for image analysis.

2.3 Low-level Vision and Markov Random Field

Various interpretation and recognition tasks can be divided into three modules based on processing similarities, level and scale *viz. low-level*, *mid-level* and *high-level* vision. Low-level vision problems

include primitive processing, estimation and enhancements. It aims to recover meaningful description of the input intensities. Examples includes simple problems like noise removal, stereo, motion analysis, and inferring shape and reflectance from images. Mid-level vision tasks include fitting parameters to data (e.g. image segmentation, tracking, clustering, etc.) High-level vision problem includes recognition, classification and meaningful interpretation tasks.

Some of the low-level vision tasks involve modeling the relationship of intensity values in a context. Most of the low-level tasks are ill-posed. By regularizing solutions or by providing explicit hypothesis underlying scene details can be estimated uniquely. But these methods lack generalizability. Another challenge is towards adaptability of the algorithm according to the underlying information and the context.

In contrast to hypothesis based methods, learning based low-level vision algorithms aims to learn the model from the sample-data itself. The framework allows to adapt to the underlying image data and the context. *Markov Random Field* is a popular probabilistic framework which is suitable for modeling contextual information. It allows the relationship to be expressed locally and information is propagated globally through neighborhood locations. In this section, we provide an overview and sufficient understanding of Markov Random Field. Detailed theoretical analysis can be found in [86, 87, 88]. We also discuss the framework provided by Freeman *et al.* [37] for learning low-level vision tasks.

2.3.1 Graphs and Neighborhoods

Let $\mathcal{S} = \{s_1, s_2, ..., s_N\}$ denote the set of *sites* and $\mathcal{G} = \{\mathcal{G}_s, s \in \mathcal{S}\}$ be a *neighborhood system* for $\mathcal{S}$ consisting of subsets of $\mathcal{S}$ satisfying two properties: a) $s \notin \mathcal{G}_s$, a site is not neighboring to itself; b) $s \in \mathcal{G}_r \Leftrightarrow r \in \mathcal{G}_s$, the neighboring relationship is mutual. $\{\mathcal{S}, \mathcal{G}\}$ is the graph and $\mathcal{G}_s$ is a set of neighbors of s.

For computer vision and image processing applications, graphs and neighborhoods are defined on an integer lattice $\mathcal{Z}_{mn}$. Let $\mathcal{S} = \mathcal{Z}_{mn} = \{(i,j) : 1 \leq i \leq m, 1 \leq j \leq n\}$ be the $m \times n$ integer lattice defined over an image space. A homogeneous neighborhood $\mathcal{G}$ is defined as,

$$\mathcal{G} = \{\mathcal{F}_{i,j}, (i,j) \in \mathcal{Z}_{mn}\}, \tag{2.20}$$

and

$$\mathcal{F}_{i,j} = \{(k,l) \in \mathcal{Z}_{mn} : 0 < (k-i)^2 + (l-j)^2 \leq c\}, \tag{2.21}$$

where c denote the order of neighborhood. Fig. 2.5(a), 2.5(b) and 2.5(c) show the neighborhood configurations for $c = 1, 2$ and 8 respectively.

A *clique*, denoted by $\mathcal{C}$, is a subset of $\mathcal{S}$ such that every pair of distinct sites in $\mathcal{C}$ are neighbors i.e., a clique is a fully connected graph; $\mathbb{C}$ is the set of such cliques. A clique of size t has exactly t nodes with each pair of node connected by an edge. A clique system should not be confused with the neighborhood system. Neighborhood configuration at $c = 1$ can have cliques as shown in Fig. 2.5(d). Similarly, neighborhood configuration at $c = 2$ has cliques as shown in Fig. 2.5(d) and Fig. 2.5(e). The complexity of the clique type grow rapidly with c.

2.3.2 Markov Random Fields

Markov random field is a type of stochastic process. A very specific form of Markov random field, called Ising Model, was proposed in physics to explain certain empirically observed facts of ferromagnetic materials. To understand Markov random field intuitively, an example from sociology [89] is usually given. Consider a group of people, who at a given moment can take either of two stands

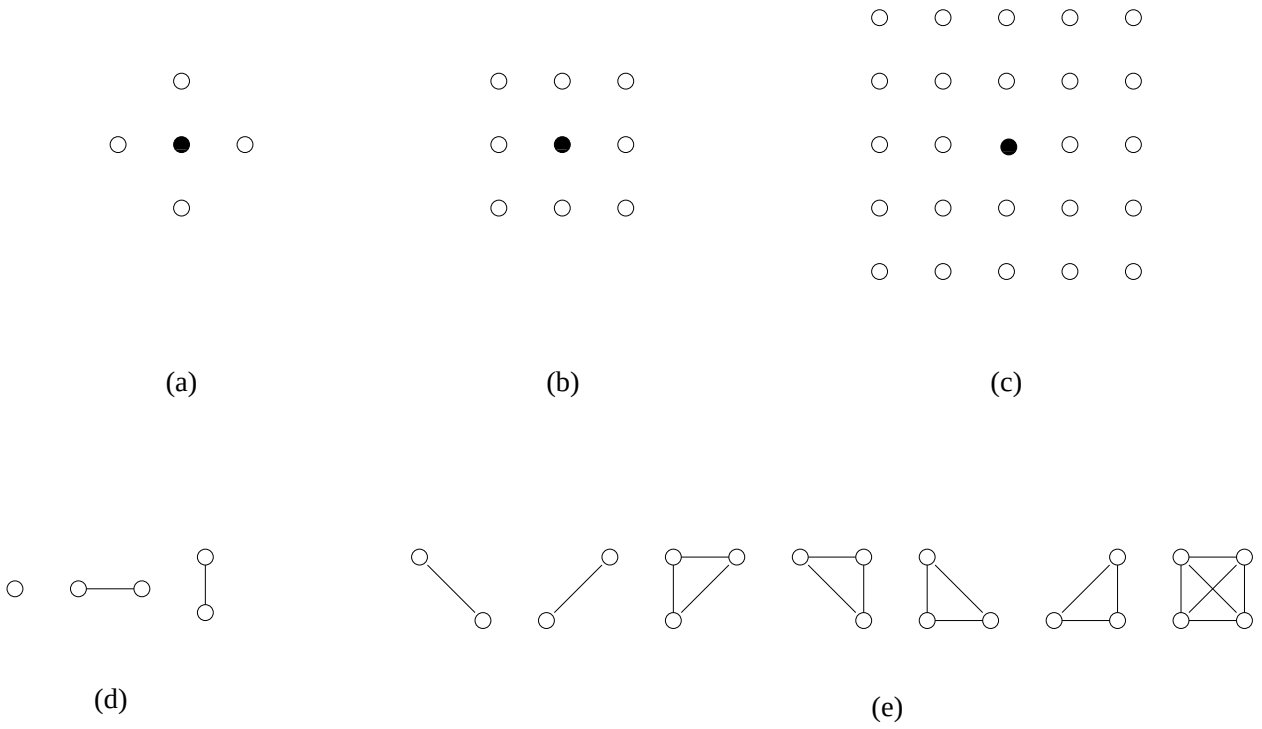


Figure 2.5: Neighborhood configurations at (a) $c = 1$, (b) $c = 2$ and (c) $c = 8$; (d),(e): various clique types on a lattice of regular sites.

up (↑) or down (↓). The total energy of the system is equivalent to the amount of tension in the system which is a sum of two terms. First, due to interaction with people they know on the basis of extent they dis-agree. Interaction level and hence the tension may vary from person to person. Second, on the current state of the government either up (↑) or down (↓). Tension can be minimized by agreeing with the people they know and/or agreeing with the stand of the government. Minimum tension occurs if maximum number of people agree with the government. The goal would be to know what would be the total tension of the system after a series of interactions are allowed. Also, what would be the tension of the system if society is of a more liberal or of totalitarian nature. Long range interactions are possible in a liberal system and in totalitarian system interactions are short range and phase transition can happen. It may so happen that attitude of one person by chance can take hold and it can spread to all over the society. Liberal nature of the system is equivalent to high temperature which allow long range interaction. Totalitarian system is represented by low temperature values. A neighborhood system of a person are group of people he know.

Markov random field is widely used framework to model various problems in computer vision. It provides an easy framework to incorporate contextual constraints. A *field* is a map that labels or assigns every point in a space by a function. Examples of field include Newtonian gravitational field, magnetic field, etc. A *random-field* is a set of random numbers whose values are mapped to a n-dimensional space. Values in a random-field are usually spatially-correlated with each other. Examples of random-fields include Markov random field, Gibbs random field, Conditional random field, etc.

Let $\{\mathcal{S}, \mathcal{G}\}$ denote an arbitrary graph, $\mathcal{X} = \{\mathcal{X}_s, s \in \mathcal{S}\}$ denote any family of random variables indexed by $\mathcal{S}$, Λ denote a set of all possible numbers or labels that can be assigned to a site. i.e.,

$\mathcal{X}_s \in \Lambda$ for all s and let Ω denote the set of all possible *configurations* such that,

$$\Omega = \{\omega = (x_{s_1}, ...x_{s_N}) : x_{s_i} \in \Lambda, 1 \leq i \leq N\}, \tag{2.22}$$

Any configuration is abbreviated as $\{\mathcal{X} = \omega\}$. A configuration is any possible assignment of labels to each node of the graph. $\mathcal{X}$ is a MRF with respect to $\mathcal{G}$ if following two conditions hold,

- $P(\mathcal{X} = \omega) > 0, \quad \forall \omega \in \Omega \qquad$ *(Positivity)*
- $P(\mathcal{X}_s = x_s | \mathcal{X}_r = x_r, r \neq s) = P(\mathcal{X}_s = x_s | \mathcal{X}_r = x_r, r \in \mathcal{G}_s), \qquad$ *(Markovianity)*

for every $s \in \mathcal{S}$ and $\omega \in \Omega$. The first condition implies that each configuration of assignment is possible. Second condition implies that the probability of a random variable $\mathcal{X}_s$ at site s depends only on the nearest neighbors and the site is conditionally independent of all other vertices in the graph. Various problems in computer vision (e.g. denoising, stereo, etc.) require such frameworks. The collection of functions on the left-hand side of second condition is called the *local characteristics* of the MRF.

Labeling Problem: Many problems in computer vision can be modeled as labeling problem. The labeling problem is to assign a label from the label set Λ to each of the site in $\mathcal{S}$. For example, image segmentation is a two label problem *viz.* foreground and background and stereo reconstruction is a multi-label problem with number of possible disparities as number of labels. In these problems, neighborhood labels are correlated. Labeling problem can be modeled in a MRF framework and given the images, the Maximum a-Posteriori(MAP) estimate of the labels can be computed.

Gibbs Distribution

There are multiple reasons to include Gibbs distribution in discussion and compare the equivalence with MRF. Gibbs random field is characterized by its global property (Gibbs distribution) whereas MRF is characterized by its local property (Markovianity). At each site in the network, the probability of a label depends only on the labels of neighboring sites. Joint probability distribution of the $\mathcal{X}_s$ is not apparent because it depends on neighboring states which itself is unknown and in-turn depend on their respective neighbors. Also, it is difficult to impose a desired local behavior i.e. it is difficult to visualize when a given set of functions could be a conditional probability distribution on Ω.

A *Gibbs distribution* relative to $\{\mathcal{S}, \mathcal{G}\}$ is a probability measure π on Ω with,

$$\pi(\omega) = \frac{1}{Z} e^{-\mathcal{U}(\omega)/\mathcal{T}} \tag{2.23}$$

where Z is a normalizing constant (also known as *partition function*), $\mathcal{T}$ is the *temperature* and $\mathcal{U}$ is the *energy function*, which is a sum of *clique potentials* mathematically defined as,

$$\mathcal{U}(\omega) = \sum_{\mathcal{C} \in \mathbb{C}} V_{\mathcal{C}}(\omega), \tag{2.24}$$

$V_{\mathcal{C}}$ functions represent contributions to the total energy from external fields (clique of size 1), pair interactions (clique of size 2) and so on. There were many limitations of MRF formulations which were perceived by many authors. But Markov-Gibbs equivalence theorem also known as *Hammersley-Clifford* theorem provides a simple way to specify joint probability. According to the theorem:

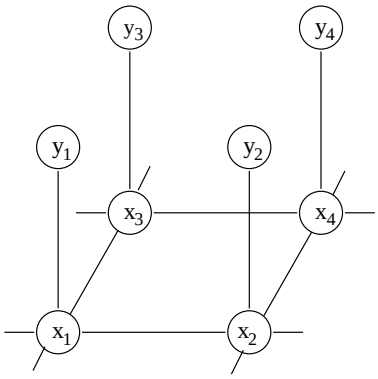


Figure 2.6: Markov network for low-level vision problems. Each node corresponds to a patch of a scene or an image. Edges connecting nodes indicate the statistical dependency between nodes.

Theorem 1. *Let $\mathcal{G}$ be a neighborhood system. Then $\mathcal{X}$ is an MRF with respect to $\mathcal{G}$ if and only if $\pi(\omega) = P(\mathcal{X} = \omega)$ is a Gibbs distribution with respect to $\mathcal{G}$.*

Proof can be found in [87] and the references therein. As a consequence of this theorem, joint probability can be specified only by its potentials instead of local characteristics. MAP estimation problem reduces to energy minimization problem. There are explicit formulas which can calculate $\mathcal{U}$ from the local characteristics but obtaining it from the clique potentials is much simpler and practical.

2.3.3 Learning Low-Level Vision

As mentioned before, low-level vision problems are ill-posed. Prior information is learned from the training data rather than hypothesized. Algorithms are highly adaptable to the underlying scene and context. Markov random-field is core to model such requirements. In the context of low-level vision problems, an *image* is given as input and the desired characteristics to be estimated is called as *scene*. Let $\mathbf{x}$ is the given image and $\mathbf{y}$ is the scene to be estimated. The MAP estimation of the underlying scene is,

$$\mathbf{x}_{MAP} = \arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}) \tag{2.25}$$

$$= \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})P(\mathbf{x}). \tag{2.26}$$

Such formulation poses a tremendous burden on the learning and inference phase. The image and scene is divided into small patches under Markov assumption as $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ and $\mathbf{y} = \{y_1, y_2, ..., y_N\}$. Each node in a Markov network corresponds to a patch of an image or a scene. Figure 2.6 shows a toy Markov network for such problems. Each scene node is statistically dependent on the underlying image patch and 4 nearest neighbors. Patches are shown by circles and statistical dependencies are shown by edges. The likelihood and the prior probability term can be expanded as,

$$p(\mathbf{y}|\mathbf{x}) = \prod_k \Phi(x_k, y_k) \tag{2.27}$$

and

$$P(\mathbf{x}) = \prod_{(i,j)} \Psi(x_i, x_j) \tag{2.28}$$

where k is over all image nodes and the pair (i, j) indicates neighboring nodes, Φ and Ψ are pairwise compatibility functions learned from the training data. The joint probability is given by,

$$P(\mathbf{x}|\mathbf{y}) = \prod_{k} \Phi(x_k, y_k) \prod_{(i,j)} \Psi(x_i, x_j) \tag{2.29}$$

Above energy function is maximized to estimate the scene.

Inference

Due to the presence of cyclic dependencies in the network, it is computationally infeasible to solve it for global minima. Applying Bayesian inference methods in a network has been found to be useful even in presence of cycles by Freeman *et al.* [37]. The equation is optimized by iterating the following steps. The MAP estimate at node j is,

$$\hat{x}_{jMAP} = \arg\max_{x_j} \Phi(x_j, y_j) \prod_{k} M_j^k, \tag{2.30}$$

where k is over all neighbors of node j, and M_j^k is the message from node k to node j, given by,

$$M_j^k = \max_{[x_k]} \Psi(x_j, x_k) \Phi(x_k, y_k) \prod_{l \neq j} \tilde{M}_k^l, \tag{2.31}$$

where $\tilde{M}_k^l$ is M_j^k from the previous iteration. It is difficult to optimize the equation 2.29 because of high dimensionality of the scene variables and large number of patches. Freeman *et al.* [37] favored to obtain a local minimal solution, which approximates the solution. First, a small set of similar patches (usually 10-20) are obtained using Approximate Nearest Neighbor data-structure [90]. After that belief propagation rules are used to solve the network. In section 2.4.3, learning based super-resolution technique is described along with how compatibility functions are learned.

In summation, we described how Markov random-field is used to solve ill-posed low level vision problems. The technique shows that a large database can be used to solve scene interpretation problems. This framework is used in chapter 4 to capture right amount of scene information for image super-resolution.

2.4 Super-Resolution

The *resolution* of an image refers to the scene details an image can hold. In other words, it is the ability to distinguish scene details in an image. A *high-resolution* image has finer scene details as compared to a *low-resolution* image. Image *Super-resolution* (SR) is a process of simulating a high-quality and high-resolution camera from the image(s) captured by a low-quality and low-resolution camera. Discrete sampling and box averaging of an irradiance field at each pixel of the sensor limits the capturing of detailed scene information. Additionally, sensor and environmental noise and blurring further degrade the quality of a captured image. Super-resolution algorithms combine information from either multiple captured images or prior information stored in a training database to compute high-frequency information and remove various imaging artifacts.

Generating high-resolution images from degraded low-resolution images has a variety of applications in space imaging, medical imaging, surveillance and commercial videography. Space imaging tasks require the image to be captured at highest possible resolution. Existing imaging apparatus and telescopes are limited to capture details up to a particular limit. Super-resolution algorithms

can enhance the resolution of these images by a certain magnification factor (usually up to 3-4). For surveillance tasks, it may not be possible to capture all finer scene details. During identification and analysis of objects/subjects multiple frames of the videos are fused to obtain a single high-resolution image with finer details using such a technique. Old NTSC format videos can be converted into new high quality HDTV format.

The aim of this section is to understand basic image degradation process. We also provide a basic overview of multi-frame and single-frame super-resolution algorithms. For comprehensive understanding of the topic several references of various super-resolution algorithms are also provided.

2.4.1 Imaging Model

Imaging model describes a general relationship between the scene (or a high-resolution image) and the images captured by the camera's CCD (or low-resolution images). There are numerous imaging artifacts and optical-distortions that can be incorporated in the imaging model. In Super-resolution literature, for simplicity, only a few artifacts are modeled. The imaging model is formulated in terms of blurring artifacts of the camera and box-averaging of irradiance field on the sensor, noise and non-ideal sampling of irradiance field on image sensor. Multiple images are captured for a class of SR algorithms at sub-pixel displacements. So, all these artifacts are modeled separately for each of the captured image. Also, the illumination variations across frames is also incorporated in the model. Mathematically the process can be described as,

$$\mathbf{y}_k = \mathbf{L}_k\mathbf{D}_k\mathbf{B}_k\mathbf{F}_k\mathbf{x} + \mathbf{n}_k, \quad 1 \leq k \leq n, \tag{2.32}$$

where, n is the total number of images captured. $n = 1$ for single frame super-resolution algorithms. $\mathbf{x}$ is the ideal high-resolution image, $\mathbf{y}_k$ are degraded, low resolution images. These low-resolution and high-resolution images are mathematically represented as column matrices by concatenating multiple columns of the image into a single column. $\mathbf{F}_k$ represents the motion parameters of frame k with respect to a reference frame. $\mathbf{B}_k$ is the blurring matrix. It captures the blurring due to camera lenses and sensor averaging of irradiance field. The irradiance field falling on a pixel of the chip is averaged (usually box averaging or Gaussian averaging) out and a single intensity value is assigned for that pixel region. This is also known as camera's point spread function (PSF). $\mathbf{D}_k$ represents the decimation operator. $\mathbf{L}_k$ is a diagonal matrix, which designate illumination variations with respect to a reference image. Illumination change is represented by multiplication of pixel values by a number at each pixel location. The multiplication factor vary only slightly around the pixel neighborhood. For simplification, $\mathbf{L}_k$ is adjusted at the beginning of image formation equation rather than placing with $\mathbf{x}$. $\mathbf{n}_k$ is a noise vector. Fig. 2.7 shows pixel structures on low-resolution and high-resolution chip.

2.4.2 Multi-frame Image Super-Resolution

All super-resolution algorithms incorporate information from other sources apart from the information available in the given image. In multi-frame image super-resolution, multiple images of the same scene are captured. These images are taken at non-integer relative pixel displacements to gain non-redundant information in each capture. These images are registered with respect to a common image at sub-pixel level. Blur parameters are estimated and decimation matrix is specified by the user. Equation 2.32 is solved to get high-resolution, high-quality image. Apart from the movement of camera, movements of the object in different frames can also provide non-redundant information. Tsai and Huang [91] first proposed improving image resolution

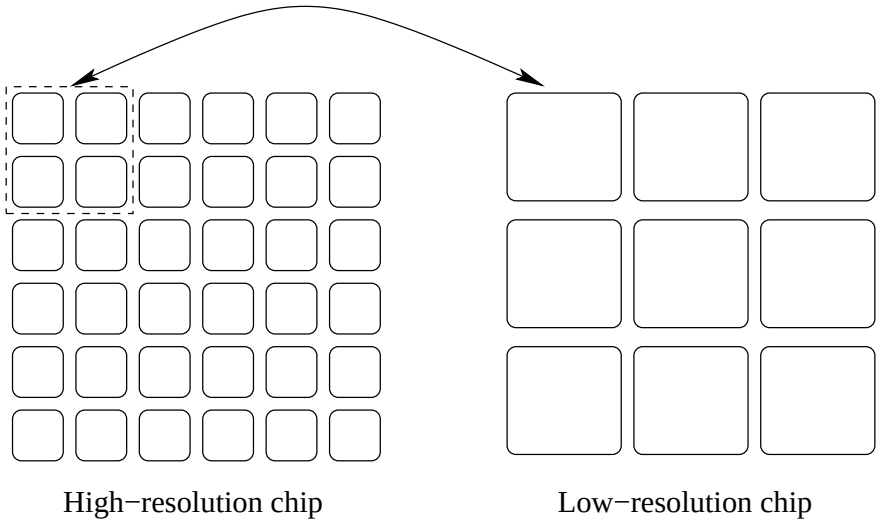


Figure 2.7: High-resolution images are captured on a dense high quality chip having more pixels per unit area whereas low-resolution image is captured on a low-quality and less-dense chip.

by using multiple satellite images. Thereafter many super-resolution algorithms have been proposed [92, 93, 94, 27, 28, 29, 30, 31, 32, 33, 34]. [35] and [36] provide a comprehensive literature survey on multi-frame SR algorithms.

Fig. 2.8 shows an example of multi-frame image super-resolution. Images captured at sub-pixel displacements provide scene irradiance information at locations on pixel grid, where the information was missed in initial captures. As scene information is box averaged at each pixel, multiple frames provide a system of equations to solve for high-resolution image pixels.

Super-resolution enhancements are divided into two categories *viz.* *frequency domain* and *spatial domain* algorithms. Frequency domain techniques use the shifting property of the Fourier transformation to model global translation and take the advantage of sampling theory for image restoration. Frequency domain techniques are simple and computationally faster. But these methods fail to incorporate wider range of transformations between frames. Degradation models, in frequency domain, can not incorporate spatially varying artifacts. On the contrary, spatial domain algorithms are slow but can accommodate a wider range of image degradations. Additionally, prior information or regularization can be incorporated with ease.

Estimating High-Resolution Image

We briefly describe maximum likelihood estimate of high-resolution frame. Let's assume that there is no illumination variation across frames, so the term $\mathbf{L}_k$ in equation 2.32 can be dropped. Let $\mathbf{H}_k = \mathbf{D}_k\mathbf{B}_k\mathbf{F}_k$. The image formation equation can be re-written as,

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x} + \mathbf{n}_k. \tag{2.33}$$

Images are assumed to have been corrupted with Gaussian white noise. It means that noise $\mathbf{n}_k$ is assumed to have occurred independently at each pixel location and that they are distributed according to a Normal distribution with zero mean and unknown variance σ^2. As $\mathbf{H}_k\mathbf{x}$ is a non-random quantity, consequently, $\mathbf{y}_k$ has mean $\mathbf{H}_k\mathbf{x}$ and variance σ^2. The Gaussian white noise has a auto-correlation matrix $\Sigma_k = E\{\mathbf{n}_k\mathbf{n}_k^T\} = \sigma^2 I$. The maximum likelihood estimate of $\mathbf{x}$ is thus

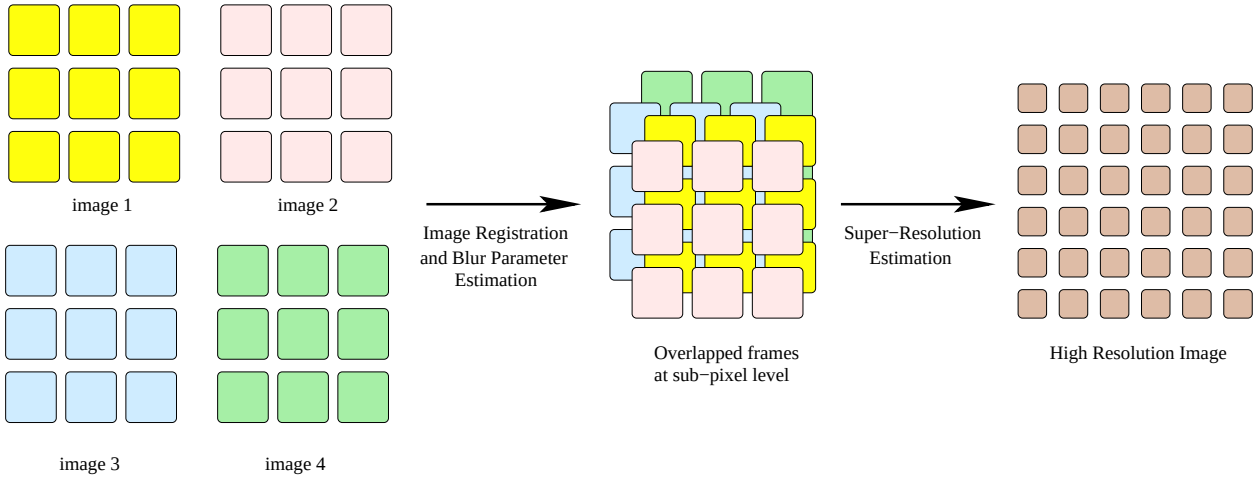


Figure 2.8: Example showing four images of the same scene captured at sub-pixel displacements. These images are registered at sub-pixel level and the high-resolution image is computed. Each of the square pixel represents the effective pixel size of camera's CCD while capturing an image.

given by,

$$\mathbf{x}_{ML} = \arg\max_{\mathbf{x}} \prod_k p(\mathbf{y}_k|\mathbf{x}) \tag{2.34}$$

$$= \arg\max_{\mathbf{x}} \prod_k \frac{1}{Z} e^{-\frac{1}{2}[\mathbf{y}_k - \mathbf{H}_k\mathbf{x}]^T \Sigma_k^{-1} [\mathbf{y}_k - \mathbf{H}_k\mathbf{x}]} \tag{2.35}$$

Rather than maximizing the above expression we maximize its logarithm, which is simple. After substituting $\Sigma_k = \sigma^2 I$ and taking the logarithm, the above equation is reduced to a least square estimate as,

$$\mathbf{x}_{ML} = \arg\min_{\mathbf{x}} \sum_k ||\mathbf{y}_k - \mathbf{H}_k\mathbf{x}||^2. \tag{2.36}$$

Let,

$$\mathbf{L}(\mathbf{x}) = \frac{1}{2} \sum_k ||\mathbf{y}_k - \mathbf{H}_k\mathbf{x}||^2 \tag{2.37}$$

To minimize the above expression we take the derivative of $\mathbf{L}(\mathbf{x})$ with respect to $\mathbf{x}$ as,

$$\nabla\mathbf{L}(\mathbf{x}) = \sum_k \mathbf{H}_k^T(\mathbf{H}_k\mathbf{x} - \mathbf{y}_k) \tag{2.38}$$

The gradient-based iterative minimization method updates the solution in each iteration as,

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \lambda\nabla\mathbf{L}(\mathbf{x}), \tag{2.39}$$

where λ is a scale factor defining the step size in the direction of the gradient. This technique is also called as *simulate and correct*. In each iteration, given an estimate of $\mathbf{x}$ from previous iteration, the degradation process is simulated. The error between the simulated low-resolution frames and observed low-resolution frames is computed. The error is projected back on the high-resolution

grid. This estimated high-resolution image is corrected from this error. Over iterations we get the desired high-resolution image.

Super-resolution reconstruction is an ill-posed problem. Additional smoothness constraints or regularization measures are introduced in the least square estimate (equation 2.36). For example, combination of Total Variation and Bilateral Filter is used in [34], Tikhonov cost function is used in [30], etc. Zhao [29] used shape from shading framework and synthesized high-resolution image in presence of illumination variations across multiple images. Also they proposed wavelet/pyramid based adjustment for generating super-resolved images for general scenarios.

Limits on Multi-frame Super-Resolution

Baker and Kanade [39] provided theoretical analysis that shows that the reconstruction constraints provide less and less useful information as the magnification factor increases. Any smoothness prior leads to overly smooth results with little high-frequency content. One of the key result is that for square point spread functions and integer magnification factors, the reconstruction constraints are not even invertible. Later Lin and Shum [74] provided the exact theoretical bounds on the magnification factors under local translations. They concluded that under practical scenarios, the magnification limit is 1.6. The first choice if one wants to try a magnification larger than 1.6 is 2.5. The theoretical limit has been found to be 5.7. Also, effective magnification factors can only lie on some disjoint intervals. Smaller magnification limits on multi-frame super-resolution algorithms require all the underlying parameters to be computed as accurate as possible. Image registration is one such underlying factor which has been addressed in detail in chapter 3.

2.4.3 Learning Based Super-Resolution Algorithms

One of the main disadvantage of multi-frame super-resolution algorithms that it require multiple image captures. For dynamic scenes the performance of such algorithm deteriorates. As shown by Zhao and Sawhney [32] that errors from the traditional optical flow algorithm can render the reconstruction infeasible. Smoothness constraints [34, 30] or prior information lack generalizability and are usually hypothesized.

Learning based super-resolution algorithms learn the prior probabilities or constraints from the training data. Application specific priors are learned and high-frequency details are inferred. Usually only a single image is used but the availability of multiple images further improves the quality of a reconstructed image. Freeman *et al.* [37] first provided a general framework for learning low-level vision problems which includes super-resolution. Super-resolution is proposed as an inference problem. High-resolution information is inferred from a low-resolution image in a Markov framework. Several other similar and other single frame super-resolution algorithms include [38, 39, 40, 41].

High-Resolution Image Inference

High-resolution image data is inferred rather than solved. We briefly describe the super-resolution algorithm by Freeman *et al.* [37]. The algorithm is proposed in a Markov network. Let $\mathbf{x}$ be the high-frequency component to be inferred and $\tilde{\mathbf{y}}$ is the mid-frequency component of a low-resolution image $\mathbf{y}$ interpolated to the size of $\mathbf{x}$. Only mid-frequency components of low-resolution image is considered to remove variability of patches in training dataset. Low-frequency components have little influence on high-frequency information prediction. MAP estimation of high-frequency content

is given as,

$$\mathbf{x} = \arg\max_{\mathbf{x}} P(\mathbf{x}|\tilde{\mathbf{y}}) \tag{2.40}$$

$$= \arg\max_{\mathbf{x}} p(\tilde{\mathbf{y}}|\mathbf{x})P(\mathbf{x}) \tag{2.41}$$

As mentioned in section 2.3.3 predicting such an information at image level poses tremendous challenges in training and inference phase. The inference problem is decomposed into a patch based network under Markov assumption. Nodes labeled as x_i in Fig. 2.6 denote a high-frequency patch and nodes labeled as y_i denote a mid-frequency patch. The equation 2.41 is expanded as described in section 2.3.3 and the equation 2.29 is maximized to obtain a super-resolved image.

Training Data Generation

Training patches are generated as follows. A set of high quality high-resolution images are chosen. Each of these images are blurred and down-sampled. The down-sampled images are interpolated to the original magnification factor. Only the mid-frequency information of these images are retained. Images are divided into small square patches of equal sizes. High-resolution and low-resolution patch pairs are stored in the training dataset.

Learning Compatibility Functions

Compatibility functions (equation 2.27 and 2.28) measure the degree of consistency between the predicted patch and the underlying patch, and between a predicted patch and the context. Context here refers to the predicted patches in the neighborhood. In the inference phase, the patches are selected such that they themselves estimate the compatibility function $\Psi(x_j, x_k)$ between neighbors. This is defined in the region of overlap as a function of sum of squared intensity differences between predicted patches in neighbors. Assuming that the scene patches differ from ideal training samples by Gaussian noise, the compatibility matrix is defined as,

$$\Psi(x_k^l, x_j^m) = e^{-\frac{|d_{jk}^l - d_{kj}^m|^2}{2\sigma_s^2}}, \tag{2.42}$$

where x_a^b denote the bth scene candidate at location a from the training data set, d_{jk}^l are the pixels of the lth scene candidate of patch j in the overlap region between patches j and k, and let d_{kj}^m, are the corresponding pixels at location k. This function force the algorithm to have minimum deviation in intensity values in the region of overlap. Another compatibility function is defined between the low-resolution image patch y_k and lth scene candidate x_k^l. Assuming that scene patches differ from ideal training samples by Gaussian noise, the compatibility function is,

$$\Phi(x_k^l, y_k) = e^{-\frac{|x_k^l - y_k|^2}{2\sigma_i^2}}. \tag{2.43}$$

Energy is minimized and high-resolution information is inferred using loopy belief propagation rules defined in section 2.3.3. MRF provides local spatial consistency constraints. However, maximum-likelihood estimate of the high-resolution information may not be consistent throughout.

Limits on Learning Based Super-Resolution Algorithms

Lin *et al.* [75] established limits on learning based super-resolution algorithms for general natural images. The limit is roughly around 10 though it is not a very well defined and strict limit.

The limits are calculated with respect to intensity values that can be recovered within an error limit instead of the actual content and meaningful details that can be super-resolved back. This category of algorithms perform well for natural objects, where the perceptual quality is more important than accurate reconstruction of reality. However, the performance drops significantly on man-made structures where even with a magnification factor of 3 (see Fig. 4(a), in Lin *et al.* [75]), the actual content need not be resolved in the final result. In chapter 4, we look into how much scene information should be captured so that further magnification enhancements can be achieved using off the shelf single frame super-resolution algorithms.

Chapter 3

Accurate Registration for Super-Resolution using Local Phase

3.1 Introduction

Image Registration is a process of geometrically aligning two or more images obtained from different views. In many computer vision and image processing applications, we need highly accurate image registration under differing noise and illumination conditions. For example, in applications such as generation of super-resolution images from multiple images, the output quality depends mostly on the registration accuracy. In large scale mosaicing, a small error in registration of two images can lead to large errors at later stages. In medical image analysis, registration accuracy is required to predict diseases based on the image comparison. Image understanding algorithms such as 3D reconstruction from videos also needs accurate registration. In this chapter, we discuss the problem of accurate image registration, particularly for image super-resolution.

Generating high-resolution images from multiple low-resolution, degraded images has a variety of applications in space imaging, medical imaging, commercial videography, surveillance, etc. Any super-resolution algorithm assumes accurate blur and registration parameters. Most of the existing registration algorithms perform well in presence of uniform illumination across frames as well as limited and uniform blur and noise. However, these conditions are frequently violated in real-world imaging, where specular reflections and strobe lights create large variations in illumination of the scene. Moreover, non-uniform blur often results from depth variations in the scene, while high noise levels are seen in images generated from compact sensors in mobile devices (see Figure 3.1). Interestingly, these are the exact situations, where one would like to employ super-resolution algorithms.

The primary factor that controls the quality of the super-resolved image is the accuracy of registration of the low resolution frames. Park et al. [36] has shown by example that small error in registration can considerably affect the super-resolution results. Most multi-image super resolution algorithms assume that the exact registration parameters between the constituent frames are known. However, as mentioned before, the image artifacts can affect the accuracy of estimation of these parameters. Typically, two characteristics of registration have been considered in the past:

- *Accuracy:* Super-resolution algorithms require extremely precise alignment of the constituent low-resolution frames; accurate to the order of a tenth of a pixel. However, most of these algorithms tend to be sensitive to illumination and blur variations and noise.
- *Robustness:* Registration algorithms that are robust to image artifacts are available and have

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

Figure 3.1: Image degradations: (a) and (b) have spatially varying blur, while (c) and (d) have different illuminations due to use of flash in (c).

been used in application such as registering multi-modal medical and space images. The primary concern of such algorithms is to handle extremely large variations in the image, while being moderately accurate.

The first category of registration algorithms, which work in the spatial or pixel domain, are commonly employed in super-resolution (SR) algorithms due to their accuracy. The most successful ones are RANSAC [33] or gradient descent based [95] methods that minimize the difference between pixel intensity values of the registered images. Robinson et al. [96] also proposed a statistically optimal registration technique based on intensity values. The registration parameters in such approaches converges at incorrect values under image artifacts, specifically, non-uniform illumination. Although the RANSAC-based registration algorithms are robust in presence of outliers but the performance of such algorithms is restricted by the reliability of feature detectors. The reliability of feature detectors drops considerably with increase in image artifacts. The second category of algorithms mentioned above, are meant to deal with large variations between images to be registered such as Sonogram and MRI images of a body part [73]. As noted before the accuracy of such approaches is too low to be considered for super-resolution applications.

A second class of approaches use frequency domain processing to compute the registration parameters. It is well known that these approaches are relatively stable under various image artifacts. However, they are limited in the class of transformations that can be estimated between two images [97]. Further reviews of the registration algorithms for super-resolution can be seen in review papers by Park et al. [36] and Borman and Stevenson [35].

The image formation process used in Super Resolution (SR) reconstruction is given by a linear system,

$$\mathbf{y}_k = \mathbf{L}_k\mathbf{D}_k\mathbf{B}_k\mathbf{F}_k\mathbf{x} + \mathbf{n}_k, \tag{3.1}$$

where $1 \leq k \leq n$, $\mathbf{x}$ and $\mathbf{y}_k$ are the high and low resolution images respectively. The registration parameters are captured by the geometric transformation, $\mathbf{F}_k$, and $\mathbf{B}_k$ is the blurring matrix. The registration algorithms mentioned above, try to estimate the matrix $\mathbf{F}_k$. To accommodate the effect of illumination [29], diagonal matrix $\mathbf{L}_k$ has been introduced into the linear equation. However, [29] only deals with the non-uniform illumination variation during the SR phase, while assuming accurate registration. One of the solutions to deal with registration error is to incorporate this error as noise, while computing the super-resolution image [32]. A second approach is to incorporate the registration phase and the high-resolution image computation phase into a single optimization framework [98]. However, with larger amounts of registration error and outliers, the results will degrade fast, or will not converge in more complex optimizations.

In this chapter, we explore at alternate solution to the problem of robustness in the registration step of a SR algorithm. We formulate the registration as optimization of the local phase alignment

at various spatial frequencies and directions. The local phase in an image has been used for problems such as estimation of stereo disparity [77], and optical flow field estimation [78]. We extend its scope to estimate accurate registration parameters and use it for computing super-resolved images. In this chapter,

- We propose a registration framework using local-phase, which is known to be robust to noise and illumination parameters. The approach is correspondenceless and yields results that are an order of magnitude better compared to the conventional schemes.
- A method for estimating local translation components and estimation of image registration parameters from these estimates is outlined. The results are shown for affine transformation, although it can be extended to any class of image transformations.
- We derive the theoretical error rate of the approach introduced by limitations of Finite Impulse Response (FIR) filters and show that the algorithm converges to the exact registration parameters,
- We show that the algorithm is not sensitive to a large class of blur kernel functions.
- Finally, we present experimental results of SR reconstruction, that demonstrates the advantages of this approach as compared to other popular techniques.

3.2 Homography

Image homography [99] is a special class of image transformation which relates two images of a *planar scene* captured from different viewpoints. The homography transfers points from one view to the other. If $\mathbf{x}$ and $\mathbf{x}'$ are the images of the same world points, represented in the homogeneous coordinate system, then

$$\mathbf{x}' = \mathbf{H}\mathbf{x} \tag{3.2}$$

and

$$\mathbf{H} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \tag{3.3}$$

where $\mathbf{H}$ is a 3×3 homography matrix that maps all points of a scene in the first view to the second view. The homography depends on the intrinsic and extrinsic parameters of the camera and the 3D plane equations. Homography matrix is defined up to scale. Full *projective* transformation has 8 free parameters. *Translation* transformation moves every point by a fixed distance in the same direction. *Rotational transformation* moves all the points by a particular angle about an axis keeping the distance from the axis constant. *Scaling* results in the enlargement or diminishing of an object along the axis. Scaling transformation could involve separate scale factor for each direction. *Shearing* effectively rotates one of the axis so that the axis are no longer parallel. *Similarity transformation* captures translation, rotation and scaling. *Affine transformation* captures translation, rotation, scaling and shearing in a plane.

3.3 Image Registration : Related Work

As mentioned earlier, various image registration algorithms are divided into two categories *viz.spatial domain* and *frequency domain*. Spatial domain algorithms usually involve feature detection and

feature matching. Various image characteristics determine the kind of features to be estimated. Fourier domain techniques are basically build upon phase correlation for translation. Fourier domain techniques have several advantages in terms of its computational speed and the robustness against noise and illumination variations. But they are limited in the class of image registration parameters that can be estimated. In this section, we provide references and a brief overview of key algorithms in both the domains. We also provide an overview of image registration algorithms modified specifically for super-resolution reconstruction. Survey papers by Lisa G. Brown [100] and Zitova and Flusser [73] provide a comprehensive overview of general image registration techniques. Agarwal [101] provides a brief survey on computing image homography.

	Algorithm	**Remarks**
(a)	*Spatial Domain Algorithms*	
•	Feature Extraction and Matching	Distinct features like points, lines or curves are extracted from the image pair. Correspondences are computed between features in images. Typically area or point descriptors around each point in one image is matched with all the points in the other image. Point pairs resulting in the least score are the corresponding points. These estimates are refined further using DLT or RANSAC.
	Transform model estimation	
	• DLT (Direct Linear Transformation) [99]	DLT assumes that the error in spatial location of corresponding points or mismatches follows a Gaussian distribution. Transformation model computed from large number of points produces accurate results.
	• RANSAC (RANdom SAmple Consensus) [102]	Mismatches or the error in actual spatial location of the corresponding points need not follow a Gaussian distribution. In RANSAC, after computing the correspondences, a small set of points are selected randomly. Homography is computed from these points. All correspondences are divided into two parts, inliers and outliers, based on how well they satisfy the current homography matrix. Homography matrix is recomputed from the inliers and the process is repeated till convergence.
•	Intensity Minimization using Gradient Descent [95]	One of the images is taken as a reference image and the co-ordinate of the other image is specified as a function of transformation parameters. Sum of squared differences of intensity values is minimized between two images and the transformation parameters are estimated. The registration is highly accurate except in presence of illumination variations, where accuracy go down significantly.
(b)	*Frequency Domain Algorithms*	

•	Phase correlation for translation parameters [103, 104].	Classical phase correlation techniques [103] register the translated image pair at pixel level. Foroosh *et al.* [104] extended the idea of phase correlation to calculate highly accurate parameters at sub-pixel level.
•	Reddy and Chatterjee [105] for similarity transformation parameters	They extended basic idea of phase correlation for translation to image pairs related by rotation and scaling as well. The algorithm is robust. But to compute the scaling parameters the co-ordinates of the Fourier transformed image is transferred into logarithm domain. Scaling parameters computed in logarithm domain produces inaccurate results because a) non-linear scaling of the co-ordinate system, b) a small error in the logarithm domain is a large error in the primary domain.
•	Fourier Mellin Transformation [106] for similarity transformation parameters	The scale invariance property of Mellin transform is analogous to the Fourier transform's shift invariance property. A scale change in spatial domain is equivalent to phase change in the Mellin domain.
•	Line features in Fourier Magnitude [107] for affine transformation parameters	This method depends on the presence of line features in image texture. After correction in Fourier domain the image is converted into logarithm coordinates to compute the scaling parameters. This algorithm also suffers from the deficiencies of logarithm domain i.e. non-linear scaling and small error in logarithm domain is a large error in the primary domain.
(c)	*Registration Algorithms specifically designed for SR*	
•	Joint high-resolution image estimation and registration [98].	High-resolution image and registration parameters are estimated iteratively. With larger amounts of registration errors and outliers, the result degrade fast, or will not converge in a more complex optimization. The method is dependent on initial guesses.
•	Registration using the concept of variable projection [96].	Using concepts of variable projections a joint registration/reconstruction is performed. The method is more accurate than previous approaches. Algorithm uses intensity information and hence prone to wrong registration parameters in presence of non-uniform illumination.
•	Incorporating error in image registration in SR phase [108].	Super-resolution reconstruction algorithm is modified to incorporate error in image registration. However, high degree of errors can not be neglected and can significantly degrade the quality of the reconstructed image.
•	Dropping frames with large registration errors [109].	More number of frames need to be captured.

3.4 Local Phase

Accurate registration can be achieved with the exact knowledge of degradation parameters such as blur and non-uniform illumination. However, in practice, this information is rarely available. We overcome this problem by using local phase to estimate registration parameters. Local phase is robust towards noise and smoothly varying illumination [84]. We prove the invariability of local phase information to a class of blur kernels. Due to these characteristics, our registration algorithm can easily by-pass these image artifacts, which are difficult to estimate accurately.

Local phase can be computed using any FIR band pass filter. The phase, as opposed to magnitude of the filter response, is robust [84] to Gaussian white noise. Existing registration algorithms routinely achieve up to pixel-level accuracies. However, for finer registration, features should be calculated with sub-pixel accuracy, even under various image artifacts. Local phase based registration can achieve this without explicit signal reconstruction, sub-pixel feature detection or correspondence computation. Local phase has been effectively used to solve similar problems such as stereo disparity computation [77] and optical flow [78] for noisy images. Sanger [77] used it to compute stereo disparity which is the difference in local phase in two images divided by the frequency of the signal. Gautama and Hulle [78] tracked constant phase in subsequent images to calculate optical flow.

Gabor filters are popular band pass filters as they achieve the theoretical minimum product of spatial width and bandwidth, desirable for better localization and accurate phase computation, respectively. Mathematically, a Gabor filter is a multiplication of a complex harmonic function with a Gaussian envelope [85],

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-(\frac{x^2}{2\sigma_x^2} + \frac{y^2}{2\sigma_y^2})} e^{j(\omega_x x + \omega_y y)}, \tag{3.4}$$

where, (ω_x, ω_y) is the angular frequency of the filter, σ_x and σ_y controls the spatial width of the filter, and j is $\sqrt{-1}$. Local phase is computed at angular frequency (ω_x, ω_y) at each pixel location by convolving the image with Gabor wavelet $g(x,y)$. The argument of the complex output is local phase. In our algorithm, we have assumed that the image pair to be registered are partially overlapping, so that the local phase information almost remains the same in two images. Phase difference is computed by taking the difference of phase values at each location of the image pair at the given angular frequency. A detailed description of local phase is already provided in chapter 2.

Confidence measurements: Errors could be introduced in phase difference computation due to noise and the absence of the local frequencies with which the images are convolved. Sanger [77] has described the degree of match in the amplitude values as a confidence measure. The value of confidence is high if the amplitudes of the Gabor filter response at (x,y) in both the images are close. In addition, if the amplitude falls below a particular threshold, the confidence value is set to zero. Let $|s_1|$ and $|s_2|$ be the amplitudes of the Gabor filter response. The confidence value is computed as:

$$r = min\left[\frac{|s_1|}{|s_2|}, \frac{|s_2|}{|s_1|}\right] \tag{3.5}$$

3.5 Local Phase Based Image Registration Algorithm

Our local phase based registration algorithm is robust to noise, illumination, blur and sub-sampling. We convolve the partially overlapping images with Gabor filters at multiple frequencies. The reason of convolving with multiple frequencies is that in case a particular frequency is not present

at the corresponding locations then, that observation could be pruned. The local translation parameters are computed at each spatial location from the robust phase difference estimation. An overdetermined system of equation is formed and from these estimates the registration parameters are computed. The transformation parameters are updated iteratively so that errors due to uncertainty in the frequency estimation of the band-pass filter is minimized. For our algorithm, we define partially overlapping images as the image pair where in any small 2D window at location (x, y) the corresponding point lie within the cycle of the sinusoid. This condition should hold true at most of the image locations, for our algorithm to converge.

3.5.1 2D Local Translation

In the 1D case, the shift between two sinusoids of the same frequency is estimated by measuring the phase difference at the same spatial location and then dividing it by the frequency of the signal (see Fig. 3.2).

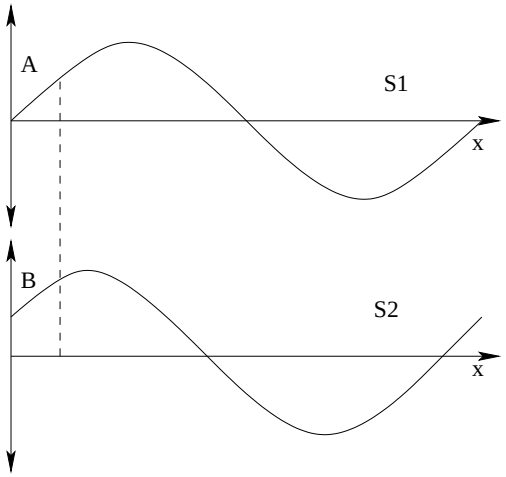


Figure 3.2: Computation of shift from two 1D signals as (*phase difference/frequency*) of the signal

The computation of translation components can be formulated on the basis of Fourier Shift theorem, according to which, a shift of Δx in the spatial domain would produce a phase difference of $\Delta x \omega_x$ at ω_x. This is extended in 2D as, a shift of $(\Delta x, \Delta y)$ in the spatial domain would produce a phase difference of $(\Delta x \omega_x + \Delta y \omega_y)$ i.e., if

$$i_2(x, y) = i_1(x + \Delta x, y + \Delta y), \tag{3.6}$$

then in Fourier domain at (ω_x, ω_y) the relationship is given by:

$$I_2(\omega_x, \omega_y) = I_1(\omega_x, \omega_y) e^{j(\omega_x \Delta x + \omega_y \Delta y)} \tag{3.7}$$

As local phase is computed using a non-ideal bandpass filter, the above relationship holds true within a certain error. This issue has been addressed in detail in the next section. By computing the phase difference at least at two different angular frequency pairs we can estimate $(\Delta x, \Delta y)$. Choosing two different angular frequency pairs is slightly tricky. Not every combination leads to a stable and accurate solution. We consider two different scenarios:

- **Frequency pairs with arbitrary values:** Multiple frequency pairs are used to solve for local translation parameters from phase difference values ($\Delta\phi = \Delta x\omega_x + \Delta y\omega_y$). Through experiments we noticed that the error increased with increase in noise values because: a) solving equations in two variables is very sensitive if both the frequency pairs are close. Also phase difference values need not be correct because of non-ideal band-pass behavior of the filter, b) In case, if one of the frequency component is absent, it would lead to wrong calculations altogether.

- **Frequency pairs with exactly one of them being zero:** Among various combinations $(\omega_x, 0)$, $(0, \omega_y)$ are more suitable as it does not involve solving two equations simultaneously. Computing local translation parameters along the axes involve the division of phase difference with the frequency. Gabor filters acts like a low-pass filter in its orthogonal direction, reducing the effect of noise considerably. Also, the inclusion of confidence parameters in the final computation is straightforward.

The phase difference is computed at multiple frequency pairs in each dimension and is combined by taking the average of estimates weighted by the confidence values. A pixel is removed from consideration for computing the registration parameters if there is not sufficient response of Gabor filters at all frequencies. This approach is correspondenceless. The local translation parameters thus estimated are accurate at sub-pixel level and computation from multiple frequencies make the estimation robust.

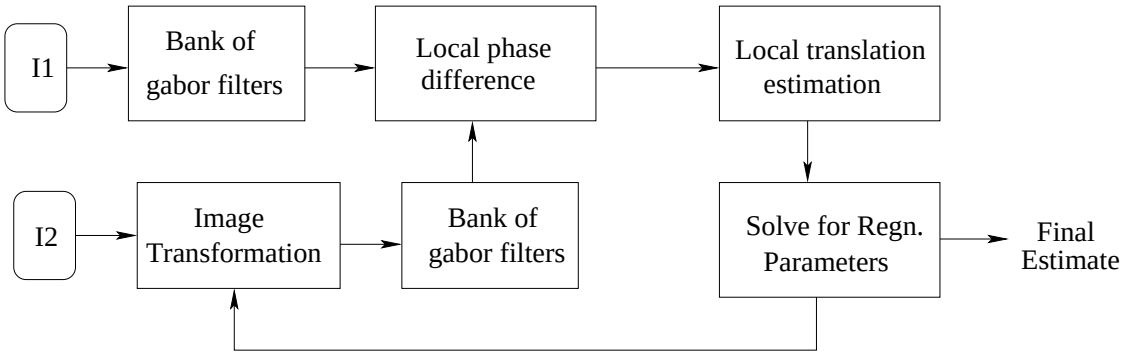


Figure 3.3: Block diagram showing different steps of the registration algorithm.

3.5.2 Frequency Selection at each Iteration

From the phase of the convolution product, as given by equation 3.9, the observation is that for a constant spatial window width, local phase is more accurate at higher frequency. But at higher frequencies the domain of convergence decreases. The frequency of the band-pass filter is changed from low to high as the algorithm converges. At each iteration, various angular frequencies of Gabor filters are selected such that they are close.

3.5.3 Registration Parameters

Local translation parameters thus computed at various spatial image locations can be thought as point correspondences with high accuracy. Given many such corresponding pairs, the image transformation parameters can be estimated by solving an overdetermined system of equations. This framework allows calculation of any class of transformation parameters. For our experiments,

we limit the class of registration algorithms to that of planar views related by affine transformation. We concentrate on affine transformation because most of the partial overlap can be approximated by affine transformations. An affine transformation is a linear transformation in in-homogeneous coordinates followed by a translation and captures translation ,rotation, scaling and shearing in a plane. Mathematically, under affine transformation two views of an object are related by

$$x = ax' + by' + c$$
$$y = dx' + ey' + f$$

At each location, we estimate the translation parameters, which is related to the correspondence of a point (x, y) in one image with (x', y') in the other. We form an overdetermined system of equations in $(a, b, c, d, eandf)$ and estimate the accurate registration parameters.

The local translation parameters, calculated at each spatial location, are approximately correct. This is because in a small window points need not be related by pure translation. Moreover, the two points need not lie within the cycle of the signal. However, over iterations, as the corresponding points come closer, the effect due to these assumptions would be negligible. We iteratively update the transformation parameters till convergence. The overall algorithm is presented below,

Algorithm 1 Local phase based registration algorithm

Input: An image pair.

Output: Parameter describing the geometric relationship between two images accurately.

1: Compute the approximate registration parameters using traditional approaches.
2: **repeat**
3: Obtain the overlapping image pair using the current registration parameters.
4: Convolve both the images with a bank of Gabor filters and calculate the phase difference values.
5: Calculate the translation parameters at each location by solving for Δx and Δy from phase differences with sufficient confidence.
6: Form an over-determined system of equations using the translation estimates and solve it to update the registration parameters.
7: **until** convergence

3.6 Convergence, Error and Robustness Analysis

Any Super-Resolution algorithm is highly dependent on the accuracy of underlying registration algorithm. Noise, blur and illumination affects the accuracy of any registration algorithm. We analyze the performance of registration algorithm that uses local phase under these artifacts. Analysis is performed for 1D signals but extensions to 2D is simple and in the same direction. We consider a 1D Gabor filter, $g(x)$, with angular frequency ω_0 and a sinusoidal signal given by,

$$i(x) = \cos((\omega_0 + \Delta\omega)(x + t)), \tag{3.8}$$

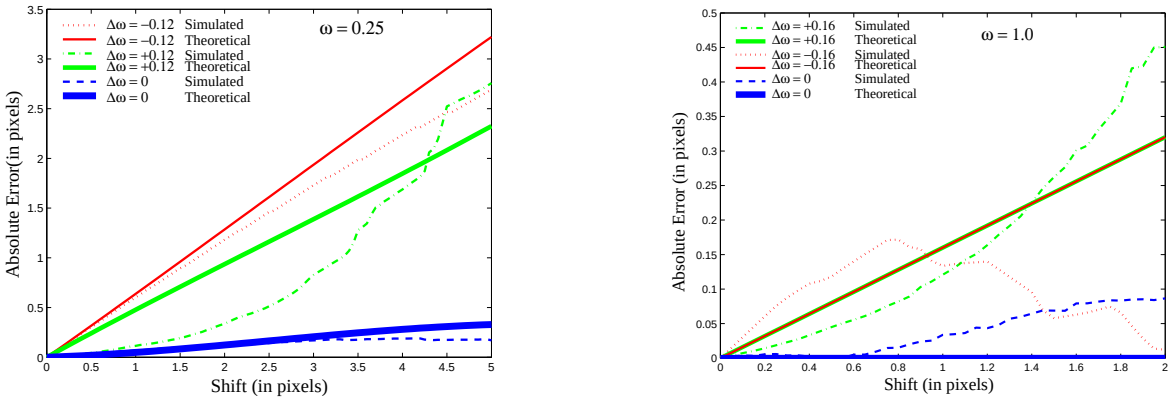
where $\Delta\omega$ at cut-off frequency is the half band-width of the filter and captures the non-ideal band pass behavior of the filter, t is the initial shift. The convergence and error bound is computed by analyzing the sinusoids at the cut-off frequencies of the Gabor filter.

3.6.1 Non-Ideal Band Pass Behavior of the Gabor Filter

Gabor Filter has the minimal value for the product of spatial width and bandwidth, which is a constant. Hence, there is a trade-off in selecting their sizes. Smaller spatial width does help in better localization, but at the cost of non-zero bandwidth. We convolve the Gabor filter, $g(x)$, with the sinusoid (equation 3.8). The phase of the convolution product is (see derivation),

$$\phi_t(x)=\tan^{-1}\left[\tan((\omega_0+\Delta\omega)(t+x))\left(\frac{1-e^{-2(\omega_0^2+\omega_0\Delta\omega)\sigma^2}}{1+e^{-2(\omega_0^2+\omega_0\Delta\omega)\sigma^2}}\right)\right], \tag{3.9}$$

From the above equation, it is easy to see that at infinite width or at very high frequency the local phase computed is accurate. But the domain of convergence decreases at very high frequencies. To show the convergence of a phase based registration algorithm, we only show that the local translation parameters are computed accurately over iterations at cut-off frequencies of the Gabor filter (cut-off frequency is calculated using equation 2.8). The error is calculated for each value of shift as the absolute difference between the actual shift and the shift computed using ($d = (\phi_2 - \phi_1)/\omega$) ($\phi_2$ and ϕ_1 are the local phase computed using equation 3.9 and we assume that only one sinusoid is present). This theoretical error rate is plotted against the simulated convolutions. Simulated scenario is generated by plotting a 1D sinusoid of the same frequency. The sinusoid is quantized and sampled on a grid after its magnitude is scaled by 128. From the error graphs (Fig. 3.4), we conclude that the error drops to zero over iterations. Note that even for ideal band-pass filter ($\Delta\omega = 0$) the error is not zero at low frequency.



(a) $\omega_0 = 0.25$,$\sigma = 5$ (b) $\omega_0 = 1.0$, $\sigma = 4$

Figure 3.4: Error in shift calculation due to non-ideal bandpass filter at various pixel locations (a) $\omega_0 = 0.25$ and $\sigma = 5$ (b) $\omega_0 = 1.0$ and $\sigma = 4$. Solid lines show the theoretical behavior as given by equation 3.9 and dotted lines show the behavior of the simulations on 1D sinusoids which are quantized after scaling by a factor of 128

Calculation of Phase (equation 3.9)

We rewrite $i(x)$ in Euler form as,

$$i(x) = \frac{e^{j(\omega_0+\Delta\omega)x}e^{j(\omega_0+\Delta\omega)t} + e^{-j(\omega_0+\Delta\omega)x}e^{-j(\omega_0+\Delta\omega)t}}{2} \tag{3.10}$$

The convolution product with the Gabor filter is

$$\begin{aligned} r(x) &= i(x) * g(x) \\ &= \frac{1}{2}\int_{-\infty}^{+\infty} (e^{j(\omega_0+\Delta\omega)x'}e^{j(\omega_0+\Delta\omega)t} + e^{-j(\omega_0+\Delta\omega)x'}e^{-j(\omega_0+\Delta\omega)t}) \\ &\qquad .e^{-\frac{(x-x')^2}{2\sigma^2}}e^{j\omega_0(x-x')}dx' \\ &= \frac{1}{2}\left(e^{j(\omega_0+\Delta\omega)t}\int_{-\infty}^{+\infty} e^{-\frac{(x-x')^2}{2\sigma^2}}e^{j\omega_0(x-x')+j(\omega_0+\Delta\omega)x'}dx'\right) \\ &\quad + \frac{1}{2}\left(e^{-j(\omega_0+\Delta\omega)t}\int_{-\infty}^{+\infty} e^{-\frac{(x-x')^2}{2\sigma^2}}e^{j\omega_0(x-x')-j(\omega_0+\Delta\omega)x'}dx'\right) \\ &= \frac{1}{2}\left(e^{j(\omega_0+\Delta\omega)t}a(x) + e^{-j(\omega_0+\Delta\omega)t}b(x)\right), \end{aligned} \tag{3.11}$$

where

$$\begin{aligned} a(x) &= \int_{-\infty}^{+\infty} e^{-\frac{(x-x')^2}{2\sigma^2}}e^{j\omega_0(x-x')+j(\omega_0+\Delta\omega)x'}dx' \\ &= \int_{-\infty}^{+\infty} e^{-\frac{(x-x')^2}{2\sigma^2}}e^{j\omega_0(x-x')-j(\omega_0+\Delta\omega)(x-x')+j(\omega_0+\Delta\omega)x}dx' \\ &= e^{j(\omega_0+\Delta\omega)x}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}((x-x')^2+j\Delta\omega(x-x')2\sigma^2+(j\Delta\omega\sigma^2)^2-(j\Delta\omega\sigma^2)^2)}dx' \\ &= e^{j(\omega_0+\Delta\omega)x}e^{-\frac{1}{2}(\Delta\omega\sigma)^2}\int_{-\infty}^{+\infty} e^{-\frac{(x-x'+j\Delta\omega\sigma^2)^2}{2\sigma^2}}dx' \\ &= \sigma\sqrt{2\pi}e^{j(\omega_0+\Delta\omega)x}e^{-\frac{1}{2}(\Delta\omega\sigma)^2} \qquad \left(\because \int_{-\infty}^{+\infty} e^{-\frac{(x'-x)^2}{c^2}}dx = c\sqrt{\pi}\right) \end{aligned} \tag{3.12}$$

and

$$\begin{aligned} b(x) &= \int_{-\infty}^{+\infty} e^{-\frac{(x-x')^2}{2\sigma^2}}e^{j\omega_0(x-x')-j(\omega_0+\Delta\omega)x'}dx' \\ &= \int_{-\infty}^{+\infty} e^{-\frac{(x-x')^2}{2\sigma^2}}e^{j\omega_0(x-x')+j(\omega_0+\Delta\omega)(x-x')-j(\omega_0+\Delta\omega)x}dx' \\ &= e^{-j(\omega_0+\Delta\omega)x}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}((x-x')^2+j(2\omega_0+\Delta\omega)(x-x')2\sigma^2+(j(2\omega_0+\Delta\omega)\sigma)^2-(j(2\omega_0+\Delta\omega)\sigma)^2)}dx' \\ &= e^{-j(\omega_0+\Delta\omega)x}e^{-\frac{1}{2}(2\omega_0+\Delta\omega)^2\sigma^2}\int_{-\infty}^{+\infty} e^{-\frac{(x-x'-j(2\omega+\Delta\omega\sigma)^2}{2\sigma^2}}dx' \\ &= \sigma\sqrt{2\pi}e^{-j(\omega_0+\Delta\omega)x}e^{-\frac{1}{2}(2\omega_0+\Delta\omega)^2\sigma^2} \qquad \left(\because \int_{-\infty}^{+\infty} e^{-\frac{(x'-x)^2}{c^2}}dx = c\sqrt{\pi}\right) \end{aligned} \tag{3.13}$$

After substitution 3.12 and 3.13 into 3.11 we get,

$$r(x) = \sigma\sqrt{\frac{\pi}{2}}\left(e^{j(\omega_0+\Delta\omega)(x+t)}e^{-\frac{1}{2}(\Delta\omega\sigma)^2} + e^{-j(\omega_0+\Delta\omega)(x+t)}e^{-\frac{1}{2}(2\omega_0+\Delta\omega)^2\sigma^2}\right)$$

Let $\theta = (\omega_0 + \Delta\omega)(x+t)$, $A = e^{-\frac{1}{2}(\Delta\omega\sigma)^2}$ and $B = e^{-\frac{1}{2}(2\omega_0+\Delta\omega)^2\sigma^2}$

$$\begin{aligned} r(x) &= \sigma\sqrt{\frac{\pi}{2}}\left(Ae^{j\theta} + Be^{-j\theta}\right) \\ &= \sigma\sqrt{\frac{\pi}{2}}\left(A(\cos\theta + j\sin\theta) + B(\cos\theta - j\sin\theta)\right) \\ &= \sigma\sqrt{\frac{\pi}{2}}\left((A+B)\cos\theta + j(A-B)\sin\theta\right) \end{aligned} \tag{3.14}$$

Phase of the response $r(x)$, after substitution and simplification

$$\begin{aligned} \phi(x) &= \tan^{-1}\left[\tan\left((\omega_0+\Delta\omega)(t+x)\right)\left(\frac{e^{-\frac{1}{2}(\Delta\omega\sigma)^2} - e^{-\frac{1}{2}((2\omega_0+\Delta\omega)\sigma)^2}}{e^{-\frac{1}{2}(\Delta\omega\sigma)^2} + e^{-\frac{1}{2}((2\omega_0+\Delta\omega)\sigma)^2}}\right)\right] \\ &= \tan^{-1}\left[\tan\left((\omega_0+\Delta\omega)(t+x)\right)\left(\frac{1 - e^{-2(\omega_0^2+\omega_0\Delta\omega)\sigma^2}}{1 + e^{-2(\omega_0^2+\omega_0\Delta\omega)\sigma^2}}\right)\right] \end{aligned} \tag{3.15}$$

3.6.2 Blur

Accurate computation of spatial features are difficult in presence of blur. Blur parameters are difficult to obtain in many real applications. Given a sinusoid, $i(x)$ (equation 3.8), and an even and real blur kernel, $b(x)$, the local phase is independent of all parameter of blur kernel. Except that the magnitude is scaled . Two images can be compared by local phase information in presence of blur. However, higher frequency information is degraded because of sampling on a grid. The blur parameters varies due to variation in depth and for a planar scene, and the variation is smooth. It can safely be assumed that in a small window the blur parameters are constant.

Blur Invariance

Local phase is independent of all blur kernels that are real and even. Let $b(x)$ be such a blur kernel. Let $i(x)$ be the sinusoid given by the equation 3.10. Let $B(\omega)$ be the Fourier transformation of $b(x)$ and $I(\omega)$ is the Fourier transformation of $i(x)$ which can be shown to be:

$$I(\omega) = e^{j\omega t}(\pi\delta(w - (\omega_0+\Delta\omega)) + \pi\delta(w + (\omega_0+\Delta\omega))) \tag{3.16}$$

Using the fact that if $b(x)$ is real and even then $B(\omega)$ will be real and even [81] and the convolution in spatial domain is equivalent to multiplication in frequency domain, we obtain:

$$\begin{aligned} R(\omega) &= B(\omega)I(\omega) \\ &= e^{j\omega t}(\pi B(\omega_0+\Delta\omega)\delta(w - (\omega_0+\Delta\omega)) + B(-(\omega_0+\Delta\omega))\pi\delta(w + (\omega_0+\Delta\omega))) \\ &= B(\omega_0+\Delta\omega)e^{j\omega t}(\pi\delta(w - (\omega_0+\Delta\omega)) + \pi\delta(w + (\omega_0+\Delta\omega))) \end{aligned} \tag{3.17}$$

The inverse Fourier transformation of the above equation is given by

$$r(x) = B(\omega_0+\Delta\omega)\cos((\omega_0+\Delta\omega)(x+t)) \tag{3.18}$$

Going by same steps as for derivation of equation 3.15, it is evident that the phase information will remain invariant but the amplitude will get multiplied by $B(\omega_0+\Delta\omega)$.

For the special case of Gaussian blur kernel, $b(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}$, the phase information remains invariant to the blur kernel, but the amplitude gets multiplied by $e^{-\frac{((\omega_0+\Delta\omega)\sigma)^2}{2}}$.

3.6.3 Illumination

Illumination change, in the image space, is the multiplication of a pixel value by another value. Smooth illumination can be modeled by the multiplication of a constant in a window. The phase information computed at these two locations will remain unchanged as compared to the magnitude of the signal, which will be scaled by the illumination constant.

3.6.4 Noise and Quantization Errors

Fleet and Jepson [84] has shown that the phase is more robust for image matching than the amplitude of the filter response in presence of noise. Quantization errors can also be modeled as noise. The quantization error results from the mapping of irradiance field onto integer grid. For band-limited noise, the error in the estimation is reduced by considering the phase output of those filters that do not allow those frequencies to pass through. This is done by assigning low scores to those phase difference estimates where there is a significant amplitude mismatch in both the signals detected.

3.7 Experiments and Results

3.7.1 Performance Metric

To compare the performance of the algorithms, we generate the image pairs having known transformation parameters from synthetic and real images. We define the mean shift error, ϵ_{ms}, as the average distance between corresponding points after registration. Mathematically, it can be expressed as:

$$\epsilon_{ms} = \frac{1}{N} \sum_{(x,y)} d(\mathbf{H'}[x\ y\ 1]^T, \mathbf{H}[x\ y\ 1]^T), \tag{3.19}$$

where $[x\ y\ 1]$ is the homogeneous representation of the image coordinate of image $i_1(x, y)$, $\mathbf{H}$ and $\mathbf{H'}$ are 3×3 transformation matrices representing the ideal and estimated image registration parameters respectively, N is the total number of points and $d(,)$ is the Euclidean distance between two spatial points. The mean shift error indicates how close the overlay is.

We perform experiments on synthetic and real images. As discussed in section 3.5.1, not every combination of frequency pairs lead to stable equations. We experimentally find that choosing frequency combinations of form $(\omega_x, 0)$ and $(0, \omega_y)$ is more stable as compared to pairs having arbitrary values to calculate the local translation parameters.

3.7.2 By Choosing Arbitrary Frequency Pairs for Gabor Filters

On real images, we test and compare the performance of our algorithm on images corrupted by Gaussian white noise with RANSAC [102], Fourier-Mellin Transform (FMT) [106] and an algorithm that minimizes the sum of squared differences of intensity values between two images using gradient descent (GD [95]). RANSAC is robust in estimating the transformation parameters in the presence of outliers. Fourier-Mellin Transform is robust to noise and varying illumination conditions, though it can estimate only up to similarity transformation. In the absence of illumination variations in the images, an image registration algorithm based on minimization of the intensity value differences can register the images accurately.

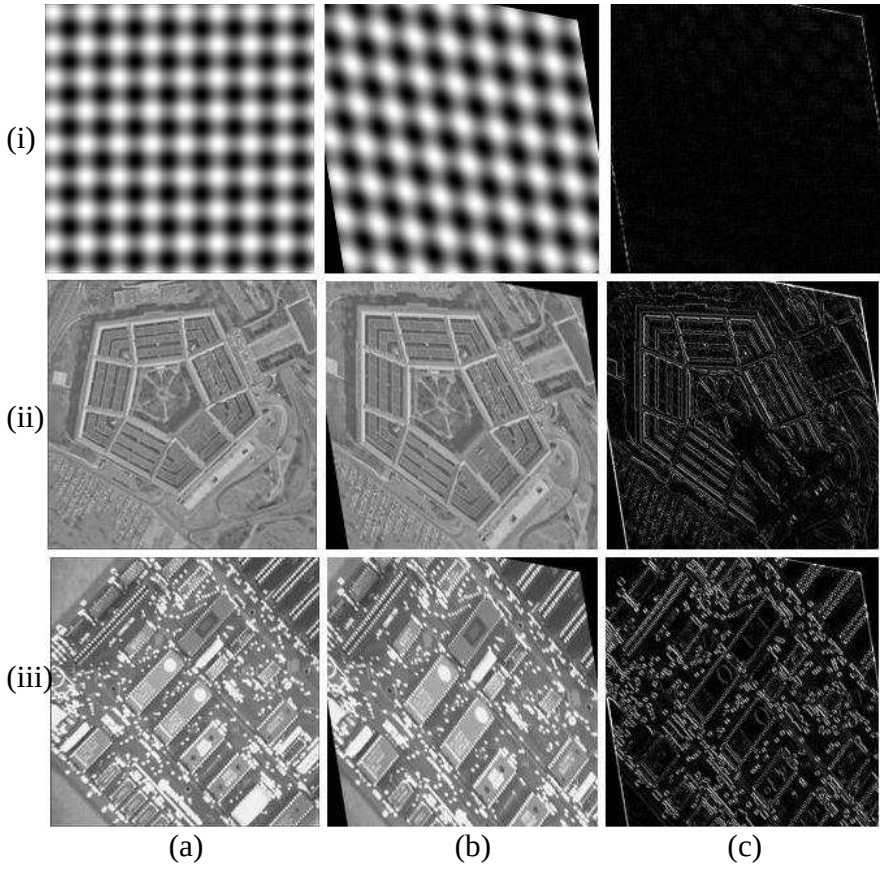


Figure 3.5: (a) and (b) are the images to be registered related by affine transformation. (c) shows the absolute image difference after using our algorithm

We first experimentally show that our algorithm produces highly accurate registration in the absence of any artifacts.

- **Sinusoidal Image:** The first row of Fig. 3.5 show two synthetic 2D sinusoidal image pairs that are to be registered. These sinusoidal images has very limited frequency components. Hence, accurate registration parameters can be computed with minimum effort. We use four Gabor filters having central angular frequencies in the range 0.37 and 0.75. The algorithm converged in 4 iterations with a final mean shift error of 0.08 per pixel. Fig. 3.5(c) show the absolute image difference after using our algorithm between the reference and the registered images.
- **Textured Images:** The second and third rows of Fig. 3.5 show the pentagon and apple chip image pairs. Multiple frequencies are present in these images. Experimental steps are same as mentioned for sinusoidal image pairs. The algorithm converged in 7 steps and the mean of shift error after using our algorithm was 0.108.

Image Corrupted by Gaussian White Noise

We add Gaussian white noise with zero mean and standard deviation varying from 0 to 6 in both the images. For image pairs related by an affine transformation, we compared our proposed approach

	Affine			Similarity	
σ	Proposed	RANSAC	GD	Proposed	FMT
0	0.10	0.18	0.068	0.06	1.19
1	0.18	0.35	0.070	0.11	1.19
2	0.31	0.41	0.071	0.18	1.19
3	0.46	1.09	0.073	0.29	1.19
4	0.52	0.76	0.073	0.33	1.19
5	0.59	0.80	0.075	0.38	1.19
6	0.65	1.05	0.078	0.45	1.19

(a) Errors on Pentagon Image Pair

	Affine			Similarity	
σ	Proposed	RANSAC	GD	Proposed	FMT
0	0.15	0.22	0.121	0.09	1.19
1	0.21	0.35	0.123	0.14	1.19
2	0.27	0.46	0.123	0.18	1.19
3	0.32	0.62	0.126	0.26	1.19
4	0.45	0.80	0.129	0.29	1.19
5	0.51	0.81	0.133	0.36	1.19
6	0.59	0.97	0.132	0.41	1.19

(b) Errors on Apple Chip Image Pair

Table 3.2: Comparison of the proposed scheme with other image registration algorithms under Gaussian white noise.

with RANSAC and the gradient descent based image registration algorithm(GD). For images related by similarity transformation we compare the performance of our algorithm with Fourier Mellin Transformation (FMT). Table 3.2 summarizes the mean of shift error values computed.

Discussions

For the images related by affine transformation, our algorithm performs better than RANSAC in presence of Gaussian white noise. However, in presence of Gaussian white noise, GD registered the images more accurately. Note that Gaussian white noise has zero mean and hence does not affect the minimum of mean squared error. For the images related by similarity transformation, our algorithm performs far better than FMT within the given noise limits. FMT is very robust to noise and illumination conditions, but the accuracy is limited because of the detection of impulse response at non-integer locations and the factors involving coordinate transformations.

In short, we note that the proposed algorithm can handle differing noise, whereas existing approaches fails to perform well. Transform domain techniques are robust to these variations, but the class of image transformation that can be calculated are very limited. The proposed algorithm can estimate the local translation parameters as long as there are sufficient locations, where the corresponding points lie within a cycle of the signal. Increase in error in image transform parameters can be attributed to solving linear equations in two variables. The equations computed need not be highly accurate because of non-ideal bandpass behavior of the filter and the potential absence of one of the frequency component.

	Bicubic	Ideal Reg.	RANSAC		GD		Proposed	
σ	ϵ_{re}	ϵ_{re}	ϵ_{ms}	ϵ_{re}	ϵ_{ms}	ϵ_{re}	ϵ_{ms}	ϵ_{re}
1	18.279	6.867	0.563	14.712	0.189	9.228	0.195	9.388
2	18.338	7.276	0.767	15.567	0.189	9.829	0.201	10.069
3	18.426	8.079	0.787	16.098	0.187	10.208	0.216	10.437
4	18.551	8.394	0.829	17.112	0.187	10.721	0.221	11.113
5	18.717	8.974	0.648	17.842	0.186	11.173	0.223	11.673
6	18.907	9.476	0.916	19.005	0.191	11.865	0.223	12.272

Table 3.3: Comparison of the proposed scheme with other image registration algorithms under Gaussian white noise (with 0 mean and standard deviation, σ, from 1-6). (Ideal) denotes the error when actual registration parameters are given as input for SR reconstruction.

3.7.3 By Choosing Frequency Pairs with Exactly One of them Zero for Gabor Filters

By choosing various frequency pairs of type $(\omega_x, 0)$ and $(0, \omega_y)$ to compute the local translation component along horizontal direction and vertical directions respectively has several advantages as discussed before. We experimentally found that the error in image registration almost remains constant with increasing noise values which was not the case previously. Super-resolved images have been constructed to show that accurate image registration results in high quality images.

Experiments have been performed on synthetic low-resolution frames and on more challenging real-life images captured using a mobile phone camera (Nokia 3320). The results of our algorithm has been compared with registration algorithm which is based on minimization of mean squared intensity differences [95](GD) using gradient descent and RANSAC [33]. Intensity minimization algorithm has been chosen because it is very accurate in presence of noise and widely used. RANSAC is robust in presence of outliers. Comparison with Fourier domain methods have been ignored in this section as we have already seen that due to various reasons it provides highly accurate transformation parameters only up to pixel level and classes of image transformation parameters that can be estimated are limited. We use the algorithm mentioned in [34] for super-resolution reconstruction without any regularization term. For synthetic data-sets the registration algorithms have been compared by using absolute mean shift error (ϵ_{ms}). Super-Resolved image has been compared with the ground truth using root mean square reconstruction error (ϵ_{re}) on intensity values.

We generated three different kind of data sets of low resolution(LR) frames, which are related by affine transform, from a single high resolution(HR) frame. In first data-set, Gaussian noise of various levels were added all under same Gaussian blur of window size 4 and variance 2. In second data set, smooth spatially varying blur with window size smoothly varying between 5 and 9 in different directions was added in each LR frame. In third data-set, noise having variance 3 and uniform Gaussian blur was added. Non-uniform illumination was synthetically generated which degrades radially from a randomly selected point source for each of the LR frames. Table 3.3 summarizes the result for noisy data-set. For the second data-set the absolute mean shift error was 0.379, 0.674, 0.285 for GD, RANSAC and our algorithm respectively. For the third data set, where each low resolution frame has different kind of illumination variations the registration error was 5.849, 1.391 and 0.210 respectively. Images were magnified by a factor of 1.8 in all the cases. Fig. 3.6 shows the super-resolved output of frames registered under varying illumination conditions.

To show the effectiveness of of our algorithm on real world scenes, we captured the video of facade of Charminar. Some frames are blurry due to lens blur. There is a significant amount of

noise present in all the images and a small variation in illumination condition across frames. Eight arbitrary frames were chosen from the video. Fig. 3.7 summarizes the super-resolution result. Images registered using intensity minimization algorithm did not produce high quality images because of slight variations in illumination conditions across frames. Registration results of our proposed algorithm is slightly better than RANSAC. Because of good number of features present in images, RANSAC performed well.

More challenging real-world video was taken using a mobile phone camera (Nokia 3320). (Video was taken such that all the LR frames are related by affine transformation only by keeping the camera in one plane). Different part of the scene was illuminated during the video capture by using a flashlight. 8 frames(each of size 128×96) were selected out of the video for scene super-resolution. Compression artifacts are clearly visible in all LR frames. Fig. 3.8 summarizes the result. The magnification factor was 2.2. Our algorithm has performed significantly better than the other image registration algorithms. The main reason is lack of features in such a small and heavily degraded image and strong illumination artifacts.

Discussions

Our algorithm performs better than RANSAC for all 3 synthetic data sets and is comparable to intensity minimization algorithm (GD) under Gaussian white noise. The optimization framework for GD can be shown to be independent of Gaussian white noise, and hence its performance is marginally better than the proposed one. However, our algorithm clearly outperforms GD in presence of non-uniform illumination and non-uniform blur. SR applied on the images taken from the video of the mobile phone camera shows the robustness and practical applicability of our algorithm. The compared algorithm fails miserably in such cases due to the lack of feature points, compression artifacts, small size of the image, and high level of degradations. Moreover, our algorithm is correspondence-less and does not require the calculation of feature points. Experiments were performed for images related by an affine transformation. However, our algorithm is easily extensible to a general class of image transformations. The computation of local phase requires a minimum amount of texture in the image. As we need not compute exact feature locations, the absolute intensity values need not be preserved across the image, and hence can deal with varying blur and illumination. Existing transformed domain techniques are more robust to these artifacts. But they solve a very small class of image transformations. Our algorithm can estimate the local translation accurately, given that the corresponding points lie within the cycle of the signal (8-10 pixels apart in practice). By quick registration using any existing image registration algorithm we can overcome this limitation. As phase difference and translation parameters can be computed at each pixel location independently, the algorithm is suitable for parallel implementations.

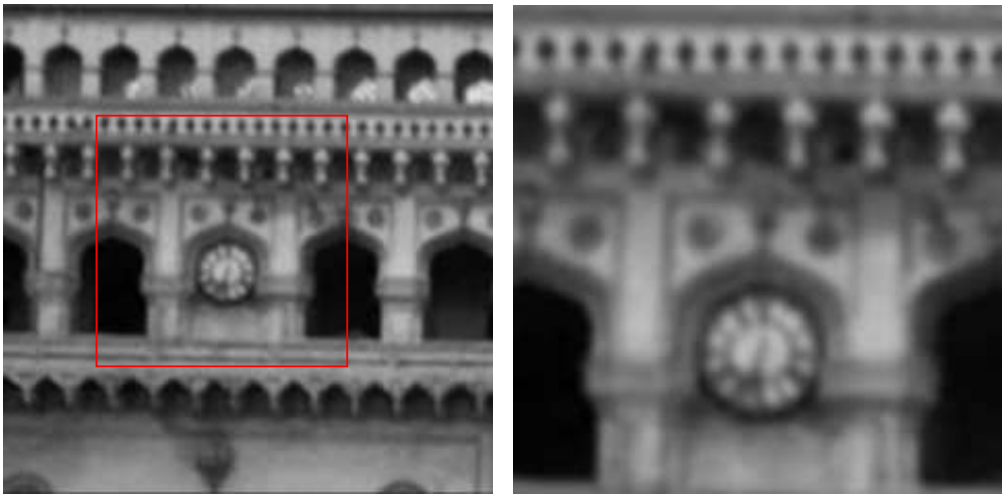
3.8 Summary

We proposed an algorithm for image registration, which is robust in presence of noise, non-uniform blur and illumination. The performance of traditional algorithm decreases due to these factors. All super-resolution reconstruction algorithms demand very high degree of accuracy in image registration. We have shown that our algorithm based on local phase is independent of blur and illumination artifacts. Our approach is also correspondence less, and hence there is no need of calculating features, explicitly. We have proven the convergence of the algorithm, even when it is impossible to identify the exact frequency of the underlying signal. Our algorithm is extensible to any general class of image registration which is not the case with other transformed domain approaches though both provides similar robustness.

(a) LR_1 (b) LR_2 (c) bicubic (d) original HR

(e) SR - actual (f) SR - intensity (g) SR - RANSAC (h) SR - proposed

Figure 3.6: Effect of registration inaccuracies on super-resolution of images corrupted with non-uniform illumination. (a) One of the low resolution frame decimated by a factor of 1.8; (b) LR image with non-uniform illumination; (c) bi-cubic interpolation of a part of the LR frame; (d) original HR image; (e-h) super-resolution with registration parameters calculated with different methods: (e) actual registration parameters, (f) intensity minimization, (g) RANSAC, (h) our algorithm.



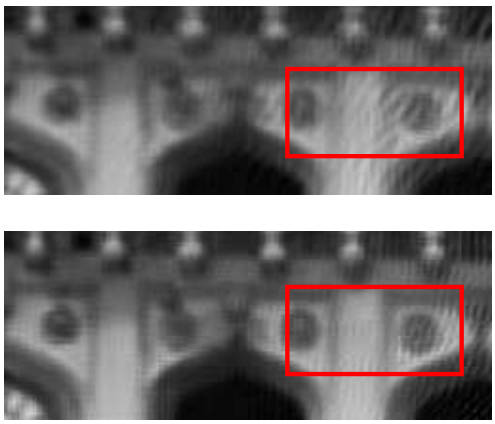
(a) LR (ROI)

(b) bicubic

(c) SR - intensity

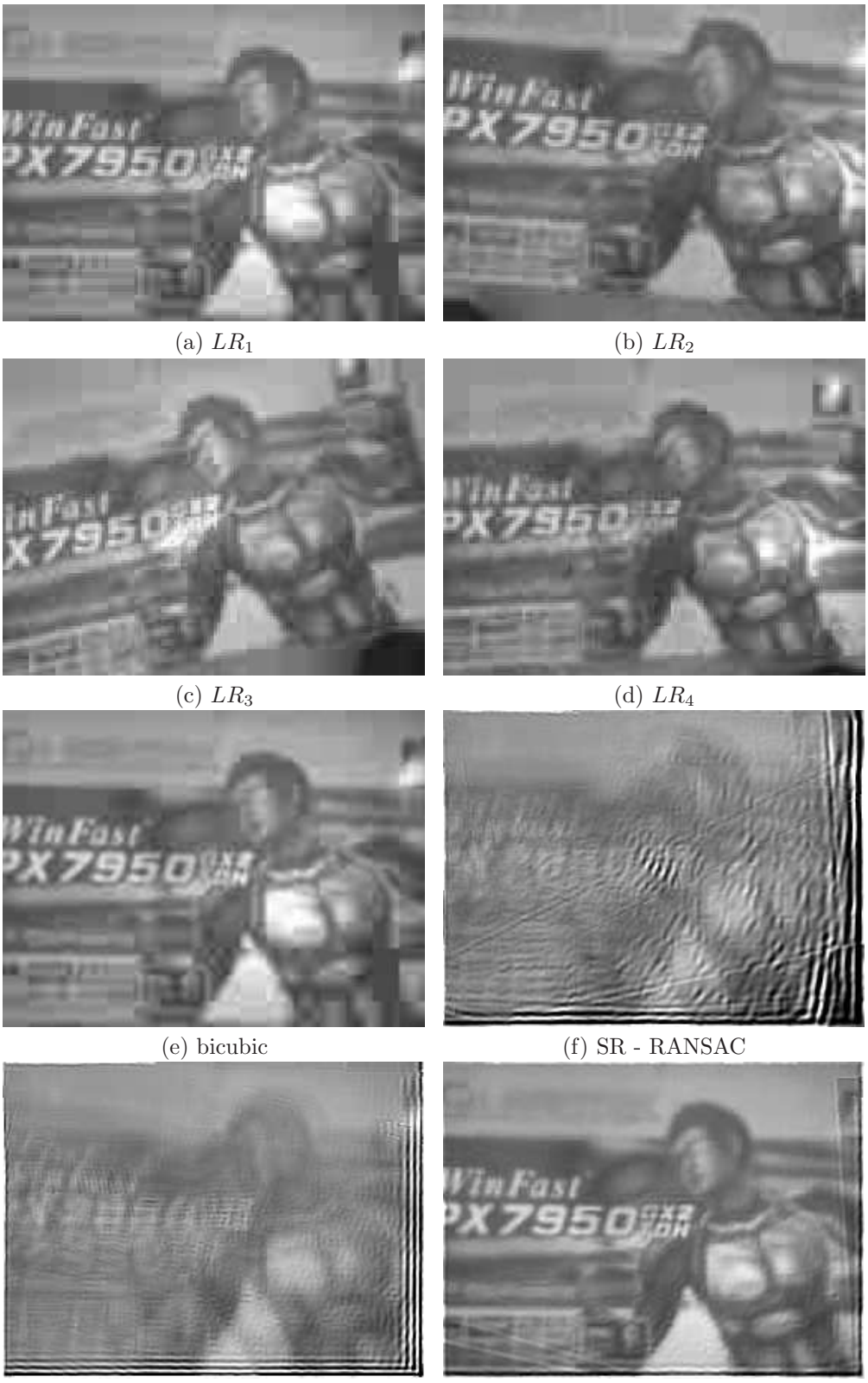
(d) SR - RANSAC

(e) SR - proposed



(f) SR (proposed vs RANSAC)

Figure 3.7: (a) One of the Low-Resolution frame (b) bi-cubic interpolation; SR reconstruction results using different registration algorithms (c) intensity minimization (d) RANSAC (e) phase-based method; (f) closer evaluation of SR reconstruction with registration using RANSAC (first) and phase-based method(second)



(a) LR_1 (b) LR_2

(c) LR_3 (d) LR_4

(e) bicubic (f) SR - RANSAC

(g) SR - intensity (h) SR - proposed

Figure 3.8: (a)-(d) LR input frames with varying illumination. (e) bicubic interpolation of (a); SR reconstruction results using different registration algorithms: (f) RANSAC, (g) intensity minimization, (h) phase-based method.

Chapter 4

Optimal Zoom Imaging: Capturing Images for Super-Resolution

4.1 Introduction

High quality image generation is an important problem that finds various applications in computer vision and image processing. Super-resolution (SR) [36] as we saw in earlier chapter deals with generating a high-resolution (HR) image from low-resolution (LR) image(s). Super-resolution algorithms are commonly divided into two categories, *viz.* multi-frame SR [30] and learning based SR [37].

- Lin and Shum [74] showed that the theoretical limit on magnification for multi-frame SR is 5.7, and in practical scenarios this limit is only 2.5. For higher magnification factors, the number of images required increases exponentially, making the computational cost beyond practical limits for most applications. Multi-frame SR also requires accurate registration and blur parameter estimate, which are very difficult to obtain in many scenarios. These drawbacks limit the applicability of multi-frame SR, and it is primarily used for revealing the exact underlying details at a limited magnification. The super-resolved images are useful to achieve higher recognition rates for various vision algorithms, e.g. [110].

- In contrast, learning based single image SR, in theory, can achieve magnification factors up to 10, as shown by Lin *et al.* [75]. The HR image generation is formulated as an inference problem. Correspondences between LR and HR patches are stored during the learning phase, and the HR image is inferred in a MRF framework with contextual constraints. This category of algorithms perform well for natural objects, where the perceptual quality is more important than accurate reconstruction of reality. They also work well if the training set is optimized for specific object/scene classes, such as faces [39]. However, the performance drops significantly on man-made structures, where even with a magnification factor of 3 (see Fig. 4(a), in Lin *et al.* [75]), the actual content need not be resolved in the final result.

We note that the bottleneck of a learning based SR algorithm lies in the nature of the underlying data, and the magnification factors achievable for various types of images or regions within an image, vary considerably. In other words, in order to get uniform perceptual quality after SR, different regions of an image need to be captured at different minimum resolutions. One could be conservative, and capture the whole image at the maximum resolution required by any image patch, which is both costly and redundant. Capturing minimum number of images in the whole

process require us to use learning based approaches. In this chapter, we propose a solution to this problem by capturing the image at ideal resolutions. The minimum required resolution for every patch of the image is predicted from a low-resolution image. Different parts of the image are then captured at the correct resolution, and thus sufficient amount of scene information is gathered at the image capturing stage itself. Any further magnification of the image can be achieved using any off the shelf single image super-resolution algorithm.

The ability to predict the ideal resolution for capture of an image region also enables a variety of applications. Automatically selecting the right resolution or zoom would enable efficient mosaicing of very large panoramas. Instead of capturing all the images at a high resolution [111], the final mosaic can be generated with fewer number of images at the right zoom level. The predicted resolution would also represent the minimum amount of information that is essential to represent a scene, and hence would reduce the computational cost of many vision algorithms that attempt scene understanding. Mobile robots could use this information to interpret and navigate the world more efficiently. Removing the redundant information that could be recreated using SR would also enable effective compression.

For most man-made structures, a limit on amount of scene information gathered can be quantified empirically. Note that primitives such as step edges along smooth curves can be enhanced effectively using single-frame SR. On the contrary, for most natural scenes, a very high value of zoom is required because of their detailed and intricate structure. However, one could replace the lost information with high-quality pre-captured content, without affecting the perceptual quality. We formulate the problem of capturing an image at ideal resolution in a patch based framework, where the ideal resolution/zoom is predicted separately for every patch. The ideal resolution or zoom will thus depend on the nature of the scene, the level of detail, and the information that can be captured by learning based SR algorithms, making the prediction challenging. We note the following points about image patches to predict ideal resolution factors.

- The structures in the image are assumed to have edges along smooth curves, which lends to enhancement by SR algorithms. The basic patch provides sufficient information to predict up to *smaller magnifications*.

- For *larger magnification* factors, the context information plays an important role, which is obtained from the predicted zoom values of the neighboring patches.

- The size of the patch is appropriately selected to provide enough structural information for smaller magnification factors and simultaneously include strong context information for predicting larger magnifications.

Once the patch size is selected, we need to learn the prediction function for the zoom level of individual patches, and to model the contextual relationship with neighboring patches. We use a Markov Random Field (MRF) framework, which is popularly used to incorporate contextual constraints.

In short, we propose an approach for high-resolution image generation by capturing sufficient information at the image capturing stage itself. The image is decomposed into patches and zoom level prediction is modeled as an inference problem in a MAP-MRF framework. We use Bayesian belief propagation rules to solve the network. As the optimization function contains numerous local minima, a robust technique is proposed to initialize the solution. Various practical constraints are proposed to minimize the extent of zoom-in. The results are validated on synthetic data and experiments are performed on real scenarios as well.

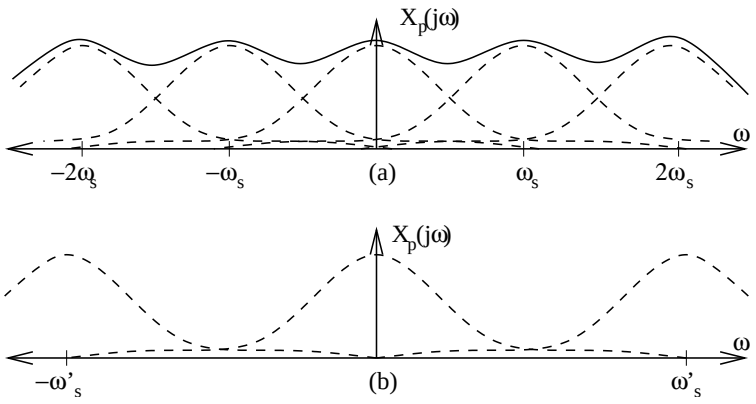


Figure 4.1: Fourier spectra of a hypothetical signal with different sampling rates; (a) sampling rate is low; (b) sampling rate is high enough so that the image can be zoomed in further easily with minimum aliasing. ω_s and ω_s' are sampling frequencies.

4.2 Related Work

There are different categories of work that address the problem of automated zoom detection from different perspectives. [112, 113] address the problem on zooming in on a pre-determined object by placing it to fill the image or by zooming-in only on the focused areas. Tordoff and Murray [114] model the zoom control for a tracking system. The goal is to zoom-in and out such that the target remains within the field of view of the camera with high confidence.

Image-cropping algorithms [115, 116, 117] can potentially be used to zoom the image to the desired target. The region of potential interest is selected from an image using a pre-defined criterion. The selected portion can be zoomed in to emulate automatic zooming. However, these algorithms do not address the resolution of the desired object and only directs the attention to it.

In computer vision literature, the term zoom has been used in different contexts. To avoid any ambiguities we mention some of them in related work. Traditionally, zoom-in refers to the change in focal length of the camera lens. Jin *et al.* [118] proposed a probabilistic model to detect zoom-in or zoom-out operations in an image sequence. Zooming-in is also used to refer to magnification of image using super-resolution algorithms, and not by camera e.g. [119].

4.3 Predicting the Right Zoom

The right zoom of the camera is such that the image captured at that zoom contains sufficient amount of information. Image can then be magnified further with simple algorithms which enhances edges and certain features. We first describe 'zooming-in sufficiently' from Nyquist view. Zoom prediction is modeled as an inference problem. The image is divided into patches and zoom factor is predicted for each patch. Both structural cues and contextual information around the patch are incorporated and modeled in a MAP-MRF framework. The network is solved using Bayesian belief propagation rules. Randomness measure is defined to initialize zooming factor in the network.

4.3.1 A Nyquist View of Zoom-in

The irradiance field observed by a camera requires very large frequency range to represent all information. One observation to be made is that the magnitude of Fourier spectra usually decreases as a function of increasing absolute frequency. According to the Nyquist theorem [81], a signal can

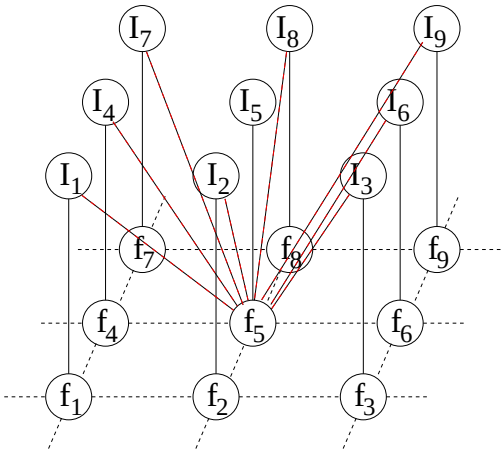


Figure 4.2: Markov Network for zoom prediction. $\tilde{I}_i$ are LR patches and the corresponding resolution front values f_i. The output value at any location is also dependent on certain information of neighboring patches and the context.

be uniquely reconstructed from its samples if the size of the band of input signal is less than the sampling frequency. Fig. 4.1 shows the Fourier spectra of a hypothetical signal at different sampling rates. If sampling rate is low, Fig. 4.1(a), signal is highly aliased and significant information is lost. On the other hand if sampling rate is high enough, Fig. 4.1(b), the aliasing is low and significant information can be recovered from the sampled information. Rest of the high frequency are usually step edges, which can be recovered by promoting step functions and edges along smooth curves while zooming-in, and noise, which can be characterized and ignored. This forms the basis of selecting the right zoom of the camera. Sufficient information is gathered at the image capturing stage so that any further resolution enhancements requires only simple feature enhancements.

4.3.2 Probabilistic Model

Image at the right zoom is captured in two steps. In the first step, a low-resolution image is captured and the zoom is predicted for each patch. In the second step the image(s) are captured at the right zoom. Before we describe probabilistic model we define the *resolution-front* of an image.

Definition 1. *Let* $\tilde{I} = \{\tilde{I}_1, \tilde{I}_2, ...\tilde{I}_N\}$ *be the image captured, represented as a concatenation of square patches,* $\tilde{I}_i$*, on a 2D grid each of size* $m \times m$ *at image locations* $1, 2...N$*.* **Resolution front** $R_f = \{f_1, f_2, ...f_N\}$ *of the image* $\tilde{I}$ *is the amount of minimum magnification* f_i *required at image patch location* i*, so that the block can be super-resolved further by using only simple feature enhancement algorithms.*

We essentially predict the resolution front rather than the absolute zoom required. It has more usability in various scenarios, some of which are discussed in experiments and results section. The prediction strategy should follow three principles mentioned before. We present our zoom prediction algorithm as an inference problem similar to inference problems presented by Freeman *et al.* [37] in a Markov Network. The Maximum-a-Posteriori (MAP) estimate of the resolution-front R_{MAP}

is given by,

$$\mathrm{R}_{MAP} = \arg\max_{\mathrm{R}_f} P(\mathrm{R}_f|\tilde{\mathrm{I}}), \tag{4.1}$$

$$= \arg\max_{\mathrm{R}_f} p(\tilde{\mathrm{I}}|\mathrm{R}_f)P(\mathrm{R}_f), \tag{4.2}$$

To simplify the inference problem, the formulation is reduced to a patch based model under Markov assumption similar to the one used in [37]. However, our patch structure incorporates intensities of all pixels of the underlying patch, $\tilde{\mathrm{I}}_i$, at higher weights and some pixels from neighboring patches at lower weights(see section 4.3.3). Let $\mathbf{y}_i$ be a column vector which contains intensity values of all such pixels. Markov Random Field (MRF) is a popular framework to include contextual constraints. Each node in the network corresponds to either an image patch or a resolution front value. Fig. 4.2 shows the graphical dependencies among nodes. To maintain compatibilities of resolution-front value predictions with neighbors, a 5-value resolution-front tuple is predicted at each location. It includes the resolution front values corresponding to underlying patch and its 4-neighbors. Let f_i^j denote the resolution front value, predicted using pixel information at patch location i, for patch at location j such that $j \in N(i)$ is one of the 4 neighbors. The maximum likelihood estimate $p(\tilde{\mathrm{I}}|\mathrm{R}_f)$ is,

$$p(\tilde{\mathrm{I}}|\mathrm{R}_f) = \prod_i p(\mathbf{y}_i|f_i, f_i^j)$$

$$= \prod_i \frac{1}{Z} e^{-\frac{1}{2}(\mathbf{x}_i-\mathbf{y}_i)^T\Sigma_1^{-1}(\mathbf{x}_i-\mathbf{y}_i)}, \tag{4.3}$$

where $\mathbf{x}_i$ is a vector from the training data for which the equation is optimized and the corresponding resolution-front assignment is the ML estimate. Σ_1 is a diagonal matrix that incorporates the weights given to different pixel values of the patch. The above equation is also known as pairwise compatibility function between input and output values in a Markov network [37]. The resolution front should be compatible and dependent on the neighboring context,

$$P(\mathrm{R}_f) = \prod_i P(f_i)$$

$$= \prod_i \prod_{j\in N(i)} P(f_i|f_j). \tag{4.4}$$

The compatibility function (equivalent to above function $P(f_i|f_j)$) between the predicted resolution front values and the neighboring values is proposed as,

$$\psi(f_i, f_j) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\left((f_i-f_j^i)^2+(f_j-f_i^j)^2\right)/2\sigma_2^2}, \tag{4.5}$$

where σ_2^2 is the variance. Substituting equation 4.3 and 4.5 into equation 4.1 we get,

$$\mathrm{R}_{MAP} = \frac{1}{Z'} \arg\max_{\mathrm{R}=\{f_1\ldots f_N\}} \left(\prod_i e^{-\frac{1}{2}(\mathbf{x}_i-\mathbf{y}_i)^T\Sigma_1^{-1}(\mathbf{x}_i-\mathbf{y}_i)} \right) \times \left(\prod_i \prod_{j\in N(i)} e^{-\left((f_i-f_j^i)^2+(f_j-f_i^j)^2\right)/2\sigma_2^2} \right). \tag{4.6}$$

Rather than maximizing the above expression we maximize its logarithm which is simple.

$$\mathrm{R}_{MAP} = \arg\min_{\mathrm{R}=\{f_1\ldots f_N\}} \left(\sum_i (\mathbf{x}_i-\mathbf{y}_i)^T\Sigma_1^{-1}(\mathbf{x}_i-\mathbf{y}_i) + \sum_i \sum_{j\in N(i)} \left((f_i-f_j^i)^2+(f_j-f_i^j)^2\right)/2\sigma_2^2 \right) \tag{4.7}$$

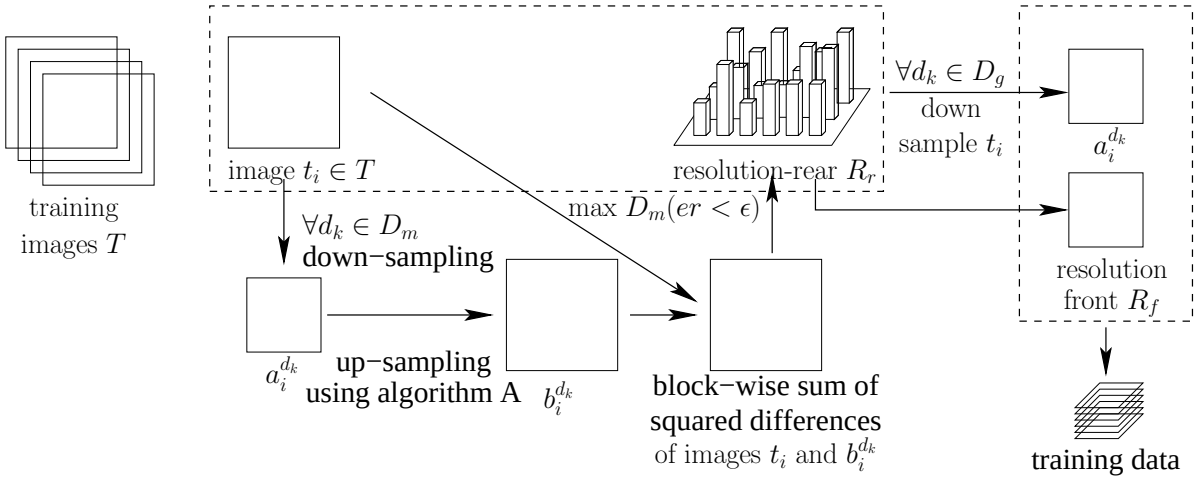


Figure 4.3: Generation process of the training data.

4.3.3 Patch Representation

Smaller patch size is desirable to increase the generalizability and larger patches for specificity. Square patch sizes having equal weights to all pixels are commonly used in a Markov network. We use slightly larger patch sizes but assign low weights to the pixels away from the center patch while computing the L_2 distance. An exponential decay function as shown in equation 4.8 is used to weight the patch pixels. This behavior is embedded in Σ_1 in the MAP formulation. The function describing the patch model is,

$$f(\mathbf{x}) = \begin{cases} c, & |\mathbf{x}| <= \mathbf{t} \\ \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp(-\frac{\mathbf{x}^2}{2\sigma_p^2}), & \mathbf{t} < |\mathbf{x}| <= \mathbf{p} \end{cases} \tag{4.8}$$

where c is a constant and $2t + 1$ is the underlying patch size.

4.3.4 Training Data Generation

Training patches are generated from selected images which a user believes that can be super-resolved further by using any single-frame SR algorithm. To simplify descriptions, we define *resolution-rear* similar to resolution-front as,

Definition 2. *Let $I = \{I_1, I_2, ...I_N\}$ be the given image, represented as a concatenation of square patches, I_i each of size $m \times m$ at image locations $1, 2...N$.* **Resolution rear** *$R_r = \{r_1, r_2, ...r_N\}$ of the image I is the amount of maximum down-sampling, r_i at image patch location i, so that the down-sampled block can be super-resolved to the original block I_i by using only simple feature enhancement algorithms. The image I has a resolution-front value $R_f = \{1\}$.*

For each image t_i, from the training images $\mathbf{T}$, we calculate how much down-sampling each block can tolerate. We downsample[1] the image at various downsampling values $D_m = \{d_1, d_2, ...d_t\}$ and then super-resolve the image using an algorithm $\mathbf{A}$. Block-wise sum of squared differences in intensity values is computed between the original and super-resolved image. If the error is greater

[1] downsampling factor=1/scaling factor

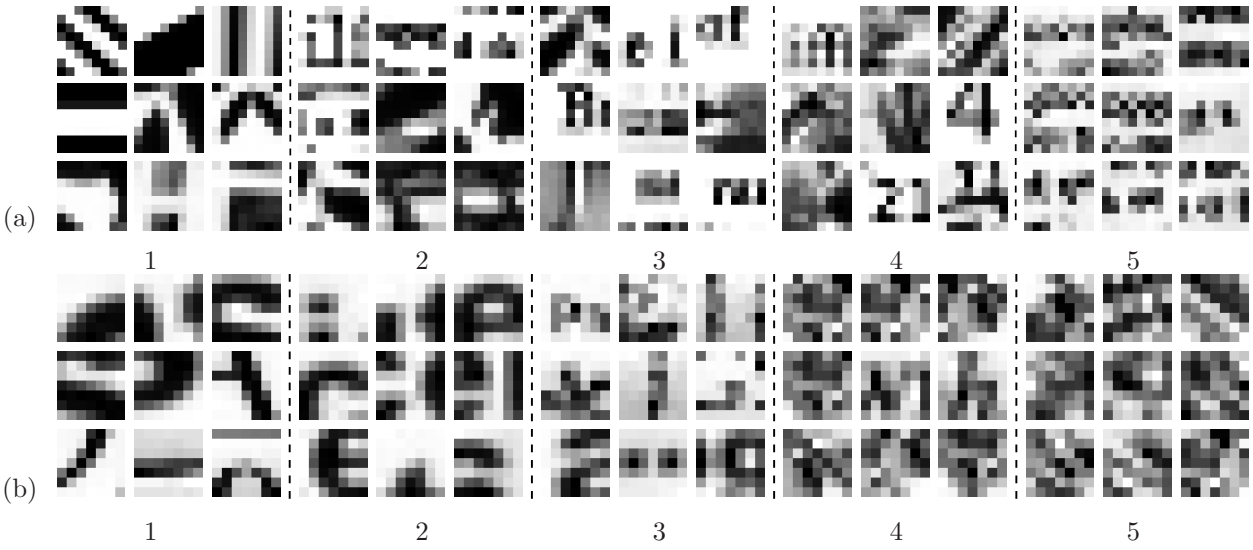


Figure 4.4: some patch structures and corresponding zoom-in values (a) computed in training phase. 4×4 is the central patch and 8×8 is overall patch with pixels from neighbors. (b) using randomness measure (sec 4.3.6).

than a threshold ϵ then then downsampling factor just smaller than current downsampling factor is assigned to r_i. For any down-sampled (by a factor k)version of image I, the resolution-front is computed from resolution-rear as,

$$f_i = \begin{cases} \frac{k}{r_i}, & \text{if } \left(\frac{k}{r_i}\right) > 1 \\ 1, & \text{otherwise} \end{cases} \tag{4.9}$$

The original image is downsampled at multiple resolution factors in D_g and resolution front is computed for each of them. 5-value tuple having resolution front value of the patch and its 4-neighbors are stored along with the patch in the training database. Fig. 4.3 explains the training patch generation process. When an image is down-sampled the block size varies at different downsampling factors. For the sake of efficiency in searching, a constant patch size is required. We take a constant block size and assign the second highest (to avoid outliers) resolution-front value. Training data is generated at various equally spaced non-integer zoom values as well. Fig. 4.4(a) shows patch intensity structures corresponding to integer zooms only.

For higher accuracy, images at various resolutions should be captured from the camera. We prefer to downsample images offline because, a) Computation of lens distortions parameters, which are different at different focal lengths, and estimation of registration parameters need to be highly accurate and the process is computationally expensive; b) Varying degree of error in measurements irradiance field and presence of noise. Certain relaxations are incorporated in error limits at various stages.

4.3.5 Energy Minimization

Solving equation 4.7 for global minima is computationally prohibitive with large number of patches. Freeman *et al.* [37] favored to obtain a local minimal solution which approximates the global minima. Using approximate nearest neighbor data-structure [90] a smaller set of similar patches (usually 20-30) are obtained. Markov network is solved using local message passing algorithm

(belief propagation). Rules are same as proposed in [37]. It was argued that these rules can be applied on graphs with loops as well without significant deviation from solution. However, presence of multiple local minimas requires a robust initialization, which is often ignored. In the next subsection, a general method is proposed to initialize the resolution-front values.

4.3.6 Robust Initialization

Each pixel value is initialized to a zoom value proportional to randomness in intensity structure in the neighborhood. Randomness measure P_i at location i is proposed as,

$$P_i = \sum_{j=[-3,3]\times[-3,3]} \boldsymbol{\sigma}_{gr_3}(i+j)\boldsymbol{\sigma}_{in_3}(i+j), \tag{4.10}$$

where $\boldsymbol{\sigma}_{in_3}$ is the variation in intensity values and $\boldsymbol{\sigma}_{gr_3}$ is the variation in gradient angle[2] in a 3×3 window. Their product at every patch location in a 7×7 window is added. Intensity of a patch is normalized before calculations. The zoom value is directly proportional to the randomness measure. High intensity variations and low angle variation imply ramp like structure. High angle variation and low intensity variation imply noise. Higher value of both imply higher zoom factor. To identify ridge like structure as regular structure, gradient angle is computed in the range $[-\pi/2, \pi/2)$ instead of full 2π range. Proposed randomness measure fails to identify step edges because just after the steep, gradient angle could be anything in presence of noise. These edges are removed from consideration using canny edge detector. Proportional to the randomness measure zoom value is assigned as successive integer levels and Markov network is initialized. Fig. 4.4(b) shows some of the patch structures and the estimated zoom values.

4.4 Calibration of Zoom Lenses

"Zoom lens model" defines the relationship between focus, zoom and aperture values. In multi-element zoom lenses, the scene magnification is controlled by moving two or more lenses along the axis and the point of focus is selected by moving the whole lens assembly to and fro. The functional relationship between the various zoom lens parameters is obtained empirically rather than mathematically [120] because of high complexity of zoom lenses, unavailability of specifications of lenses and missing markings of zoom and focus motor position. We predict zooms upto 5X in experiments. We use two zoom lenses with focal length in the range 18-55mm and 28-105mm because of unavailability of a single high zoom lens. Virtually a $105/18 \approx 6X$ zoom lens is used. A high precision scale is affixed on focus and zoom motors. Multiple images of a checkerboard pattern are captured as a function of distance (in feets) and zoom position (in motor units). Homography matrix is computed among between the base image and other images of the pattern. Average of scale factor along the two axes is the effective magnification. Fig. 4.5 shows the calibration graphs. Coupling table 4.1 is made which defines the relationship between two zoom lenses. It is the magnification achieved at minimum focal length by changing the zoom lenses.

Setting the Right Zoom: Let the first image is captured at (z_i, t_i), z_i and t_i denote the zoom and focus motor position respectively. Let m denote the required magnification factor and (z_f, t_f) denote the desired zoom lens configuration. If zoom motor position is fixed in the graph, then the focus position is a monotonous function of distance from the pattern. This result follows from the fact that only one depth point of the scene remain in focus. Let $M(,)$ and $F(,)$ be the magnification profiles(Fig. 4.5(a),4.5(c)) and focus profiles(Fig. 4.5(b),4.5(d)) of the lenses. Let

[2] $dx = x_{t+1} - x_t,\ dy = y_{t+1} - y_t,\ \theta = \tan^{-1}(dy/dx)$

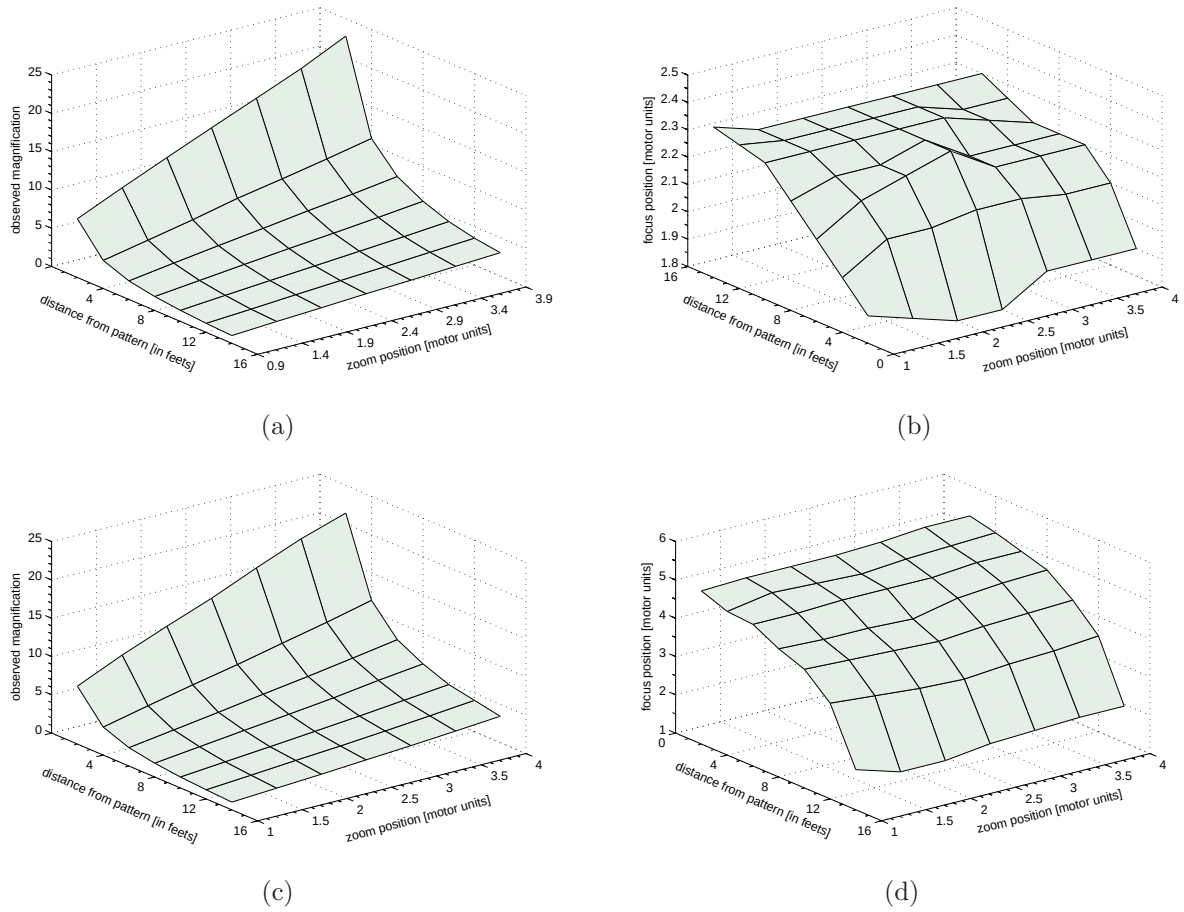


Figure 4.5: **Zoom lens calibration** (a) and (c): magnification profile of two cameras as a function of zoom motor position and distance of the camera plane from the checkerboard (measured in feets); (b) and (d) corresponding focus position in motor units.

M_k^{-1} and F_k^{-1} denote the inverse of M and F at a constant k. The required zoom-lens parameters (z_f, t_f) are obtained as,

$$d = F_{z_i}^{-1}(t_i),$$

$$z_f = M_d^{-1}(mM(d, z_i)),$$

$$t_f = F(z_f, d).$$

Intermediate values are computed by fitting higher order polynomials as described in [120]. Coupling table is used to switch to other zoom-lens and the equations are similar.

Distance (in feets)	2	4	6	8	10	12	14
Magnification	1.5617	1.5755	1.5695	1.5724	1.5770	1.5898	1.5832

Table 4.1: **Coupling table** : computed at the minimum focal length between two lenses as a function of distance.

4.5 Experiments and Results

To evaluate the performance of the proposed algorithm experiments are performed on a variety of real and simulated data-sets. As data is lost near boundary, several constraints are proposed to minimize the extent of zoom. Later in this section, several possible applications are also discussed. Around 54 images are selected which can be super-resolved further using simple SR algorithms. The randomization measure defined in section 4.3.6 is also used to check the suitability of training images. The size of the training patch is 8×8. It has 4×4 pixels from the underlying patch and other pixels from neighboring patches. Around 110,000 training patches are generated. Each training image is downsampled at various factors upto 8. 4-5 of these images are chosen and resolution-front values of them are computed. Patches are stored in the training database. Any learning based SR algorithm can be used during training phase (denoted as algorithm **A** in Fig. 4.3). [121, 40] are recent such algorithms. References therein provide further details on various similar algorithms. This algorithm is used in our experiments. The zoom is predicted upto $5X$ at intervals of 0.25. The desired zoom of the camera is calculated from the predicted resolution-front value. It is done by finding a largest rectangle (usually located at the center) for which the maximum resolution-front value is less than or equal to the size of the image divided by the size of the rectangle.

Performance on Synthetic Data: We take test images and downsample it. The resolution-front of each of them is computed as described in section 4.3.4 and also using our algorithm. The comparison with initialization and prediction in MAP-MRF framework is summarized in table 4.2.

Image	MSE (initialization)	MSE (MAP-MRF)
Book-shelf	0.3453	0.2671
Butter-fly	0.2921	0.2424
Bill-board	0.3672	0.2938
Book-text	0.3156	0.2398
Painting	0.2801	0.2250

Table 4.2: Evaluation results on synthetic data. Mean square error (MSE) is computed between the actual resolution-front value and computed using (a) randomization measure, (b) MAP-MRF.

Results on Real Data[3]: We first evaluate the performance on Snellen eye chart, which has various random alphabets printed at different font size. Fig. 4.6 summarizes the result. At various locations in Fig. 4.6(c) the resolution-front values are highly regularized. Whereas in Fig. 4.6(b) regions around the text are also marked for zoom. Prior information learned from the training data (e.g. regions above and below text should require no zoom) was useful. Fig. 4.6(f), 4.6(g) and 4.6(h) are images captured at increasing zoom. Various characters are clear at different zoom-levels.

Fig. 4.7 summarizes results on a slightly complex scene. Contextual constraints was very helpful in regularizing resolution-front values. In some of the cases, the final character size after zooming is slightly different. This is primarily because of different font types. Also predicting high zoom values from limited data could be slightly erroneous.

4.5.1 Constrained Zoom-in

To minimize the data loss near outer boundary of an image, several constraints are introduced. Scene is zoomed-in upto a level only if the constraints are met.

[3] all images in this section should be enlarged to view them properly.

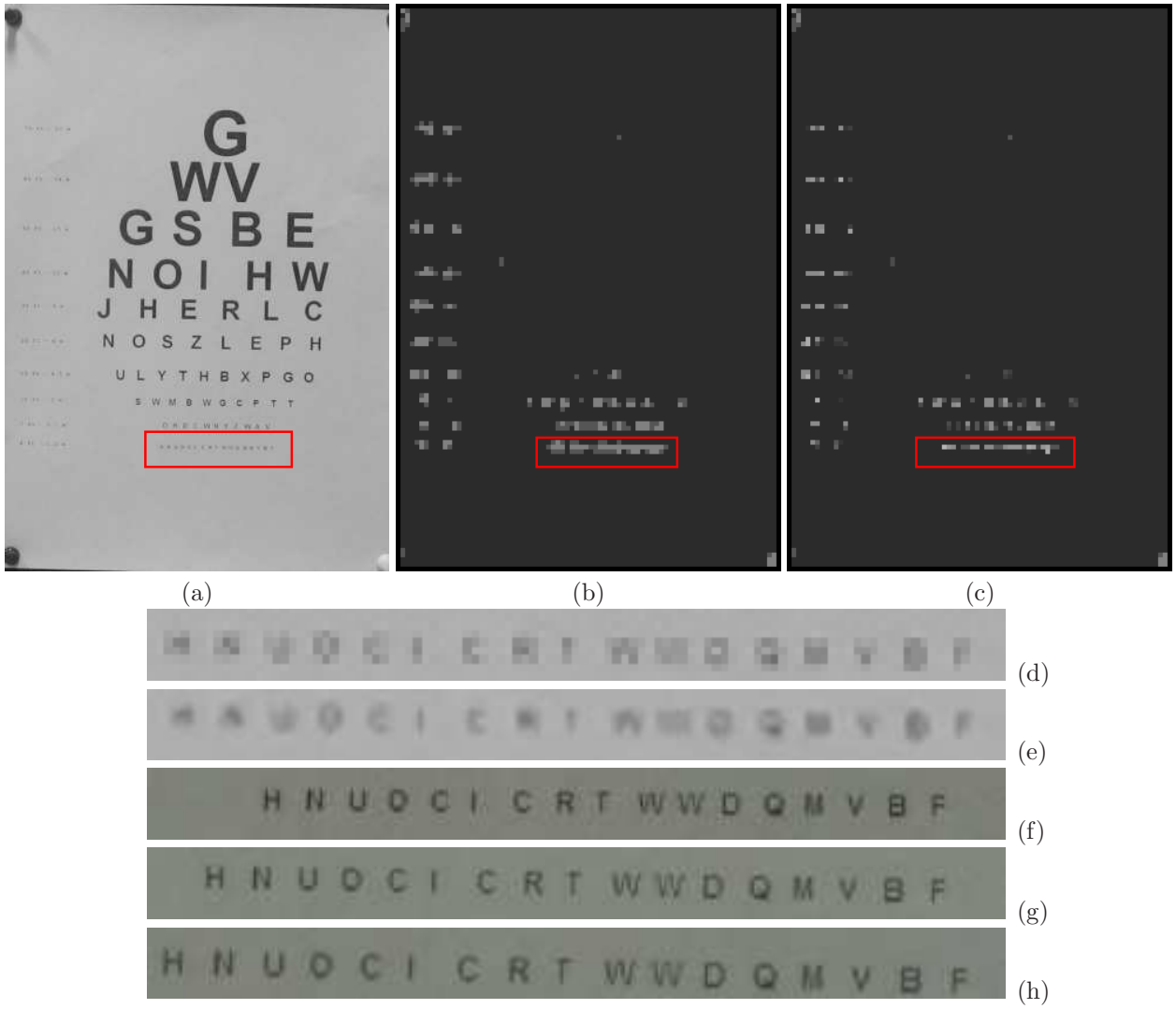


Figure 4.6: **Experiments on Snellen chart** (a) base image (b) zoom predicted using randomness measure with maximum zoom value 3 in the selected region (c) resolution-front predicted after optimizing equation 4.7 having values 3, 3.25, 3.5 and 4 in the selected region (d) selected region scaled by a factor 4 (e) super-resolved region; same patch after capturing images at zoom: (f) 3X (g) 3.5X (h) 4X.

Visually Attentive Objects : To speed up many computer vision algorithms, certain regions are preferentially processed based on their visual attentiveness. This constraint is used to preferentially treat a region which is visually attentive. Publicly available 'Saliency Toolbox' which implements the algorithm by Walther and Koch [122] is used to locate such regions. Fig. 4.8 summarizes the results.

Penalty for not Zooming-in : The zoom is costly if a few blocks require very high zoom. A graph is constructed on zoom factor versus number of blocks requiring zoom factor greater than various zoom factors. Graph is normalized and the first zoom factor where the value falls below a threshold is selected. It is also helpful to cope noise in resolution-front prediction. Fig. 4.9 summarizes the results.

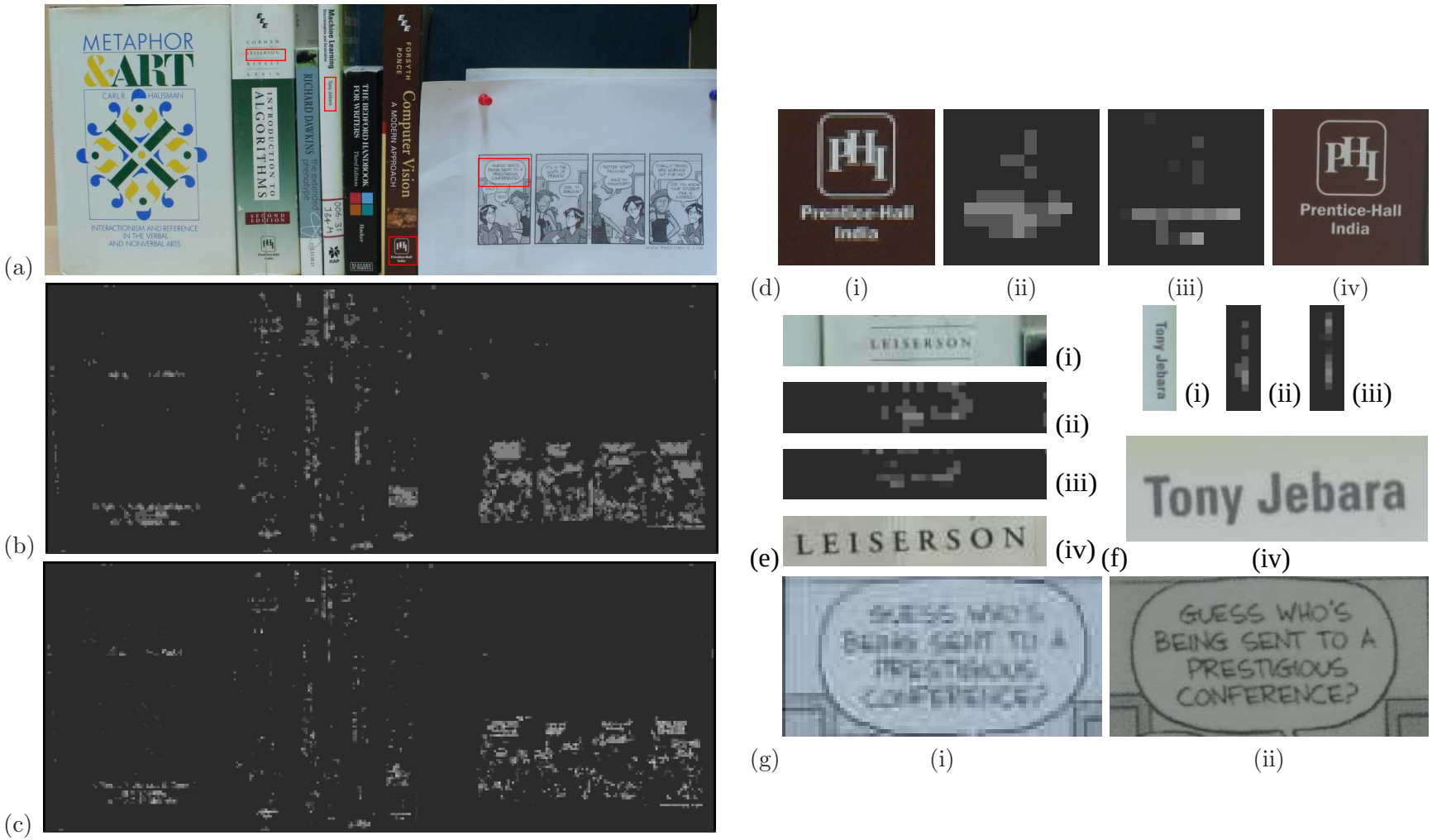
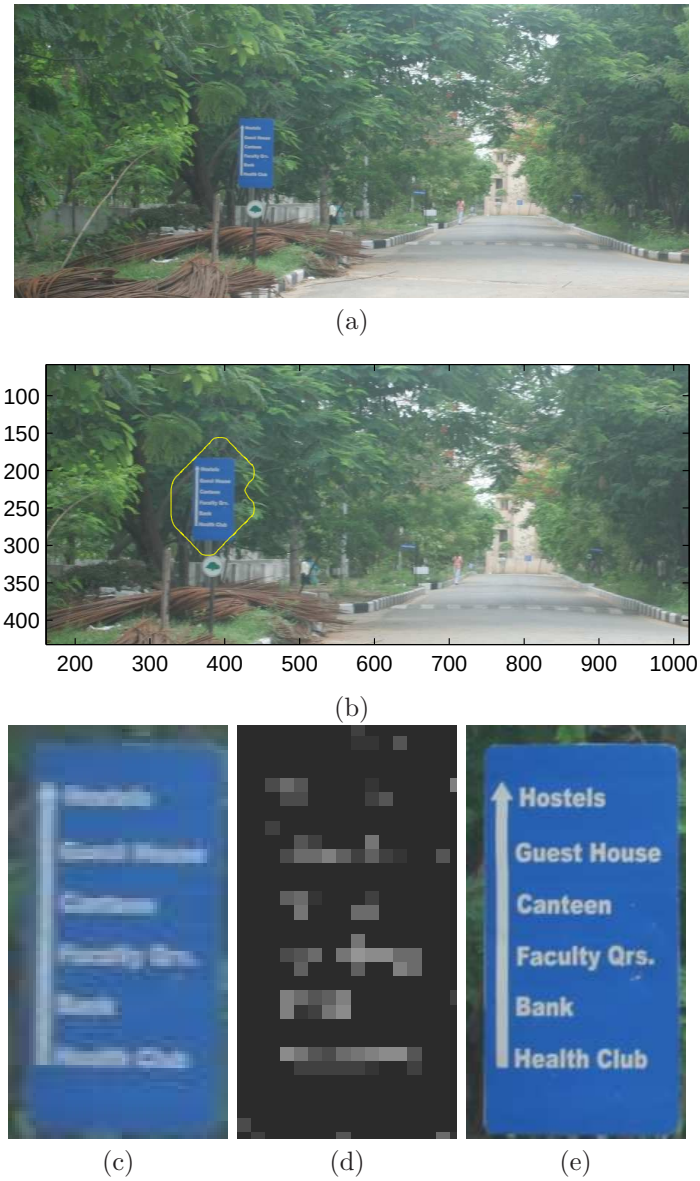


Figure 4.7: (a) base image (b) zoom predicted using randomness measure (c) resolution-front predicted after optimizing eq. 4.7; (d), (e) and (f): (i) selected regions from image (ii) initial resolution-front (iii) resolution-front after optimization (iv) regions shown at right zoom with values (d.iv) 3.5X (e.iv) 2.5X (f.iv) 2.5X (g.ii) 2.5X.



(a)

(b)

(c) (d) (e)

Figure 4.8: (a) base image (b) visually attentive region selected using saliency toolbox (c) selected LR region (d) R_f predicted (e) at right zoom (2.5X).

Other Scenarios : Pre-determined objects can be segmented and such image regions is kept at higher priority. Separating man-made and natural structures [123] also provide useful constraints for zooming. Natural objects and scenes have fine details but they convey very little useful information. Whereas man-made object usually do not have intricate structures. Natural structures can also be replaced with any high quality texture while super-resolving.

4.5.2 Applications

Integration of the proposed technique with professional or consumer cameras can provide a simple way to capture high-quality images. The algorithm can also be used to predict required magnification factor for multi-frame SR algorithms. Robotics and surveillance systems require the

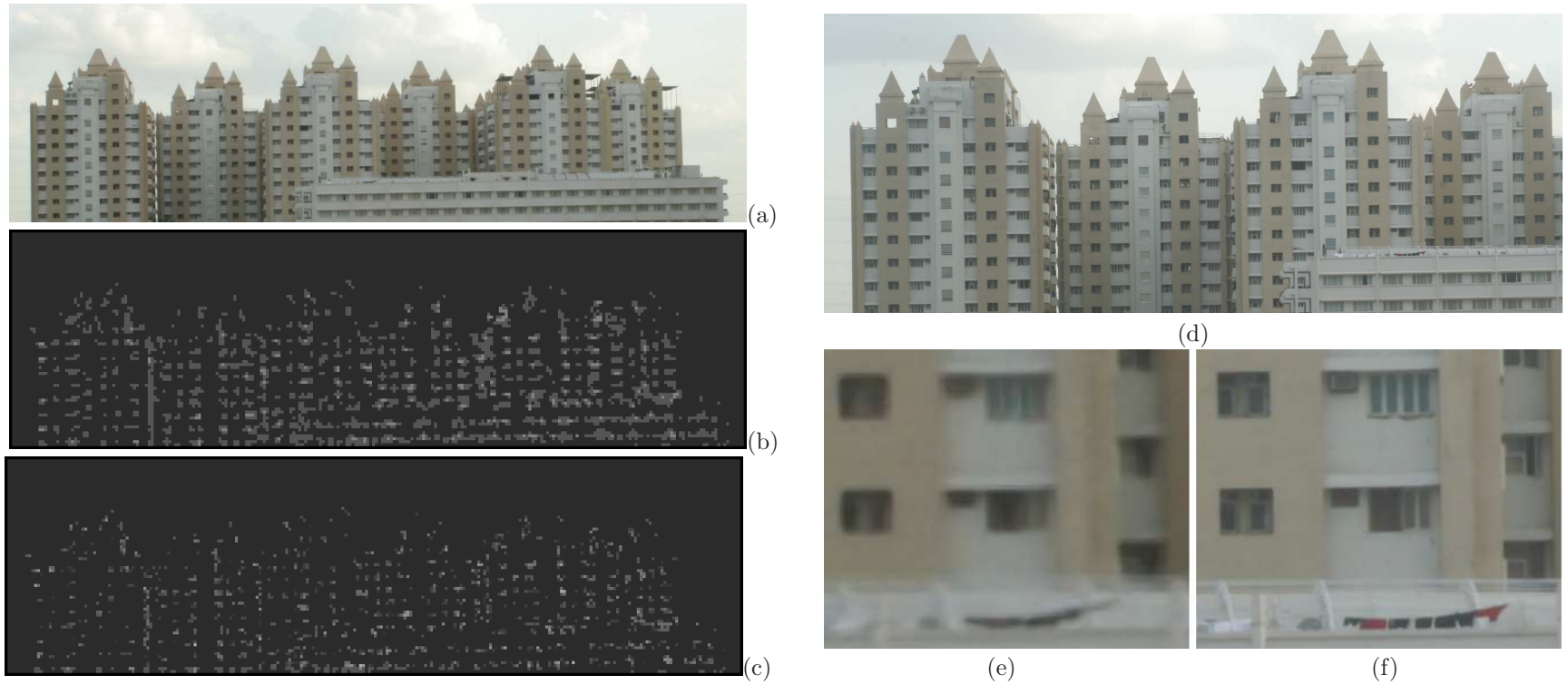


Figure 4.9: (a) base image (b) resolution-front initialization (c) resolution-front predicted (d) image captured at 2.5X zoom. Highest resolution-front value was 4.25X; (e) super-resolved image of (a) by 5X; (f) super-resolved image of (d) by 2X. Many structures are clear in (d).

interpretation of scenes which are usually unknown. It is impossible to scan scenes at maximum zoom value. Given that the most of the scene information in real world do not convey meaningful information or do not require very high zoom values. The scene can be captured optimally with minimum number of images at right zoom. For large scale image mosaicing (e.g. giga-pixel camera [111]) such algorithms can optimize the number of images captured. In automated cropping systems, regions which require very high zoom values can be removed after predicting the resolution front using our algorithm. As images captured at right zoom has almost all the information for further resolution enhancements, consequently the recognition accuracies of many systems will improve. For many real-time applications e.g. video surveillance, two camera systems can be used. One for capturing the whole scene and the other to capture only certain regions in detail.

4.5.3 Discussions

In Fig. 4.7(d.ii) and 4.7(f.ii) the initialized resolution-front values are not consistent and correct. With contextual constraints much of the regularization is brought in and resolution-front values are suppressed at unusual places. Places where underlying patch information is insufficient to predict high zoom values, context information played a significant role. In Fig. 4.8(e), the vertical line and the text have almost similar structure but the presence of context information is able to define right resolution-front values at various places. Selecting the right zoom value can as well be proposed as a high-level vision problem where a particular object is zoomed in at a pre-defined value. Proposing it as a low-level vision problem provides high degree of generalizability for a variety of scenes. Low computational speed is one of the key issues. But with additional constraints (section 4.5.1) significant speed up has been achieved. Camera shakes introduce blur in images and deteriorates the zoom prediction. But it can be controlled in autonomous environments.

4.6 Summary

In this chapter, we have presented and addressed the problem of capturing the right amount of scene information from the perspective of SR. The final captured image can be magnified further using any learning based SR algorithm. The solution is proposed in a MAP-MRF framework. MRF allows modeling of contextual constraints. Future work is towards developing complete real-time systems for zoom prediction. We also plan to address the problem of locating useful structures in images. We envision that such a functionality would be introduced in consumer cameras.

Chapter 5

Capturing Projected Image Excluding Projector Artifacts

5.1 Introduction

During the recent years, projector-camera systems have become popular for its use in computer vision [124], human computer interaction [125], improving projection quality and versatility [126, 127, 128], immersive environments [79], visual servoing [129], etc. Image projected on a scene by a projector suffers from pixelation artifacts due to gaps between the neighboring pixels on the projector's image plane, defocus artifacts due to varying scene depth and color transformation artifacts. If the camera is sufficiently close to the scene the pixelation artifacts are clearly visible. Capturing a high quality projector-scene composition in dynamic environments is even more challenging. In some of the cases, a person might not have the access to the input images (e.g., photographing a slide presentation). Interestingly, the pixelation artifact should also be useful in certain applications. In this chapter, we address two key problems.

- The *first problem*, is to accurately localize each of the projected pixels. Detection of the projected pixels in the captured image can facilitate applications such as recomposition of a projected image within a scene, which is useful in the post production stage. Procams are also useful for capturing surface properties. Considering the pixelation and blurring artifacts improves the accuracy of such estimations. Feature computations, such as SIFT, in a captured image can be inaccurate due to these artifacts. The relative spatial configuration of the localized pixels help in computing a dense shape for dynamic scenes. Usually, either a time shifted stripe pattern [124] or stereo image pairs [130] are used for static scenes.

- The *second problem*, we address is the restoration of the captured image having pixelation and defocus artifacts. Public capturing of images of various projector scene composition such as presentation slides, immersive environments [79] etc. require restoration. Projector-scene composition is also useful in movies for special effects. Images are rendered on real objects and the video is captured [80].

The localization and restoration of captured images is difficult due to a variety of factors such as spatially varying blur, background texture, noise, shapes of scene objects, and color transformations of projector and camera. The clue for the identification of projected pixels is that the sinusoids describing them share the same frequency with the neighboring pixels. The frequency, describing the sinusoids,can change over the whole image depending on the scene shape, and need to be estimated

locally. In this way we employ Gabor filter is used the frequency of the repeating sinusoid within a window. Gabor filter is widely used because of its frequency selective properties. We extend the usage of local phase, computed using the Gabor filter, to isolate each of the projected pixels distinctively. Local phase is robust to noise and intensity variations and as shown in chapter 3, it is invariant to a class of blur kernels. For restoration of the captured images, we reproject the projected pixels such that these artifacts are absent. To improve the quality further we propose a mechanism to virtualize a high-resolution projector.

5.1.1 Related Work:

To our best knowledge, this is the first work focusing on localizing projected pixels accurately and on enhancing captured images. Limited amount of work exist on display systems that reduce projector artifacts. However, they are applicable to static scenes and requires careful calibration and elaborate hardware setups. Zhang and Nayar [131] proposed a mechanism to project defocused images using a co-axial camera-projector. By slightly defocusing the projector and using the defocus compensation algorithm the pixelation artifacts are removed. Venkata and Chang [128] proposed simulating high-resolution projector using super-imposition of multiple low resolution projectors. Bimber and Emmerling [127] used multiple projectors, each having different focal planes, to project focused images at multiple depths. In all these cases, the main objective is to display a high-quality image on a surface. Note that in some situations it is not necessary or practical to display a high-quality image to improve the captured image. A seemingly related problem is restoring the halftone images from scanned documents. However, the techniques used are not applicable as the halftone images are binary images, where different configurations of dots are perceived as grayscale rendition. In our case, the captured images have varying background texture and blur.

5.1.2 Our Contributions:

We first analyze the structure of the projected pixels on the textured scene and propose a systematic approach to localize the projected pixels and remove the projector artifacts in the captured image. As our algorithm requires only one image, our system can work for dynamic scenes as well. No camera-projector calibration or co-axial camera-projector system is required. Specifically, we propose:

- An image re-formation model that describes the relationship between the display image, the projected scene, and the captured image with pixelation and defocus artifacts.
- A robust algorithm for identification of the projected pixels seen in the captured image.
- A method to remove the pixelation and blurring artifacts of the projector in the captured image.
- A mechanism to improve the quality of the captured image further by virtualizing a high-resolution projector, so that the captured image sees larger number of projected pixels.

Experiments have been performed on scenes of different complexities, under different projector settings, to show the robustness of the proposed approach.

5.2 Problem Formulation

In the formulation, we refer to the color image that is projected on the scene as the *display image*, and the image captured by the camera after the display image is projected on the scene as the

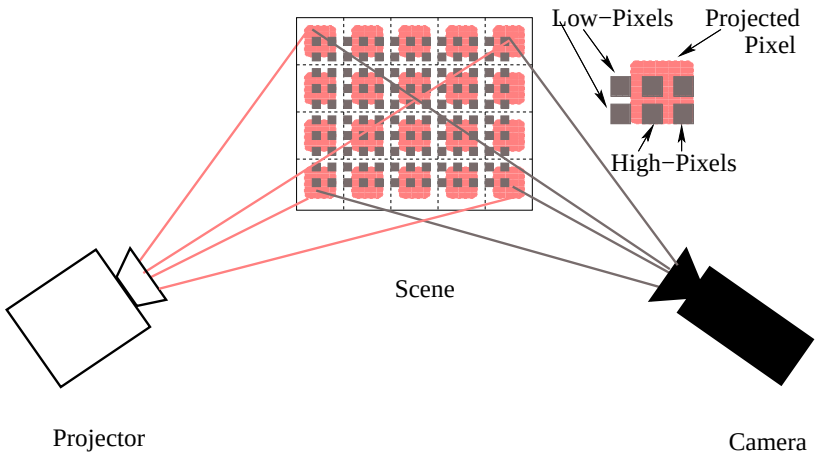


Figure 5.1: A projector-camera system: Red squares correspond to the pixels of the projector, and black pixels correspond to the pixels seen by the camera. High-pixels and low-pixels in the captured image are also marked.

captured image. We assume that the camera is sufficiently close to the scene so that each pixel projected on the scene is seen by more than one pixel in the camera's CCD. If this is not the case, pixelation would not be visible or can be treated as minor noise. The pixels of the captured image are classified into two categories: *high-pixels*, that see projected pixels of the display image, and *low-pixels*, which see the portions of the scene between neighboring projected pixels. High pixels have higher intensity values than the neighboring low pixels and hence the naming. Fig. 5.1 shows a typical projector camera system. The pixels shown are the locations seen by the pixels in the projector's LCD panel and the camera's CCD. We use the term *center-high-pixel* as the centroid of the group of high-pixels corresponding to a single display pixel. We frequently use the term *projected pixel* to mean the group of high-pixels that correspond to a single pixel of the display image.

Now we mathematically formulate the image re-formation model, which describes the relationship between the display image and the captured image. This relationship is defined in terms of color transformations at image plane, blurring artifacts of both the camera and the projector, pixelation artifacts of the projector, scene deformation and radiance due to ambient light. Given a display image $\mathbf{x}$, represented as a column vector, the image is transformed by a matrix $\mathbf{C}_p$, which models radiometric response of the display device, projector brightness, and spectral response of the projector channel (for more details see [126]). This discrete color transformed input is converted into continuous domain with pixelation, due to the gaps between the neighboring pixels on the projector plane, by a discrete to continuous transformation, $\mathrm{p}_p(.)$. The output of this function is convolved with a blur kernel b_p of projector's lens, which is a function of the scene depth, and the image is mapped on to the screen by the transformation function $\mathrm{f}_p(.)$, which is with respect to camera's co-ordinate frame. α_s models various scene surface properties and k_s is the radiance due to ambient light. The scene is mapped to the CCD of the camera by the transformation $\mathrm{f}_c(.)$ and it is blurred by the camera's lens with the blur kernel b_c, which is also a function of scene depth. The image is converted into digital form by $\mathrm{d}_c[.]$ and the color space is transformed by the matrix $\mathbf{C}_c$, which models various camera's CCD parameters. $\mathbf{y}$ is the final image captured. The process is mathematically represented as:

$$\mathbf{y} = \mathbf{C}_c \mathrm{d}_c[\mathrm{b}_c * (\mathrm{f}_c(\alpha_s \mathrm{f}_p(\mathrm{b}_p * \mathrm{p}_p(\mathbf{C}_p \mathbf{x})) + \mathrm{k}_s))]. \tag{5.1}$$

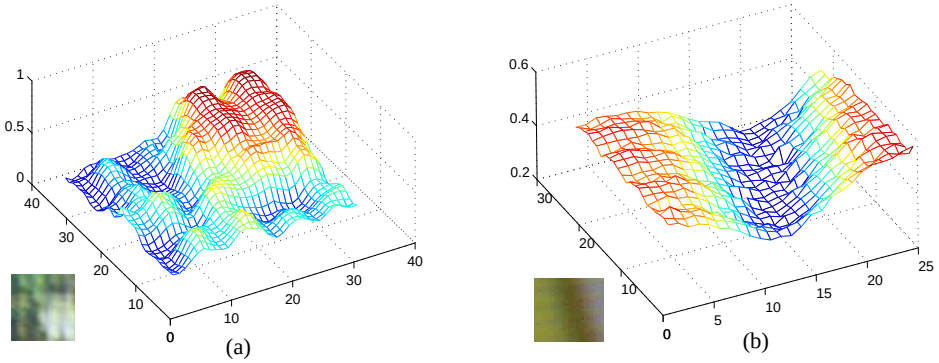


Figure 5.2: Intensity plots of patches from captured images; (a) with scene texture and no blur and (b) without scene texture and blur

Aperture value of the camera is set to the lowest so that we can get wide depth of the scene in focus and the blurring due to camera lens is negligible. The goal is to rectify a single captured image such that the deblurring and pixelation artifacts of the projector are not present. Mathematically, the problem can be described as: given an image captured using the model in equation 5.1, restore it such that it were captured using the following model:

$$\mathbf{y} = \mathbf{C}_c \mathrm{d}_c[\mathrm{f}_c(\alpha_s \mathrm{f}_p(\mathrm{n}_p(\mathbf{C}_p\mathbf{x})) + \mathrm{k}_s)], \tag{5.2}$$

where $\mathrm{n}_p(.)$ is a function that converts the discrete image pixels into the continuous space without any gap between the neighboring pixels on the projector plane. To restore the captured image, the main algorithm involved is the identification of center-high-pixel (describe in the next section). We also propose a solution for virtualizing a high-resolution projector. By virtualizing a high-resolution projector, we mean that given an image that has been captured with the image re-formation eq. 5.1, restore the captured image such that $\mathbf{x}$ is of high resolution having more number of pixels per area and it has been captured using the image formation eq. 5.2. Essentially high-resolution virtualization is simulating a high quality and high-resolution projector.

5.3 Characterizing High-Pixels

We first analyze the nature of the captured image to gain an insight on kind of algorithms suitable for robust characterization of high pixels. The problem is also analyzed from the perspective of equation 5.1 and 5.2. The assumption is made that the blurring is not extreme. The algorithm tries to determine center-high-pixel locations.

Fig. 5.2 shows the intensity plots of patches from the captured image. Local intensity sinusoidal peaks are visible in Fig. 5.2(a) and sinusoids are visible along left and right edges in Fig. 5.2(b). Another observation is that the intensity values is slightly higher between two consecutive projected pixels along horizontal or vertical direction than along diagonal direction. It allows us to decompose the problem into the detection of sinusoids in two orthogonal directions separately, for robustness rather than modeling a single pixel.

The clue for the identification of projected pixels is that the sinusoids describing their shape share the same frequency with the neighboring pixels. The repeating dominant frequency in a small window is estimated and the local phase corresponding to these frequencies are used to isolate

each of the projected pixels. In equation 5.1, for any general scene, the parameters $\mathbf{C}_c$,$\mathbf{C}_p$ and k_s remains almost constant in a small window. The change in frequency content due to this is minor. The shape parameters, f_c and f_p, are also assumed to be smoothly varying and the computation of frequencies is windowed. As the Gabor filter is a non-ideal band pass filter, it handles minute change in frequencies and orientations easily. Aperture of the camera is set so as to bring the scene in focus. Local phase is invariant to a class of blur kernels, which are even functions(see chapter 3) and is also known to be invariant to illumination and robust to noise [84]. This property helps for robust isolation of projected pixels. Background texture can change the frequency content causing errors at low intensity projection but in general scenarios the change is small. All this properties help to easily estimate $\mathrm{p}_p(.)$ in presence of all these artifacts.

5.3.1 The Algorithm

The projected pixels of the display image, as seen in the captured image, can be thought of as the intersection of two sets of equally spaced parallel lines which are approximately orthogonal. By calculating the orientation and frequency of these lines, and then taking the intersection of them, we identify the location of the projected pixels in the captured image. Before that the captured color image is converted into gray level and the local contrast is normalized.

Local Contrast Normalization: The captured image is locally normalized to a zero mean and a unit variance. This is done so as to separately highlight the high-pixels and the low-pixels uniformly in the image. Each pixel value of the captured image, $\mathcal{I}(i,j)$, is reinitialized as

$$\mathcal{P}(i,j) = (\mathcal{I}(i,j) - \mu_w(i,j)) \,/\, \sigma_w(i,j), \tag{5.3}$$

where $\mu_w(i,j)$ and $\sigma_w(i,j)$ are the mean and standard deviation in a local window of size $w \times w$ at (i,j).

Orientation and frequency estimation: We use Gabor filters [85] to calculate the orientation and frequency of these lines. Gabor filter is a band-pass filter which has frequency selective and orientation selective properties. The captured image is convolved with a bank of even symmetric Gabor filters at equally spaced angular directions and at multiple frequencies. The reason behind convolving with even-symmetric Gabor filter rather than complex Gabor filter is that they respond high to ridge like structure for the same sinusoid. The image is divided into blocks of considerable size. Along each direction, in each of the blocks, we select that frequency for which the sum of Gabor filter response is the maximum. At the next level we select two directions that has responded the maximum for any frequency based on the constraint that these two directions are at least some angle apart. To speed up the whole process, we first identify the line direction using a limited number of filters, and then refine the frequency estimate in the two orientations.

Identifying high-pixels: For each block, the exact orientations of lines and their frequencies are now available. To separate out the parallel lines (treated as a sinusoidal signal in one direction), we calculate the local phase at each point of the sinusoid and the pixels with phase in $[-\pi/2, +\pi/2]$ are set to be belonging to a line. This way we separate different lines distinctively. Note that the local intensity peaks of the projected pixels correspond to the local phase value of 0. The same process is repeated for the orthogonal set of parallel lines. Intersection of these two line maps distinctively isolate the high pixels. Local phase is computed by convolving the image $i(x,y)$ with a Gabor wavelet,$g(x,y,f,\theta)$, having frequency f and orientation θ,

$$\phi(x,y,f,\theta) = \arg[i(x,y) * g(x,y,f,\theta)], \tag{5.4}$$

where arg[] is a complex argument in $(\pi,\pi]$. As the phase information is independent of the magnitude, the limits of threshold are fixed in advance.

Center-high-pixel locations: The high pixels calculated occur as a set of connected components, where each connected component correspond to one pixel of the display image (see Fig. 5.1). The center-high-pixel of each of these connected components is calculated by taking the mean of the co-ordinate locations.

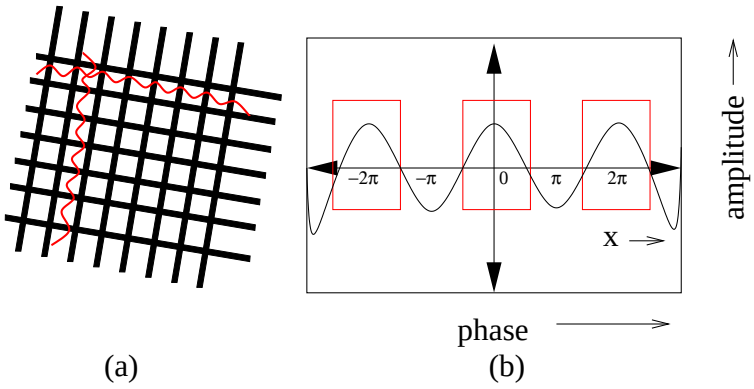


Figure 5.3: (a) captured image can be seen as super-imposition of two sets of approximate orthogonal directional sinusoids; (b) high-pixels are robustly extracted by thresholding on phase information instead of amplitude because of robustness against noise, intensity, blur

5.4 Captured Image Enhancements

The pixelation and defocus artifacts are removed by the process above from the captured image. A mechanism is described to improve the quality of captured images by virtualizing a high-resolution projector. Before that, we build the 8 neighborhood for each of the center-high-pixel. Instead of finding the eight closest points, we compute the 4 neighborhoods by utilizing the frequency and direction information computed in the previous section and then expand it to form the 8 neighborhood. For example, the north neighbor of the pixel in east would correspond to the north-east neighbor of the current pixel.

5.4.1 Depixelation and Deblurring

The intensity values of all the center-high-pixels is computed by taking the mean of the intensity values in a 3×3 window at that location. The pixels of the captured image are grouped around the center-high-pixels in the form of quadrilaterals. These quadrilaterals refer to the projected pixels, which have been captured when the projector does not have pixelation and blur artifacts. They are computed by utilizing the center-high-pixel locations of neighbors for consistency. Each pixel in the quadrilateral is then assigned the value of the corresponding center-high-pixel. For textured scenes, the value of each of the pixel of the captured image is re-initialized to the weighted mean of its original value and the corresponding center-high-pixel location value.

5.4.2 Virtualizing a High Resolution Projector

The high resolution virtualization is defined in Section 5.2. After calculating the 8 neighborhood for each of the center-high-pixel, we compute the location of new pixels to be embedded such that they lie uniformly with the neighborhood. The intensity values of the new projected pixels can be

assigned either by using various interpolation techniques or by using the one pass learning based super-resolution techniques. The restoration process is same as described before.

5.5 Experiments and Results

The proposed algorithm was tested on the images captured of scenes having different characteristics. The projector used was a HITACHI CP-S210, and the images were captured using a CANON EOS 350D camera. The aperture of the camera is set to the minimum. Gabor filters at uniform angles in eight directions were used at 3 different frequencies (0.17, 0.2, 0.23), for the initial estimation of line orientation and frequency. For refining the frequency estimates, Gabor filters with frequencies in the range $[0.16, 0.24]$ at an interval of 0.01 were used. We now describe the results with the display image projected on scenes of various properties.

5.5.1 Planar Textureless Scene

Images were captured under three different settings. In the first setting, the pixelation artifacts of the projector is clearly visible (Fig. 5.4(a)) and the projector is in focus. The images were restored at very high quality. The restored image is comparable with the display image at pixel level but some artifacts due to color transformations and brightness values of the projector can be seen. The image restored with high resolution projector virtualization is smoother than the original restored image. Fig. 5.4(g) and Fig. 5.4(j) shows images that are captured with increasing amounts of projector blur. The pixel location were calculated with high accuracy for the lower blur case and is quite satisfactory for the case with severe blur. When the blur is severe, we can observe color noise, because of the mixing of illumination of neighboring projected pixels with the low-pixels. Color noise artifacts are reduced considerably in the restored image (Fig. 5.4(l)).

5.5.2 Planar Textured scene

In the second experiment, a planar object with strong surface texture ((Fig. 5.5(1a)) is used. The center-high-pixel locations are calculated at very high accuracy even in presence of the strong scene texture. By using the mechanism mentioned in section 5.4, the background information in the image is retained. With high resolution projector virtualization the quality of the restored image is further improved. ((Fig. 5.5(1f)).

5.5.3 3D objects

In the third case, a textureless 3D object, Fig. 5.5(2b) was chosen as the scene. The restored captured image looks blocky due to smoothing of fine details on the object. However, the restored image is much better than the captured one as the pixelation artifacts are removed. Fig. 5.5(3b) shows a scene with two 3D objects, placed 6 inches apart. The texture of the captured image also get pixelated. Again with high-resolution projector virtualization the captured image is restored at a higher quality.

5.6 Discussions

The proposed algorithm for removing projector artifacts in the captured scene is robust against noise. The algorithm fails to detect the high-pixels in regions where the projected pixels are dark. This is because of the presence of high noise at very low intensity values. However, this does

not affect the quality of the restored image, even in the case of high-resolution image generation. Virtualization of high-resolution projector was also very useful to obtain the restored image at a higher quality. As the intensity value of the new pixels were calculated using the neighbouring pixels and the background region, most of the background information is retained. In restoration with textured scenes, the background sometimes becomes blocky and but with high resolution projector virtualization most of these problems are almost removed. Complete restoration in case of highly blurred projected image was not possible although the color noise was removed. We notice that the pixelation artifacts are clearly seen in a wide focus range of the projector. Our algorithm takes around 45 seconds for restoring a complete image of size 500×500 in matlab.

5.7 Summary

We have addressed the problem of restoring the captured image with projector artifacts. The solution proposed for the localization of center-high-pixels is very robust to noise and blur. We have tested our algorithm with scenes having different characteristics. Identification of center-high-pixel location is not only useful for restoring the captured image but also for other applications such as calculating scene parameters, extraction of features in the captured scene, etc. Overall, with increase in usage of projector camera systems, restoration of images with projected texture would find a wide variety of applications.

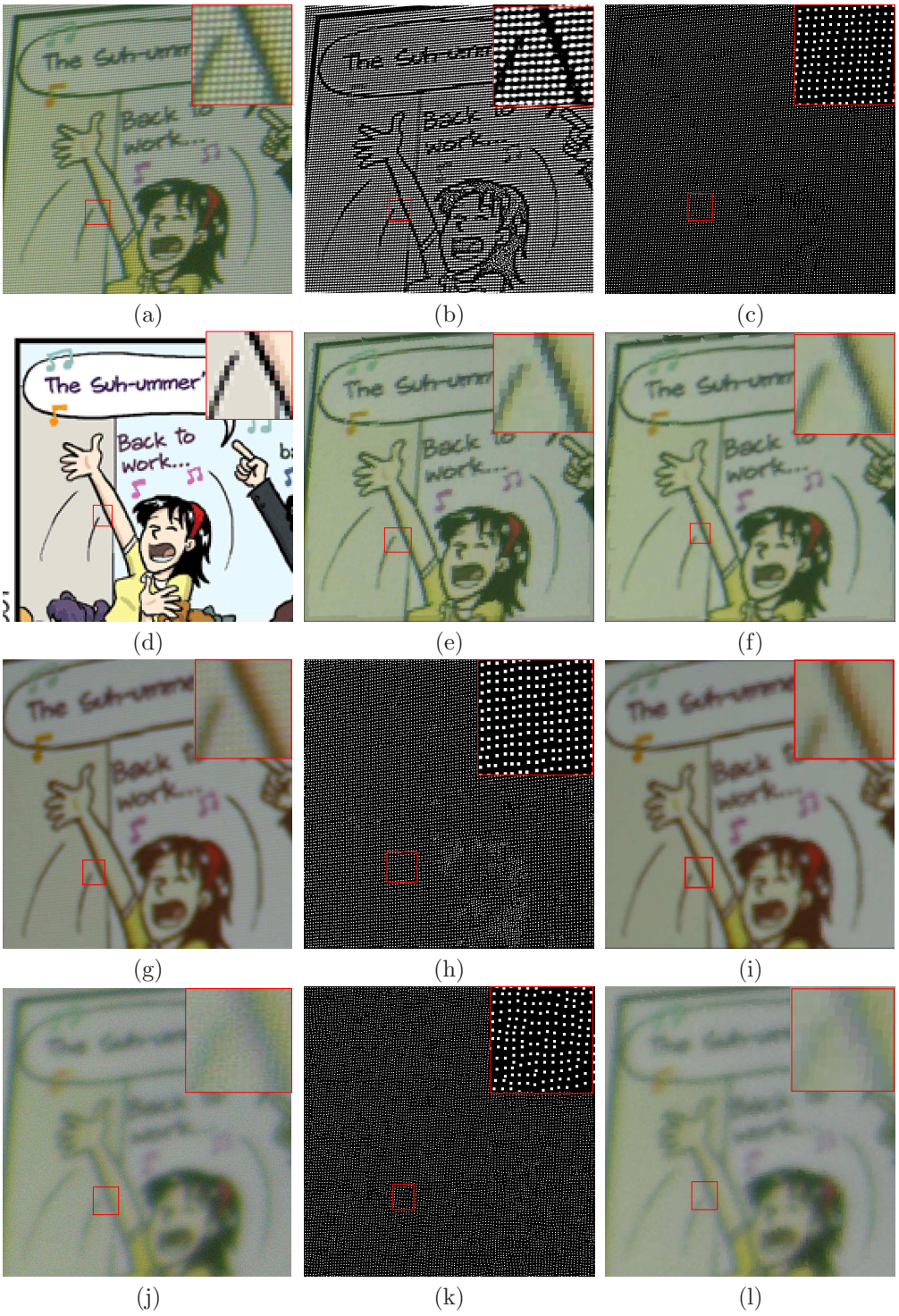


Figure 5.4: *(all images to be zoomed in)* (a) captured image (patch) with pixelation artifacts; (b) local contrast normalized image; (c) center-high-pixel location map; (d) display image patch; (e) pixelation artifacts restored image; (f) high-resolution projector virtualized image; (g) captured image, projected using a different projector with slight defocus; (h) center-high-pixel location map of image in (g); (i) restored image (j) captured image with high blurring artifacts; (k) center-high-pixel location map of the image in (j); (l) defocus artifact restored image

Figure 5.5: *(all images to be zoomed in)* (a) composite captured image(patch); (b) background object; (c) display image patch; (d) center-high-pixel map; (e) restored image; (f) high-resolution projector virtualized image

Chapter 6

Conclusions & Future Work

6.1 Conclusions

In summation, we address some of the problems towards obtaining high quality image reconstruction. Highly accurate image registration for super-resolution is a well studied problem. In this thesis, we studied this problem in real world scenarios, where images are captured from a cell-phone or low quality cameras in presence of varying illumination and blur. It is under these critical scenarios where we require high-resolution image reconstruction the most. This thesis contain two newly proposed problems as well. In literature, limits on super-resolution magnification are well defined. In some critical situations (e.g. understanding scene through a camera mounted on an automated mobile platform) detailed information of a scene need to be known before any successful processing and/or decision making. We looked into how to capture sufficient scene information such that further magnification enhancements can be done using any off the shelf single frame super-resolution algorithms. The other problem deals with capture of scene that are illuminated by a LCD projector. Image projected on a surface or a sheet suffers from pixelation and defocus artifacts. Restoring the captured image of such a scene is a problem of wide interest. If the access to power point slides or movie played through a projector is not easily available, is it still possible to capture high quality images. Such an application is highly useful during a conference talks. Projection of dots on any object also enable us to capture object shape.

For super-resolution of images, we proposed an algorithm for image registration, which is robust in presence of noise, non-uniform blur and illumination. We have shown that our algorithm based on local phase is independent of blur and illumination artifacts. Our approach is also correspondenceless, and hence there is no need of calculating features, explicitly. We have proven the convergence of the algorithm, even when it is impossible to identify the exact frequency of the underlying signal. Our algorithm is extensible to any general class of image registration, which is not the case with other transform domain approaches though both provides similar robustness.

Optimal zoom imaging for capturing sufficient scene details has been proposed in a MAP-MRF framework. MRF allows modeling of contextual constraints. With such constraints, certain amount of regularization is brought in and the errors in resolution-front values are suppressed. The places where the underlying patch information is insufficient to predict high zoom values, context information plays a significant role. Selecting the right zoom value can as well be proposed as a high-level vision problem, where a particular object is zoomed in at a pre-defined value. Proposing it as a low-level vision problem provides high degree of generalizability for a variety of scenes. Low computational speed is one of the key issues. But with additional constraints, significant speed up has been achieved. Camera shakes introduce blur in images and deteriorates the zoom prediction.

But it can be controlled in autonomous environments. We envision that such a functionality would be introduced in consumer cameras.

The solution proposed for restoring the captured image with projector artifacts, and detection of projected pixels is very robust to noise and blur. The algorithm fails to detect the high-pixels accurately in regions where the projected pixels are dark because of high noise at very low intensity values. However, this does not affect the quality of the restored image, even for the high-resolution image generation. Virtualization of high-resolution projector is useful to obtain the restored image at a higher quality. Identification of center-high-pixel location is not only useful for restoring the captured image but also for other applications such as calculating scene parameters, extraction of features in the captured scene, etc. Overall, with increase in usage of projector camera systems, restoration of images with projected texture would find a wide variety of applications.

6.2 Future Work and Scope

In the first chapter, an overview and references of various different methods towards obtaining high quality images are provided. Several common problems and challenges are also identified and listed to motivate research in this direction. In this section, certain problems and extensions to the current work are discussed.

- **Optimal Zoom Imaging for Super-Resolution at Mid-Level Vision :** Low-level vision tasks include primitive processing, estimation and enhancements. The problem of selecting the right zoom of the camera from the perspective of super-resolution has been proposed at the low-level vision. All parts of the scene need not be of high importance. For regions of the scene that are of lower importance can be captured with a large field of view in one image. This issue is introduced in chapter 4.

  The future work and scope is towards formulating the task as a mid-level vision problem. Mid-level vision tasks involve fitting parameters to data. Generalizability of the problem can be sacrificed in return for huge speed and performance gain. Additionally, a practical system need to be developed that can rank different regions in the scene based on importance.

- **Patch Matching:** To simplify the solving of Markov network for low-level vision problem, image is divided into small patches. For each patch in the image, a small number of patches are selected from the database that minimizes a distance metric. There are two problems worth considering:

  - Approximate Nearest Neighbor (ANN) data-structure proposed by Arya *et al.* [90] is computationally efficient to compute nearest neighbor in high dimensions. Though there is little theoretical scope, but problem-specific improvements would highly speed up nearest neighbor matching.
  - Commonly used patch matching algorithms minimizes a distance metric (e.g. L_k norm). L_1 norm equally penalizes deviation among all the values of a vector whereas L_∞ norm penalizes only the largest deviation. The problem with low order norms is that it does not allow smaller variations in patch intensity structures, whereas high order norms only penalizes large variations leaving small variations almost untouched. L_∞ norm measurements are highly sensitive to outliers.

    For learning based super-resolution algorithms the above metrics are appropriate. During experiments we noticed that intensity based matching metrics do not match patches

at structural or semantic level. Semantic or template matching exists at image level but not at patch level.

- **Registration of multiple degraded images in presence of various other artifacts:** In high quality image reconstruction process, various images are usually captured to calculate the inverse, robustly and reduce the size of the solution space. Inaccurate alignment of various images adversely affect the high-resolution image computation. In chapter 3, we proposed an algorithm for image registration in presence of noise, non-uniform illumination and blur. Registration parameters should be highly accurate and computed at sub-pixel level. Highly accurate image registration algorithms are desirable under heavy degradations in presence of environmental attenuation, occlusions, chromatic aberrations, etc.

- **Robust Image Features in Highly Blurred Images:** In chapter 3, we theoretically showed that phase information remains invariant to a class of blur kernel which are real and even. However, magnitude information degrades severely in presence of extreme blur. Magnitude information is vital to confirm the existence of any local spatial frequency. Approximate information on amount of blur present may probably help in determining the degree of magnitude degradation and hence can be used to extract features in highly blurred images. Computing robust features for highly blurred images is a problem of wide impact, which need to be addressed.

Related Publications

- Himanshu Arora and Anoop M. Namboodiri, "How much zoom is the right zoom from the perspective of Super-Resolution ? ", *Sixth Indian Conference on Vision, Graphics and Image Processing (***ICVGIP 2008***)*, Bhubaneshwar, India, pp 142–149, Dec. 2008.

- Himanshu Arora and Anoop M. Namboodiri, "Projected Pixel Localization and Artifact Removal in Captured Images", *IEEE Region 10 Conference (***TENCON 2008***)*, Hyderabad, India, Nov. 2008

- Himanshu Arora, Anoop M. Namboodiri, and C.V. Jawahar, "Robust Image Registration with Illumination, Blur and Noise Variations for Super-Resolution", In Proceedings of the *33rd IEEE International Conference on Acoustics, Speech and Signal Processing (***ICASSP 2008***)*, Las Vegas, Nevada, USA, pp 1301–1304 , March 31 - April 4 2008

- Himanshu Arora, Anoop M. Namboodiri, and C.V. Jawahar, "Accurate Image Registration from Local Phase Information" *Proceedings of Thirteenth National Conference on Communications (***NCC 2007***)*, IIT Kanpur, India, pp 37–41, Jan. 2007

Bibliography

[1] Wikipedia, ,". http://en.wikipedia.org/wiki/History_of_the_camera. 1

[2] R. Ng, M. Levoy, M. Brdif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Tech. Rep. CSTR 2005-02, Stanford University Computer Science, 2005. 2, 8

[3] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Transactions on Graphics (SIGGRAPH)*, 2008. ix, 4

[4] C. Liu, R. Szeliski, S. B. Kang, C. L. Zitnick, and W. T. Freeman, "Automatic estimation and removal of noise from a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 299–314, 2008. ix, 4, 5

[5] A. Gijsenij and T. Gevers, "Color constancy using natural image statistics," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA, June 2007, pp. 1–8. ix, 4, 8, 10

[6] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424. ix, 4, 7

[7] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, 2008. ix, 4, 7

[8] T. Janssen, "Computational image quality," Ph.D. dissertation, Technische Universiteit Eindhoven, Eindhoven, Netherland, 1999. 3

[9] H. Sheikh, A. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *Image Processing, IEEE Transactions on*, vol. 14, pp. 2117–2128, Dec. 2005. 4

[10] M. Motwani, M. Gadiya, R. Motwani, and J. Frederick C. Harris, "A survey of image denoising techniques," in *Proceedings of Global Signal Processing Expo and Conference*, Santa Clara Convention Center, Santa Clara, CA., September 2004. 5

[11] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006. 5, 6, 15

[12] M. Ghazel, "Adaptive fractal and wavelet image denoising," Ph.D. dissertation, University of Waterloo, 2004. 5

[13] J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *Image Processing, IEEE Transactions on*, vol. 12, pp. 1338–1351, Nov. 2003. 5

[14] S. Lyu and E. P. Simoncelli, "Statistical modeling of images with fields of gaussian scale mixtures," in *Advances in Neural Information Processing Systems 19*, Cambridge, MA, MIT Press, 2007, pp. 945–952. 5

[15] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *International Conference on Computer Vision*, 1998, pp. 839–846. 5

[16] L. Yuan, J. Sun, L. Quan, and H.-Y. Shum, "Image deblurring with blurred/noisy image pairs," in *ACM Transactions on Graphics*, New York, NY, USA, ACM, 2007, p. 1. 5, 6

[17] W. H. Richardson, "Bayesian-based iterative method of image restoration," *Journal of the Optical Society of America (1917-1983)*, vol. 62, pp. 55–59, 1972. 6

[18] D. Kundur and D. Hatzinakos, "Blind image deconvolution," *Signal Processing Magazine, IEEE*, vol. 13, pp. 43–64, May 1996. 6

[19] A. Krishnan, "Non-frontal imaging camera," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 1997. 6

[20] A. Levin, "Blind motion deblurring using image statistics," in *Advances in Neural Information Processing Systems 19*, Cambridge, MA, MIT Press, 2007, pp. 841–848. 6, 10

[21] J. Jia, "Single image motion deblurring using transparency," *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, June 2007. 6

[22] A. Rav-Acha and S. Peleg, "Two motion-blurred images are better than one," *Pattern Recogn. Lett.*, vol. 26, no. 3, pp. 311–317, 2005. 6

[23] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 787–794, 2006. 6

[24] S. Nayar and M. Ben-Ezra, "Motion-based motion deblurring," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 689–698, June 2004. 6

[25] A. Levin, P. Sand, T. S. Cho, F. Durand, and W. T. Freeman, "Motion-invariant photography," *ACM Transactions on Graphics*, August 2008. 6

[26] R. Raskar, A. Agrawal, and J. Tumblin, "Coded exposure photography: motion deblurring using fluttered shutter," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 795–804, 2006. 6

[27] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graph. Models Image Process.*, vol. 53, no. 3, pp. 231–239, 1991. 6, 30

[28] R. Schultz and R. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE T. Image Proces.*, vol. 5, no. 6, pp. 996–1011, 1996. 6, 30

[29] W.-Y. Zhao, "Super-resolution with significant illumination change," *International Conference on Image Processing*, vol. 3, pp. 1771–1774, 2004. 6, 30, 32, 36

[30] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *Image Processing, IEEE Transactions on*, vol. 6, no. 12, pp. 1646–1658, 1997. 6, 11, 30, 32, 55

[31] A. J. Patti and Y. Altunbasak, "Artifact reduction for set theoretic super resolution image reconstruction with edge adaptive constraints and higher-order interpolants.," *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 179–186, 2001. 6, 30

[32] W.-Y. Zhao and H. S. Sawhney, "Is super-resolution with optical flow feasible?," in *ECCV (1)*, 2002, pp. 599–613. 6, 30, 32, 36

[33] D. Capel and A. Zisserman, "Computer vision applied to super resolution," *IEEE Signal Processing Magazine*, vol. 20, pp. 75–86, May 2003. 6, 30, 36, 50

[34] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Robust shift and add approach to super-resolution," *In Proc. of the 2003 SPIE Conf. on Applications of Digital Signal and Image Processing*, pp. 121–130, Aug. 2003. 6, 30, 32, 50

[35] S. Borman and R. Stevenson, "Spatial resolution enhancement of low-resolution image sequences - a comprehensive review with directions for future research," *Technical Report, University of Notre Dame*, 1998. 6, 30, 36

[36] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *Signal Processing Magazine, IEEE*, vol. 20, no. 3, pp. 21–36, 2003. 6, 30, 35, 36, 55

[37] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 25–47, 2000. 6, 10, 11, 24, 28, 32, 55, 58, 59, 61, 62

[38] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comp. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, 2002. 6, 32

[39] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1167 – 1183, September 2002. (To Appear). 6, 11, 32, 55

[40] R. Fattal, "Image upsampling via imposed edge statistics," *ACM Trans. Graph.*, vol. 26, no. 3, p. 95, 2007. 6, 32, 64

[41] J. Sun, J. Sun, X. Zongben, and S. Heung-Yeung, "Image super-resolution using gradient profile prior," in *IEEE CVPR*, Jun 2008. 6, 32

[42] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, ACM Press/Addison-Wesley Publishing Co., 1997, pp. 369–378. 6

[43] T. Mitsunaga and S. Nayar, "Radiometric self calibration," *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1, pp. –380 Vol. 1, 1999. 6, 8

[44] M. Robertson, S. Borman, and R. Stevenson, "Estimation-theoretic approach to dynamic range enhancement using multiple exposures," *Journal of Electronic Imaging 12(2), 219 228 (April 2003).*, vol. 12, pp. 219–228, Apr. 2003. 6

[45] G. Ward, "Fast, robust image registration for compositing high dynamic range photographcs from hand-held exposures," *Journal of Graphics Tools*, vol. 8, no. 2, pp. 17–30, 2003. 6

[46] M. Aggarwal and N. Ahuja, "Split aperture imaging for high dynamic range," *Int. J. Comput. Vision*, vol. 58, no. 1, pp. 7–17, 2004. 6

[47] S. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, pp. 472–479 vol.1, 2000. 6

[48] M. Goesele, W. Heidrich, B. Hfflinger, G. Krawczyk, K. Myszkowski, and M. Trentacoste, "High dynamic range techniques in graphics: from acquisition to display," *Tutorial 7: Eurographics*, August 2005. 6

[49] Y. Zheng, S. Lin, and S. B. Kang, "Single-image vignetting correction," in *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, IEEE Computer Society, 2006, pp. 461–468. 7

[50] D. B. Goldman and J.-H. Chen, "Vignette and exposure calibration and compensation," in *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Washington, DC, USA, IEEE Computer Society, 2005, pp. 899–906. 7

[51] A. Litvinov and Y. Schechner, "Addressing radiometric nonidealities: a unified framework," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 52–59 vol. 2, June 2005. 7

[52] T. Boult and G. Wolberg, "Correcting chromatic aberrations using image warping," *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pp. 684–687, Jun 1992. 7

[53] S. B. Kang, "Automatic removal of chromatic aberration from a single image," *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, June 2007. 7

[54] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vision*, vol. 48, no. 3, pp. 233–254, 2002. 7

[55] K. Garg and S. Nayar, "Vision and Rain," *International Journal on Computer Vision*, pp. 1–25, Feb 2007. 7

[56] Y. Y. Schechner, S. G. Narasimhan, and S. K. Nayar, "Instant dehazing of images using polarization," *Computer Vision and Pattern Recognition (CVPR)*, vol. 1, p. 325, 2001. 7

[57] C. Ballester, V. Caselles, J. Verdera, M. Bertalmi, and G. Sapiro, "A variational model for filling-in gray level and color images," *ICCV*, vol. 01, p. 10, 2001. 7

[58] S. H. Kang, T. F. Chan, and S. Soatto, "Inpainting from multiple views," *3DPVT*, vol. 0, p. 622, 2002. 7

[59] J. Jia and C.-K. Tang, "Image repairing: robust image synthesis by adaptive nd tensor voting," *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 1, pp. I–643–I–650 vol.1, June 2003. 7

[60] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003. 7

[61] N. Komodakis and G. Tziritas, "Image completion using global optimization," in *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, IEEE Computer Society, 2006, pp. 442–452. 7

[62] J. Hays and A. A. Efros, "Scene completion using millions of photographs," in *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, New York, NY, USA, ACM, 2007, p. 4. 7

[63] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, 2004. 8

[64] L. Yatziv and G. Sapiro, "Fast image and video colorization using chrominance blending," *Image Processing, IEEE Transactions on*, vol. 15, pp. 1120–1129, May 2006. 8

[65] H. Siddiqui and C. Bouman, "Training-based color correction for camera phone images," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, pp. I–733–I–736, April 2007. 8

[66] E. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, pp. 108–128, December 1977. 8, 10

[67] G. Buchsbaum, "A spatial processor model for object color perception," *Franklin Inst.*, vol. 310, no. 1, pp. 1–26, 1980. 8, 10

[68] K. Barnard, V. Cardei, and B. Funt, "A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data," *Image Processing, IEEE Transactions on*, vol. 11, pp. 972–984, Sep 2002. 8

[69] S. Lin and L. Zhang, "Determining the radiometric response function from a single grayscale image," *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2, pp. 66–73, June 2005. 8

[70] T.-T. Ng, S.-F. Chang, and M.-P. Tsui, "Using geometry invariants for camera response function estimation," *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, June 2007. 8

[71] B. Wilburn, H. Xu, and Y. Matsushita, "Radiometric calibration using temporal irradiance mixtures," *Computer Vision and Pattern Recognition, 2008. CVPR '08. IEEE Conference on*, June 2008. 8

[72] F. Moreno-Noguer, P. N. Belhumeur, and S. K. Nayar, "Active refocusing of images and videos," in *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, New York, NY, USA, ACM, 2007, p. 67. 8

[73] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, October 2003. 9, 36, 38

[74] Z. Lin and H.-Y. Shum, "Fundamental limits of reconstruction-based superresolution algorithms under local translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 83–97, 2004. 9, 10, 11, 32, 55

[75] Z. Lin, J. He, X. Tang, and C. Tang, "Limits of learning-based superresolution algorithms," in *ICCV*, 2007, pp. 1–8. 10, 11, 33, 34, 55

[76] I. E. Abdou and N. J. Dusaussoy, "Survey of image quality measurements," in *ACM '86: Proceedings of 1986 ACM Fall joint computer conference*, Los Alamitos, CA, USA, IEEE Computer Society Press, 1986, pp. 71–78. 10

[77] T. D. Sanger, "Stereo disparity computation using Gabor filters," *Biological Cybernetics*, vol. 59, pp. 405–418, 1988. 11, 18, 22, 23, 37, 40

[78] T. Gautama and M. Van Hulle, "A phase-based approach to the estimation of the optical flow field using spatial filtering," *IEEE Trans. Neural Networks*, vol. 13, no. 5, pp. 1127–1136, 2002. 11, 37, 40

[79] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, "The office of the future: a unified approach to image-based modeling and spatially immersive displays," in *SIGGRAPH*, New York, NY, USA, 1998, pp. 179–188. 12, 71

[80] R. Raskar, G. Welch, K. Low, and D. Bandyopadhyay, "Shader lamps: Animating real objects with image-based illumination," in *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, London, UK, Springer-Verlag, 2001, pp. 89–102. 12, 71

[81] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals & systems (2nd ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996. 15, 46, 57

[82] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, pp. 529–541, May 1981. 17

[83] D. Gabor, "Theory of communication," *J. IEE*, vol. 93, pp. 429–457, 1946. 18, 19

[84] D. J. Fleet and A. D. Jepson, "Stability of phase information," *PAMI*, vol. 15, pp. 1253–1268, December 1993. 18, 23, 40, 47, 75

[85] J.-K. Kämäräinen, "Feature extraction using Gabor filters," Ph.D. dissertation, Lappeenranta Univ. of Technology, 2003. 18, 40, 75

[86] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*. American Mathematical Society, 1980. 24

[87] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *PAMI*, vol. 6, pp. 721–741, November 1984. 24, 27

[88] R. Chellappa and A. Jain, *Markov Random Fields: Theory and Applications*. Academic Press, 1993. 24

[89] W. Weidlich, "The statistical description of polarization phenomenon in society," *Br. J. Math Statistic Psychol*, pp. 251–266, 1971. 24

[90] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM*, vol. 45, no. 6, pp. 891–923, 1998. 28, 61, 82

[91] R. S. Tsai and T. S. Huang, "Multiframe image resotration and registration," in *Advances in Computer Vision and Image Processing, JAI Press Inc.*, vol. 1, 1984, pp. 317–339. 29

[92] A. Tekalp, M. Ozkan, and M. Sezan, "High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration," *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 3, pp. 169–172, Mar 1992. 30

[93] S. Kim and W.-Y. Su, "Recursive high-resolution reconstruction of blurred multiframe images," *Image Processing, IEEE Transactions on*, vol. 2, pp. 534–539, Oct 1993. 30

[94] N. K. Bose, H. C. Kim, and H. M. Valenzuela, "Recursive total least squares algorithm for image reconstruction from noisy, undersampled frames," *Multidimensional Syst. Signal Process.*, vol. 4, no. 3, pp. 253–268, 1993. 30

[95] L. Ibanez, W. Schroeder, L. Ng, and J. Cates, *The ITK Software Guide*. Insight Software Consortium, 2005. http://www.itk.org. 36, 38, 47, 50

[96] D. Robinson, S. Farsiu, and P. Milanfar, "Optimal registration of aliased images using variable projection with applications to super-resolution," *The Computer Journal*, April 2007. 36, 39

[97] P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A Frequency Domain Approach to Registration of Aliased Images with Application to Super-Resolution," *EURASIP Journal on Applied Signal Processing (special issue on Super-resolution)*, vol. 2006, pp. Article ID 71459, 14 pages, 2006. 36

[98] 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 Transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997. 36, 39

[99] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000. 37, 38

[100] L. G. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, no. 4, pp. 325–376, 1992. 38

[101] A. Agarwal, C. V. Jawahar, and P. J. Narayanan, "A survey of planar homography estimation techniques," *IIIT Technical Report*, vol. IIIT TR/2005/12, June 2005. 38

[102] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981. 38, 47

[103] C. D. Kuglin and D. C. Hines, "The phase correlation image alignment method," in *Int. Conf. Cybernet. Society*, New York, USA, 1975, pp. 163–165. 39

[104] H. Foroosh, J. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *Image Processing, IEEE Transactions on*, vol. 11, pp. 188–200, Mar 2002. 39

[105] B. S. Reddy and B. N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration.," *IEEE Transactions on Image Processing*, vol. 5, no. 8, pp. 1266–1271, 1996. 39

[106] X. Guo, Z. Xu, Y. Lu, and Y. Pang, "An application of fourier-mellin transform in image registration," in *CIT '05: Proceedings of the The Fifth International Conference on Computer and Information Technology*, Washington, DC, USA, IEEE Computer Society, 2005, pp. 619–623. 39, 47

[107] M. P. Kumar, S. Kuthirumunal, C. V. Jawahar, and P. J. Narayanan, "Planar homography from fourier domain representation," in *Proceedings of SPCOM*, Dec 2004, pp. 560–564. 39

[108] M. K. Ng, J. Koo, and N. K. Bose, "Constrained total least-squares computations for high-resolution image reconstruction with multisensors," *International Journal of Imaging Systems and Technology*, vol. 12, no. 1, pp. 35–42, 2002. 39

[109] E. S. Lee and M. G. Kang, "Regularized adaptive high-resolution image reconstruction considering inaccurate subpixel registration.," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 826–837, 2003. 39

[110] O. Arandjelovic and R. Cipolla, "A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution," in *ICCV*, 2007. 55

[111] J. Kopf, M. Uyttendaele, O. Deussen, and M. F. Cohen, "Capturing and viewing gigapixel images," in *SIGGRAPH '07: ACM SIGGRAPH 2007 papers*, New York, NY, USA, ACM, 2007, p. 93. 56, 69

[112] T. Hashimoto, M. Ikemura, K. Kimura, Y. Hata, K. Hayashi, H. Ootsuka, and M. Nakanishi, "Camera having an auto zoom function," *US Patent No. 5291233*, 1994. 57

[113] K. Aoyama, "Auto-zoom camera," *US Patent No. 5604562*, 1997. 57

[114] B. Tordoff and D. Murray, "Resolution vs. tracking error: zoom as a gain controller," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Madison, Wisconsin*, IEEE Computer Society Press, June 2003. 57

[115] J. E. Bollman, R. L. Rao, D. L. Venable, and R. Eschbach, "Automatic image cropping," *US Patent No. 5978519*, 1999. 57

[116] J. Luo, "Automatically producing an image of a portion of a photographic image," *US Patent No. 6654507*, 2003. 57

[117] J. Luo and R. T. Gray, "Method for automatically creating cropped and zoomed versions of photographic images," *US Patent No. 6654506*, 2003. 57

[118] R. Jin, Y. Qi, and A. Hauptmann, "A probabilistic model for camera zoom detection," in *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 3*, Washington, DC, USA, IEEE Computer Society, 2002, p. 30859. 57

[119] A. Belahmidi and F. Guichard, "A partial differential equation approach to image zoom," *Image Processing, 2004. ICIP '04. 2004 International Conference on*, vol. 1, pp. 649–652, Oct. 2004. 57

[120] R. Willson, "Modeling and calibration of automated zoom lenses," in *Proceedings of the SPIE No. 2350: Videometrics III*, October 1994, pp. 170 – 186. 62, 63

[121] X. Li and M. Orchard, "New edge-directed interpolation," *Image Processing, IEEE Transactions on*, vol. 10, no. 10, pp. 1521–1527, Oct 2001. 64

[122] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006. 65

[123] S. Kumar and M. Hebert, "Man-made structure detection in natural images using a causal multiscale random field," in *in proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2003, pp. 119–126. 67

[124] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz, "Spacetime stereo: A unifying framework for depth from triangulation," *IEEE PAMI*, vol. 27, pp. 296–302, Feb. 2005. 71

[125] Y. Kakehi, M. Iida, T. Naemura, Y. Shirai, M. Matsushita, and T. Ohguro, "Lumisight table: An interactive view-dependent tabletop display," *IEEE Comp. Graph. App.*, vol. 25, no. 1, pp. 48–53, 2005. 71

[126] M. Grossberg, H. Peri, S. Nayar, and P. Belhumeur, "Making One Object Look Like Another: Controlling Appearance using a Projector-Camera System," in *IEEE CVPR*, vol. I, Jun 2004, pp. 452–459. 71, 73

[127] A. Emmerling and O. Bimber, "Multifocal projection: A multiprojector technique for increasing focal depth," *IEEE TVCG*, vol. 12, no. 4, pp. 658–667, 2006. 71, 72

[128] N. Damera-Venkata and N. L. Chang, "Realizing super-resolution with superimposed projection," in *PROCAMS'2007*, Minnesota, USA, June 2007. 71, 72

[129] J. Pages, C. Collewet, F. Chaumette, and J. Salvi, "A camera-projector system for robot positioning by visual servoing," in *PROCAMS'2006*, New York, USA, June 2006. 71

[130] R. Raskar, M. Brown, R. Yang, W. Chen, G. Welch, H. Towles, B. Seales, and H. Fuchs, "Multi-projector displays using camera-based registration," in *Proceedings of the 10th IEEE Visualization Conference*, Washington, DC, USA, 1999. 71

[131] L. Zhang and S. Nayar, "Projection defocus analysis for scene capture and image display," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 907–915, 2006. 72