Image Enhancement by Super-resolution, Focus Editing and Exposure Composition

ZHANG, Wei

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Electronic Engineering

The Chinese University of Hong Kong
August 2010

Dedication

To my beloved Tingting. None of this would be possible
without your love and support.

To my well-loved parents, my grandma and the entire family,
who are the reason to try and make the world a safer place.

Acknowledgments

This thesis would never be possible without without the support, guidance and love from so many people around me. I would like to acknowledge some of them here. Of course there are many others who have helped me along the way of my working towards doctoral degree, but it is not possible to list all of them here.

First, I would like to express my sincere gratitude to my thesis advisor, Prof. Wai-Kuan Cham, for his invaluable advice and generous support during my doctoral thesis. He provided me with the freedom to pursue my own interests and the guidance to stay on track. His words of encouragements and guidance have helped me through many obstacles in the past years. He also has spent great effort to help me improve the skills of writing papers and giving oral presentations. I learned an immense amount from his genuine motive and profound wisdom for pursuing scientific truth.

Over the past few years, I have been fortunate to have many wonderful teachers. I would like to thank Prof. Thierry Blu, Prof. Xiaogang Wang, Prof. King Ngi Ngan and Prof. Hung Tat Tsui, who are the faculty members of the Image and Video Processing (IVP) lab, for their valuable comments and illuminations on my research work. I am also grateful to our lab technician Yuk Chung Wong, for his timely help with computer maintenance.

I would like to thank my colleagues and friends in the IVP lab, Dr. Jian Yao, Dr. Hongliang Li, Dr. Yu Liu, Dr. Jie Dong, Dr. Zhenzhong Chen, Dr. Chun Man Mak, Dr. Zhenyu Wei, Dr. Jie Li, Dr. Yifeng Jiang, Dr. Zhijun Zhang, Dr. Xin Jin, Dr. Haiyan Shu, Dr. Fan Zhang, Dr. Bangsheng Chen, Deqing Sun, Chi Keung Fong, Chunhui Cui, Qian Zhang, Qiang Liu, Wanli Ouyang, Renqi Zhang, Lin Ma, Songnan Li and Cong Zhao. I really cherish the time we spent together in the lab.

I also want to thank some overseas researchers, Dr. Hong Chang, Dr. Soonmin Bae, Prof. Rob Fergus, Prof. Shree K. Nayar, Prof. Brian Curless, Prof. Michael Goesele, Wei Liu, Qi Shan, Orazio Gallo, Matteo Pedone, Mateusz Markowski, Dr.

Paul Debevec, Dr. Neel Joshi, and Dr. Amit Agrawal for their providing data, source codes or kindly answering some questions.

Last but not least, I owe a great deal of thanks to my parents, my grandma, my entire family and my friends. Their love and support has been immeasurable, and I would not be here without them.

Abstract

Although significant progress has been made in imaging devices during the past few decades, the photographs acquired by digital cameras are still far from perfection due to the physical limitations of hardware such as aperture, lens and sensor. This fact brings out the demand for study on image enhancement: a computational technique that aims to improve the interpretability or perception of information in photographs for human viewers. The work in this thesis mainly focuses on three tasks in image enhancement.

Firstly, since the camera sensor has limited resolution, the acquired images cannot capture the scene very detailedly. Hence, people often resort to a postprocessing technique called super-resolution (SR) to enhance the resolution of the captured images. In the first part of this thesis, two approaches are presented to address the challenging single image SR problem, which is to recover a high-resolution (HR) image from one low-resolution (LR) input. Specifically, a novel learning-based framework is designed specifically for face image SR task from the perspective of DCT domain. In addition, an efficient two-step scheme is developed to super-resolve generic image by exploiting the salient edges of the input LR image.

Secondly, due to the limitation of lens and aperture, some cameras cannot produce pleasant photographs with desired focus setting. For example, portrait photography that requires shallow depth of field (DOF) is not allowed when using the compact point-and-shoot cameras. In the second part of this thesis, a new and complete postprocessing-based focus editing system that is able to handle the tasks of focus map estimation, image refocusing and defocusing, is developed to overcome the optical limitations and create different kinds of novel photos with desired focus setting from an imperfect photo.

Finally, since the radiance of the real world spans several orders of magnitude and its dynamic range dramatically exceeds the capability of the current digital cameras,

there often exist some undesirable over- or under-exposed regions in a photograph. The third part of this thesis aims at producing one great looking well-exposed image that is virtually impossible with a single exposure by compositing a stack of photos at different exposures taken with a conventional camera. Particularly, a simple but effective method is presented to describe how to take advantage of the gradient information to accomplish exposure composition in both static and dynamic scenes. Compared to conventional high dynamic range (HDR) imaging work, the proposed approach is quite appealing in practice since it is computationally efficient and easy to use, and frees users from the tedious radiometric calibration and tone mapping steps.

Throughout this work, extensive experiments on various real and synthetic image data are conducted to evaluate the performance of the proposed algorithms.

摘要

成像技術在過去的幾十年裏得到瞭很大的發展。但是由于受到照相機硬件包括光圈，鏡頭和傳感器的限制, 所拍攝的照片還有很多不盡人意的地方。這種現象促進瞭圖像增強（Image Enhancement）技術的研究，其目的是通過計算的手段來提高圖像承載信息的能力，增強觀者的視覺感受。本論文的目標正在于此，主要在以下三個方面上進行圖像的增強。

首先，由于照相機的傳感器分辨率有限，所以圖像往往不能很詳細的記錄場景信息。這樣，人們往往訴諸于一種稱爲超分辨率(Super-resolution)的後處理技術來提高所拍攝圖像的分辨率。本論文的第一個部分探討的正是這種技術，並提出瞭兩種不同的方法用于單幅圖像的超分辨問題，即從一幅低分別率的圖像上來恢複其對應的高分辨率圖像。其中，一個從離散余弦變換(DCT)域角度上的提出的基于學習的框架主要用來針對人臉圖像的超分辨率問題。而另一個框架是通過挖掘低分變率圖像中的強邊緣信息來解決一般圖像的超分辨率問題。

其次，由于受鏡頭和光圈的影響，一些照相機不能拍出令人滿意定焦效果的圖片。比如，肖像照片通常需要是窄景深(DOF)的，而常用的傻瓜相機(Point-and-shoot camera)卻很難滿足這個要求。因此，本論文的第二個部分提出瞭一個整套的基于後處理的圖像定焦編輯(Focus editing)系統，它可以完成定焦圖的計算，圖像重聚焦以及散焦等任務。這個系統可以幫助我們克服照相機硬件限制，把一幅定焦有瑕疵的圖像變成具有不同定焦效果的多個新圖像。

最後，由于真實世界中的光線跨度非常大，其動態範圍已經遠遠超出瞭照相機的亮度表示能力。所以，拍攝的照片經常會出現過曝光和欠曝光的問題。本論文的第三個部分研究的是用于解決此問題的曝光融合(Exposure composition)技術，即通過融合多副由常用照相機拍攝的不同曝光的照片來生成一副單次拍攝不可能得到的良好曝光圖像。尤其是，本論文描述瞭怎樣利用圖像的梯度信息來實現在靜態和動態場景中的曝光融合問題，所提方法簡單而有效。與傳統的高動態範圍(HDR)技術相比，此方法效率高，易于使用並且可以把用戶從繁瑣的相機映射標定和色調映射等步驟中解放出來。因此在實際應用中更加有吸引力。

爲瞭驗證所提算法的性能，本論文在大量真實和合成數據的基礎上進行瞭廣泛的實驗。

Publications and Patents

Journal Papers

- **Wei Zhang** and Wai-Kuen Cham, "Hallucinating Face in the DCT Domain," *IEEE Transactions on Image Processing (T-IP)*. (Accepted with minor revision)
- **Wei Zhang** and Wai-Kuen Cham, "Single Image Refocusing and Defocusing," submitted to *IEEE Transactions on Image Processing (T-IP)*. (In review)
- **Wei Zhang** and Wai-Kuen Cham, "Gradient-directed Multi-exposure Composition," submitted to *IEEE Transactions on Image Processing (T-IP)*. (In review)

Conference Papers

- **Wei Zhang** and Wai-Kuen Cham, "High Quality Artifact-free Super-resolution," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Hong Kong, Sep. 2010.
- **Wei Zhang** and Wai-Kuen Cham, "Gradient-directed Composition of Multi-exposure Images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, Jun. 2010.
- **Wei Zhang**, Jian Yao and Wai-Kuen Cham, "3D Modeling from Multiple Images," in *Proceedings of International Symposium on Neural Networks (ISNN), Lecture Notes in Computer Science (LNCS)*, Shanghai, China, Jun. 2010.
- **Wei Zhang** and Wai-Kuen Cham, "Single Image Focus Editing," in *Proceedings of IEEE International Conference on Computer Vision workshop on Color and Reflectance in Imaging and Computer Vision (ICCV-CRICV)*, Kyoto, Japan, Oct. 2009.

- **Wei Zhang** and Wai-Kuen Cham, "Image-based Modeling from Multiple Views," in *Proceedings of BJ-HK Doctoral Forum*, Beijing, China, Dec. 2008.
- **Wei Zhang** and Wai-Kuen Cham, "A Single Image based Blind Super-resolution Approach," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, San Diego, USA, Oct. 2008.
- **Wei Zhang** and Wai-Kuen Cham, "Learning-based Face Hallucination in DCT Domain," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Alaska, USA, Jun. 2008.

Patents

- **Wei Zhang** and Wai-Kuen Cham, Gradient-Directed Composition of Multi-Exposure Images, US Provisional Patent Application Number 61330237, filed on Apr. 30, 2010.
- **Wei Zhang** and Wai-Kuen Cham, Single Image Focus Editing, US Provisional Patent Application Number 61278182, filed on Oct. 2, 2009.

Nomenclature

Abbreviations

$1-D$	One-Dimensional
$2-D$	Two-Dimensional
$3-D$	Three-Dimensional
AC	Alternating Current
ACA	Accumulated Consistency Assessment
BP	Belief Propagation
$BPDN$	Basis Pursuit DeNoising
CCD	Charge-Coupled Device
$CMOS$	Complementary Metal-Oxide-Semiconductor
CPU	Central Processing Unit
CRF	Camera Response Function
DC	Direct Current
DCT	Discrete Cosine Transform
DOF	Depth of Field
$FERET$	Facial Recognition Technology
FFT	Fast Fourier Transform
GPU	Graphic Processing Unit
$HDTV$	High Definition Television
HDR	High Dynamic Range
HR	High-Resolution
HRI	High-Resolution Image
$JNBM$	Just Noticeable Blur Metric
LDR	Low Dynamic Range
LLE	Locally Linear Embedding
LR	Low-Resolution
LRI	Low-Resolution Image
MAP	*Maximum A Posterior*
ML	*Maximum Likelihood*
MRF	Markov Random Field
MSE	Mean Square Error
$NTSC$	National Television System Committee
PAL	Phase Alternate Line
PAR	Piecewise AutoRegressive
$POCS$	Projection Onto Convex Sets

PSF	Point Spread Function
RCA	Reference view guided Consistency Assessment
SAI	Soft-decision Adaptive Interpolation
SBD	Single-image Blind Deconvolution
SLR	Single-Lens Reflex
SR	Super-resolution
SSD	Sum of Squared Differences
$SSIM$	Structural SIMilarity
SVR	Support Vector Regression

Operators and Functions

$\mathbf{A}^\top$	Transpose of a matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of a square, invertible matrix $\mathbf{A}$
$\mathbf{a} \bullet \mathbf{b}$	Element-wise multiplication of two vectors $\mathbf{a} = (a_1, a_2, ..., a_m)$ and $\mathbf{b} = (b_1, b_2, ..., b_m)$, which is defined as: $\mathbf{a} \bullet \mathbf{b} = [a_1 b_1, a_2 b_2, ..., a_m b_m]^\top$
$f \otimes g$	Convolution of two functions f and g, which is defined as the integral of the product of the two functions after one is reversed and shifted. $(f \otimes g)(t) = \int f(\tau) g(t - \tau) d\tau$
$\|\mathbf{a}\|_p$	l_p-norm of a vector $\mathbf{a} = (a_1, a_2, ..., a_m)$, which is defined as: $\|\mathbf{a}\|_p = (\|a_1|^p + |a_2|^p + ... + |a_m|^p)^{1/p}$
$erf(\cdot)$	Error function: $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$
$U(\cdot)$	Unit step function
$g(x; \sigma)$	1-D Gaussian filter: $g(x; \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-x^2}{2\sigma^2})$
$g(x, y; \sigma)$	2-D Gaussian filter: $g(x, y; \sigma_f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(x^2+y^2)}{2\sigma^2})$

Notations

$\mathbf{A}$	A matrix
$B(u,v)$	The (u,v)th basis image
$\mathbf{b}$	A vector
b	Edge basis
$C(u,v)$	The (u,v)th DCT coefficient
c	Edge contrast
$D\downarrow$	Decimation matrix
f	A blurring filter
h	Camera's PSF
I_C	Composite image
I_D	Defocused image
I_F	Focused image
I_H	High-resolution output image
$\widehat{I_H}$	Intermediate high-resolution result
$\widetilde{I_H}$	Training high-resolution image
$\overline{I_H}$	Prefiltered training high-resolution image
I_H^{AC}	AC coefficients of the high-resolution image
I_H^{DC}	DC coefficients of the high-resolution image
I_L	Low-resolution input image
I_L^{AC}	AC coefficients of the low-resolution image
$\overline{I_L^{AC}}$	AC coefficients of the training high-resolution image
$\overline{I_H^{AC}}$	AC coefficients of the training high-resolution image
I^i	The i_{th} input image
I_m	Magnified image by interpolation
I_p	Predicted sharp image
I^{ref}	The reference image
I_x^i	Partial derivatives of image I^i along x direction
I_y^i	Partial derivatives of image I^i along y direction
K	Number of the input exposures
M_e	Binary edge mask
M_s	Binary smooth mask
n	Noise
T	An image block
$V(u,m)$	The m_{th} element of the u_{th} basis vector of the 1-D N-point DCT
$W_i(j)$	Weighting coefficient of $I_L^{AC}(i)$'s j_{th} nearest neighbor $\overline{I_L^{AC}}(j)$
w	Edge width
P	Prefilter
P^{-1}	Inverser of prefilter, i.e., postfilter

t_i	First order derivative filter: $t_1 = [1 \ -1]$ and $t_2 = [1 \ -1]^T$
x_0	Edge location
$\mathbf{x}$	A vector
ϵ	A small value such as 10^{-25} to avoid singularity
τ	Threshold that defines the well-exposed range
$\mathcal{N}(0, \sigma^2)$	Zero mean Gaussian distributions of variance σ^2
$I^i(x, y)$	Intensity of pixel located at (x, y) in the i_{th} image
$I_C(x, y)$	Intensity of pixel located at (x, y) in the composite image
$C_A^i(x, y)$	ACA consistency of pixel located at (x, y) in the i_{th} image
$C_R^i(x, y)$	RCA consistency of pixel located at (x, y) in the i_{th} image
$d_{ij}(x, y)$	Gradient direction change between pixels $I^i(x, y)$ and $I^j(x, y)$
$d_{i \to ref}(x, y)$	Gradient direction change between pixels $I^i(x, y)$ and $I^{ref}(x, y)$
$E^i(x, y)$	Exposure quality of pixel located at (x, y) in the i_{th} image
$S_A^i(x, y)$	ACA consistency score of pixel located at (x, y) in the i_{th} image
$S_R^i(x, y)$	RCA consistency score of pixel located at (x, y) in the i_{th} image
$V^i(x, y)$	Visibility of pixel located at (x, y) in the i_{th} image
$W^i(x, y)$	Weight of pixel located at (x, y) in the i_{th} image
$\nu^i(x, y)$	Gradient magnitude of pixel located at (x, y) in the i_{th} image
$\theta^i(x, y)$	Gradient direction of the pixel located at (x, y) in the i_{th} image
$\theta^{ref}(x, y)$	Gradient direction of the pixel located at (x, y) in the reference image

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation & Objectives

From ancient rock art to children's sidewalk drawings, we live in a visual world. According to a conception of visual experience that has been widely held by perceptual theorists, we open your eyes, and we enjoy a richly detailed picture-like experience of the world, one that represents the world in sharp focus, uniform detail and high resolution from the center out to the periphery. It can be called: snapshot conception of experience. Over the ages, human beings are trying to record the visual world constantly with different forms to keep this fascinating experience for ever. Among them, painting is the most long-history one, and it is still popular even today. The oldest painting can date back to 32,000 years ago. From then on, people begin to depict the creatures, domestic scenes, labor scenes, or nature by applying paint, pigment, color or other medium to a surface as walls, paper, canvas, wood, glass, lacquer, clay or concrete. However, painting is more like an artistic creation. It is inaccurate and time consuming. Amateur can hardly master it. Only the person with assiduous training can become a skilled painter and produce excellent works.

The advent of camera break the ice, and offers a quick and faithful depiction of things in life. By definition a camera is a object, with a lens, that captures incoming light and directs the light and results image towards film (optical camera) or the imaging device (digital camera). The first camera that is small and portable enough to be practical for photography was built by John Strognofe in 1685. Over the last hundreds of years, camera has come a long way (see Figure 1.1), from hulking to handheld, from monochrome to color, from optical to digital, from still image to video. Today, camera has become a necessity of our life. Especially, the development of computer, Internet and wireless communication greatly promoted the popularity of camera. Camera even

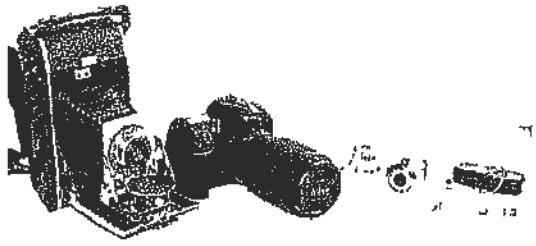

Figure 1.1· *Cameras from hulking to handheld, from optical to digital*

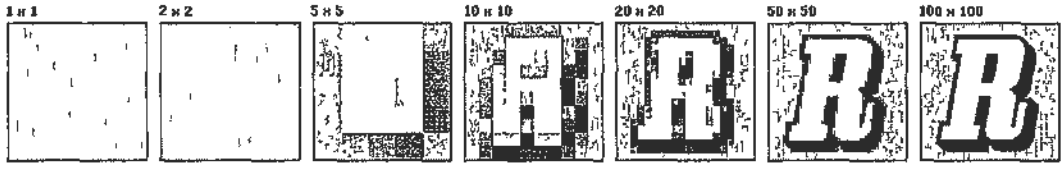

Figure 1.2· *Resolution illustration*

results in a new language that everyone can understand. The language is photography, through which we could recall a moment frozen in time and could share it with others.

Although nowaday camera is quite powerful, it is not a patch on our eyes and cannot capture what we see exactly. In most cases, the acquired photographs are still far from perfection due to the physical limitations of hardware such as aperture, lens and sensor. In this thesis, three aspects of the hardware limitations are addressed as follows:

First, camera sensor like CCD (Charge-Coupled Device) and CMOS (Complementary Metal-Oxide-Semiconductor) can only allow a limited number of spatial pixels, which results in a limited image resolution [Choi et al , 2004]. Although these sensors are suitable for most imaging applications, the current resolution level and consumer price will not satisfy the future demand. In most cases, images with high resolution are desired and often required. Especially, the recent popularity of HDTV (High Definition Television) brings out the need for resolution enhancement of NTSC and PAL formats. High resolution means that pixel density within an image is high, and therefore as shown in Figure 1 2, a high-resolution (HR) image can offer more scene details that may be critical in various applications.

Second, due to the limitation of lens and aperture, some cameras cannot produce pleasant photographs with desired depth of field (DOF). As illustrated in Figure 1.3,

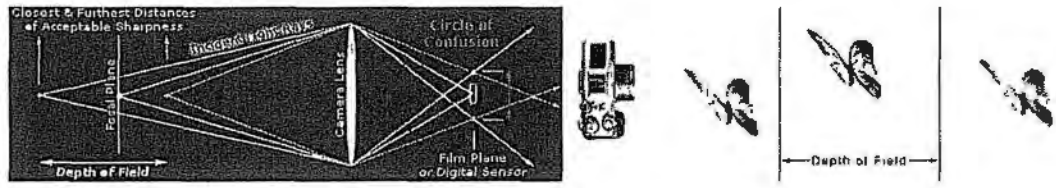

Figure 1.3 *DOF illustration The area within the DOF appears sharp, whilst the areas in front of and beyond the DOF appear blurry*

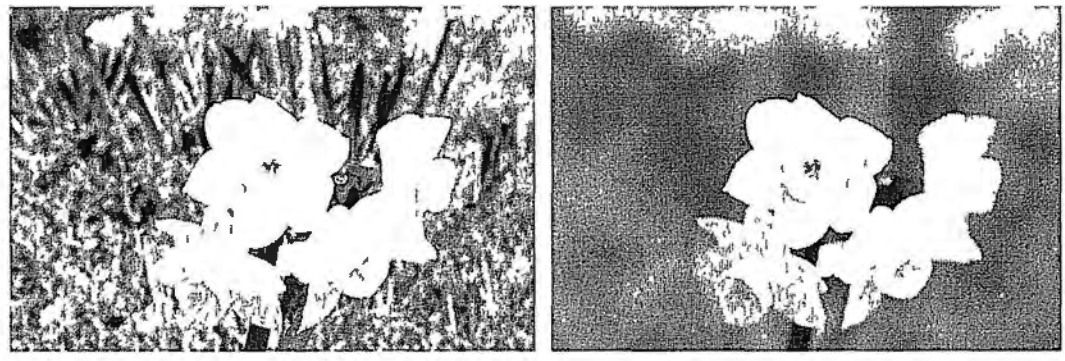

Figure 1.4 *The flowers captured with different apertures Left f/32; Right f/5 6*

DOF is the range of distance within the subject that is acceptably sharp. It is controlled by the lens aperture diameter specified as camera's f-number - the ratio of lens focal length to aperture diameter. As shown in Figure 1.1, reducing the aperture diameter (increasing the f-number) increases the DOF, while a larger aperture (smaller f-number) produces a shallower DOF In some cases, such as landscapes, it may be desirable to have the entire image sharp, and thus a large DOF is appropriate. In other cases, such as portrait, a small DOF is preferred for highlighting a subject while de-highlighting the foreground or background. However, a normal lens can only offer a limited DOF. As a result, one common complaint about cameras is that when using one sometimes it is hard to get nice out-of-focus background or all-focused objects.

Third, the light of real world spans several orders of magnitude and thus its dynamic range - the ratio between the brightest and darkest parts of the scene, dramatically exceeds the capability of camera sensor as shown in Figure 1 5. As a result, there often exist some undesirable over- or under-exposed regions in an image when the dynamic range of the latent scene is too high to be reproduced with a consumer camera at a single aperture and shutter speed as illustrated in Figure 1 6. In fact, not only cameras, but also display devices like most monitors and printers, do not have the capability of

Figure 1.5 *Dynamic range comparison of the scene of real world and camera*

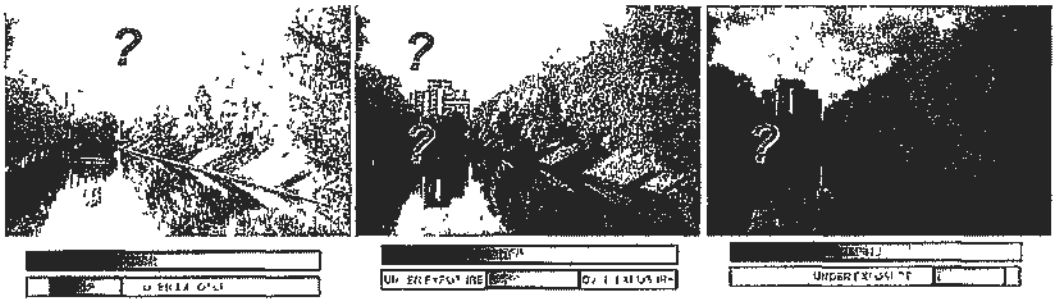

Figure 1.6 *Image capture with different exposure times Left: long exposure; Middle: medium exposure; Right: short exposure*

dealing with high dynamic range (HDR) content, either.

The objective of this thesis is to propose a series of image enhancement methods to remedy the aforementioned issues and make photography beyond the physical limitations possible. Firstly, two kinds of approaches are presented to address the resolution enhancement. One aimed at face images, the other is for generic images. Secondly, a focus editing system is presented which can yield images with different focus effects from an imperfect image. Finally, a simple but effective approach is presented to generate a tonemapped-like HDR image where all parts appear well-exposed by multi-exposure composition.

1.2 Previous Work

In this section, we give a brief overview of the existing work relevant to the three topics of this thesis: resolution enhancement, focus editing and HDR. More overviews of the related work will be presented in the Introduction section of each chapter.

Generally speaking, there are two ways to relieve the three camera limitations mentioned above. One is hardware solution which relies on the improvement of device physics and circuit technology. The other is through image enhancement which is the process of improving the quality of a digital image by manipulating the image with software. Therefore, in the next sections, the existing techniques of each topic are

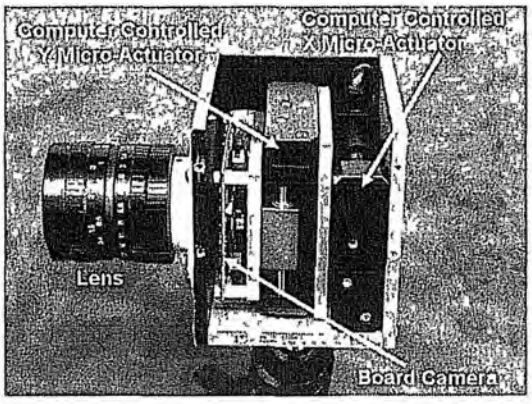

Figure 1.7 *The jitter camera prototype which is composed of a lens, a board camera and computer-controlled micro-actuators [Ben-Ezra et al. 2004, Ben-Ezra et al., 2005]*

divided into two classes: hardware solution and software solution.

1.2.1 Resolution Enhancement

Hardware Solution

As the pixel size of an image sensor ultimately determines the resolution of the captured image, the most direct solution to increase spatial resolution is to reduce the pixel size (i.e., increase the number of pixels per unit area) by sensor manufacturing techniques [Agranov et al., 2007; Fife et al., 2007]. The recent development of CCD and CMOS sensors has made HDTV production possible. However, the miniaturizing the pixel also reduces its light sensitivity and thus makes the sensor much more prone to shot noise that severely degrades the image quality. Therefore, there needs to be a limitation in the pixel size reduction, and the current image sensor technology has almost reached this level [Park et al., 2003; Choi et al., 2004].

Another approach for enhancing image resolution is to increase the chip size, which leads to an increase in capacitance [Komatsu et al., 1993]. Since large capacitance makes it difficult to speed up a charge transfer rate [Choi et al., 2004], this approach is considered ineffective The high cost for high precision optics and image sensors is also an important concern in many commercial applications regarding HR imaging. Apart from the above two ways, Elkhatib and Salama [Elkhatib and Salama, 2008a; Elkhatib and Salama, 2008b] recently presented a new system that can achieve a high resolution digital imaging independent of the pixel size by integrating a nanohole in each pixel of the image sensor.

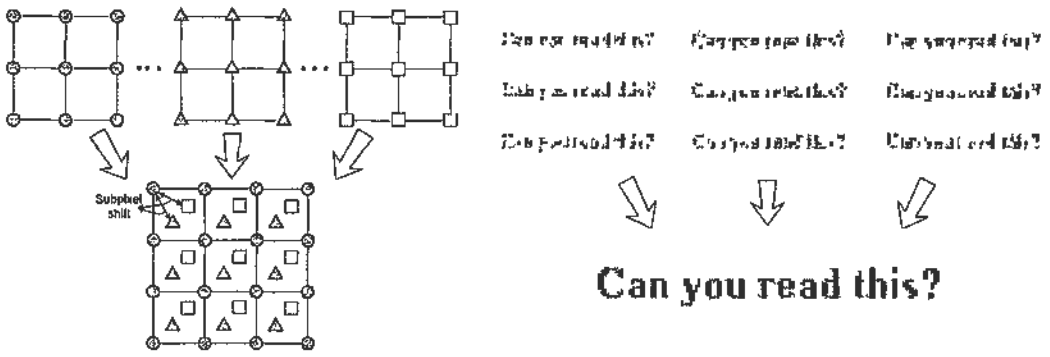

Figure 1.8 *Illustration of multi-image SR*

In addition, Ben-Ezra et al. [Ben-Ezra et al., 2004; Ben-Ezra et al., 2005] developed a novel camera called the "jitter camera" shown in Figure 1.7. The jitter camera produces shifts between consecutive video frames by shifting the video detector instantaneously and timing the shifts to occur between pixel integration periods. Then, the captured videos are further processed by an adaptive resolution enhancement algorithm to achieve resolution enhancement.

Software Solution

Image processing based software solution is a promising alternative to achieving resolution enhancement, since it costs less and the current imaging systems can be still utilized. This kind of solution is normally referred as super-resolution (SR) [Park et al., 2003] whose goal is to produce a HR image or a sequence of HR images from a low-resolution (LR) image or a sequence of LR images. It can be widely applied in various fields, including image compression, medical imaging, satellite imaging, and video applications. According to the number of input images, SR can be further categorized into two groups.

Multi-Image SR Most existing work was presented based on the premise of the availability of multiple LR images that capture different looks of the same scene. As shown in Figure 1.8, the LR images have different subpixel shifts from each other and each provides some new information that cannot be captured from the other [Park et al., 2003; Protter and Elad, 2009; Takeda et al., 2009]. Hence, after registering these images, a HR image can be obtained by combining all the new information together [Irani and Peleg, 1991; Irani and Peleg, 1993].

Based on the generative model of a camera which describes how a latent scene is transformed, filtered and sampled to form an observed image, a *maximum likelihood* (ML) estimator provides a simple way to reverse these degradations in order to estimate a HR representation of the scene [Capel, 2004]. However, SR is a well recognized ill-posed problem and a multiplicity of possible solutions exists given a set of observation images. Therefore, the ML estimator is extremely sensitive to noise in the observed images and to errors in registration. To solve this problem, it is necessary to introduce a prior model that imposes constraints on the form of the SR image, such as local smoothness, edge preservation, positivity and energy boundedness [Borman and Stevenson, 1998]. Thus, a *maximum a posterior* (MAP) estimator can be obtained and the solution is accepted only when it is both a good fit to the observations, and also has a high likelihood with respect to the prior model [Cheeseman et al., 1994; Schultz and Stevenson, 1996; Hardie et al., 1997; Capel, 2004; Protter et al., 2009]. Besides, projection onto convex sets (POCS) provides a convenient way for the inclusion of prior constraints and seek to solve the SR inverse problem iteratively using a full generative image model and arbitrary motion model [Eren et al., 1997; Patti et al., 1997; Elad and Feuer, 1999; Patti and Altunbasak, 2001]. All above methods can also be regarded as the reconstruction-based approach whose performance deteriorates as the magnification factor becomes a bit large [Lin and Shum, 2004].

Single-Image SR Recently, some efforts were made on inferring a HR image from a single LR input. Compared to the multi-image methods, single-image SR is more challenging and inherently limited by the amount of data available in an image.

The most popular way of enhancing image resolution in the graphics software is through interpolation-based methods such as Bilinear and Cubic B-Spline, but they suffer from severe blurring problem. There also existed some reconstruction-based single-image methods proposed with the aid of advanced prior models, where besides the global sparse priors [Rudin et al., 1992; Black and Sapiro, 1998; Tappen et al., 2003; Levin and Weiss, 2007; Roth and Black, 2009], local edge-based priors were developed to further preserve edge sharpness such as [Fattal, 2007; Sun et al., 2008; Dai et al., 2009].

Learning-based methods attract a lot of attention in the recent years. Usually,

the unknown HR image is inferred by making use of a training set directly or indirectly. In comparison with the interpolation-based methods and the reconstruction-based methods, learning-based methods can achieve higher magnification factor and better visual quality especially for single-image SR problem [Liu et al., 2008]. Baker and Kanade [Baker and Kanade, 2000; Baker and Kanade, 2002] presented a pioneering work on super-resolving face images based on a Bayesian formulation. Capel and Zisserman [Capel and Zisserman, 2001] extracted eigenfaces from a collection of training face images as a prior model to constrain and super-resolve LR face images. Freeman et al. [Freeman et al., 2000] proposed a well-known parametric MRF (Markov Random Field) based inference model to learn the statistics between the underlying scene and the observed image data. This framework was applied to the SR problem as well as other low-level vision problems. Such framework was extended and adopted by Sun et al. [Sun et al., 2003], Bishop et al. [Bishop et al., 2003], Wang et al. [Wang et al., 2005], Liu et al. [Liu et al., 2007], Ma et al. [Ma et al., 2008] and Xiong et al. [Xiong et al., 2009]. For instance, Liu et al. [Liu et al., 2007] developed a two-step statistical modeling approach for face hallucination which integrates a global parametric model and a local nonparametric model. Wang et al. [Wang et al., 2005] proposed a combination model that integrates the SR constraint and the patch based image co-occurrence constraint for the SR problem. But as analyzed in Lin et al. [Liu et al., 2008], the disadvantage of learning-based method is that the performance often replies on how well the input LR image matches the training samples. Therefore sufficient number of appropriate training samples should be provided to ensure the SR performance.

In addition, without using external data, Glasner et al. [Glasner et al., 2009] presented promising single-image SR results by integrating the reconstruction-based model and the learning-based model into an unified SR framework.

1.2.2 Focus Editing

Image focus editing is an interesting research topic and has received a lot of attention in recent years. Two tasks are mainly involved in this topic. One is image refocusing which is to recover the sharpness of the blurry defocused objects in an input image and generate a virtual all-focused image. The other is defocusing which is to blur an image and create defocus effects.

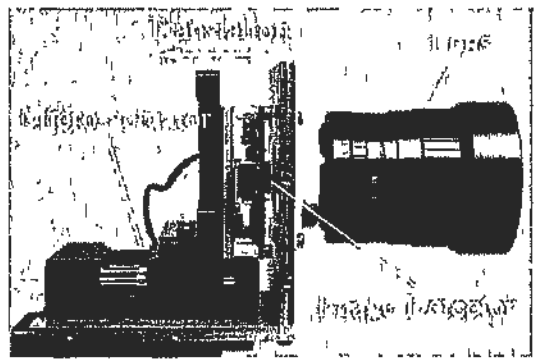

Figure 1.9 *A flexible DOF camera prototype which is composed of a lens and a detector mounted on a translation stage that is controlled by a micro-actuator [Nagahara et al., 2008]*

Hardware Solution

To tackle the image refocusing and defocusing problem, a large number of algorithms were presented with the aid of additional optical elements or devices that are used to capture more information about the target scene. For instance, Ng et al. [Ng et al., 2005] created a plenoptic camera by placing a microlens array between the sensor and the main lens. Thus synthetic images focused at different depths can be computed with the extra information captured by the microlens. Alternatively, one can place a positive lens array in front of the camera [Georgiev et al., 2007]. Veeraraghavan et al. [Veeraraghavan et al., 2007] used a cosine mask rather than lens array for computational improvement. A coded aperture is designed in [Levin et al., 2007] by inserting a patterned occluder within the aperture of the camera lens. Depth and the all-focused image can be recovered from a photograph taken by this modified camera. In [Moreno-Noguer et al., 2007], the depth map and the refocused image are produced with the aid of a grid of dots projected on the scene. Liang et al. [Liang et al., 2008] presented a new imaging system which can produce different focusing images by including a novel component called programmable aperture and two associated post-processing algorithms. Nagahara et al. [Nagahara et al., 2008] addressed the flexible DOF photography with a prototype camera (see Figure 1.9) that uses a micro-actuator to translate the detector along the optical axis during image integration

Software Solution

As an alternative, the other methods achieve focus editing by using only image processing. A natural way is to capture multiple images of the scene with different focus setting and then combine them to create synthesized images with new focus effects [Kubota et al., 2004; Kubota and Aizawa, 2005; Hasinoff and Kutulakos, 2007]. An early method was presented by Subbarao et al. [Subbarao et al., 1995] which showed that a focused image can be obtained from only two blurred images taken with different camera parameter settings. More recently, Yang et al. [Yang et al., 2008b; Yang and Schonfeld, 2010] presented a method that is able to produce in-focus image sequences by processing blurred videos captured with out-of-focus cameras. Hasinoff and Kutulakos [Hasinoff and Kutulakos, 2008; Hasinoff, 2008; Kutulakos and Hasinoff, 2009] proved that capturing a focal stack at the press of a button, instead of a single photo can boost significantly the optical performance of a conventional camera. Generally speaking, the focal stack photography has two performance advantages: first, it allows us to capture a given DOF much faster than one-shot photography, and second, it leads to higher signal-to-noise ratios when capturing wide DOF with a restricted exposure time.

Recently, the more challenging single-image-based work has attracted much attention. For example, the single image defocusing problem was addressed in [Yan et al., 2009] and [Bae and Durand, 2007]. Yan et al. [Yan et al., 2009] developed an interactive system for defocusing by constructing the depth information of an input image with user interaction. Bae and Durand [Bae and Durand, 2007] contributed at proposing an automatic focus map estimation method by estimating the edge blurriness with a brute-force fitting strategy. The defocusing there is handled with the aid of the lens blur tool in *Photoshop*. The method proposed by Bando and Nishita [Bando and Nishita, 2007] can tackle the single image refocusing task but it requires lots of user intervention to determine the blur kernel from a number of predefined candidates.

1.2.3 High Dynamic Range Imaging

Hardware Solution

To extend the dynamic range of conventional camera, some new HDR camera prototypes have been developed during the past years. Normally, this kind of methods require

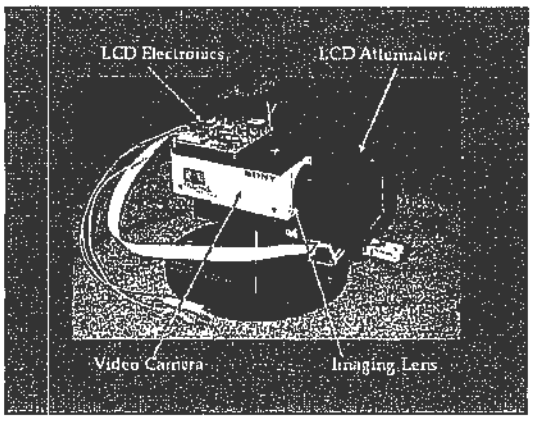

Figure 1.10: *A HDR camera prototype which is composed of a color video camera, an imaging lens, an LCD spatial attenuator and electronics to control the attenuator [Nayar and Branzoi, 2003].*

additional optical elements or devices to help the camera sensor record more dynamic range of the target scene. For example, some HDR camera prototypes such as [Saito, 1996; Kimura, 1998; Ikeda, 1998; Aggarwal and Ahuja, 2004] can split the incoming light to several detectors which have different exposures. Some methods like [Wen, 1989; Murakoshi, 1994; Street, 1998] were presented to achieve HDR imaging with a different CCD design where each detector cell includes two sensing elements of different sizes. When the detector is exposed to the scene, two measurements are made within each cell and they are combined on-chip before the image is read out. With only one image detector, some researchers attempted to give the pixels different exposures adaptively to the scene by using additional hardware such as a computational element that can measure the time each pixel takes to attain full potential well capacity [Brajovic and Kanade, 1996], an optical mask with a pattern of cells with different transparencies [Nayar and Mitsunaga, 2000] or a controllable liquid crystal light modulator whose transmittance can be varied [Nayar and Branzoi, 2003] (see Figure 1.10). Instead of direct pixel intensity measurements as output, Tumblin et al. [Tumblin et al., 2005] presented a rather new camera design that first measures the differences between adjacent pixel pairs and then quantize the obtained differences appropriately to capture the HDR scene.

However, compared to conventional camera, HDR camera is still unavailable to consumers and has three main limitations. First, it is expensive as some additional hardware is required. Second, it normally takes longer time to finish a shot. Third, it

has limited resolution and faces the challenge of scaling to high resolution while keeping fabrication costs under control.

Software Solution

The software solutions seek to produce a HDR image from a stack of images taken by a conventional camera with different exposure times. This kind of techniques can be referred as multi-exposure HDR, and can be furthered classified into two types according to whether the scene contains moving objects or not.

Static HDR The standard HDR technology prevalent in the current graphics software belongs to static HDR and require all objects stay stationary while capturing. It normally consists of two steps. First, recover the camera response function (CRF) and estimate the radiance maps from the multiple exposed images and their exposure settings [Debevec and Malik, 1997; Grossberg and Nayar, 2003]. Combine all radiance maps will result in a HDR image encoded specially to store the pixel values that span the whole tonal range of the real world scene. Second, since the commonly used display devices can only allow a low dynamic range (LDR), tone mapping is necessary to remap the HDR image to a LDR image [Durand and Dorsey, 2002; Fattal et al., 2002; Reinhard et al., 2002; Drago et al., 2003; Li et al., 2005; Shan et al., 2010]. As an alternative, the other kind of work attempted to produce the desired tonemapped-like HDR image directly by compositing the multiple exposures in the image domain [Goshtasby, 2005; Mertens et al., 2009; Shanmuganathan and Chaudhuri, 2009]. These methods skip the typical HDR process, and no intermediate HDR image needs to be generated. Therefore, they are more efficient and do not require tone mapping.

However, the major problem of above static methods is that the target scene is required to be completely still throughout the image capture. Any object movement in the exposure sequence can cause ghosting artifacts in the resulting image. This drawback severely affected the application of HDR in practice, since for most scenarios, it is hard to guarantee all objects involved stay stationary from one capture to the next. For instance, there often exist crowds of people moving around in tourist resorts. There are windblown trees in nature scenes.

Dynamic HDR Recently, lots of efforts have been made to address how to achieve ghost-free HDR imaging in dynamic scenes. In brief, they first detect motion regions explicitly or implicitly, and then combine all calibrated radiance maps without the pixels corrupted by moving objects to create an artifact-free HDR image. For instance, Kang et al. [Kang et al., 2003] proposed to compute the optical flow between successive frames and then warp pixels to create ghost-free HDR results. To find the pixels corrupted by moving objects, Reinhard et al. [Reinhard et al., 2005] proposed to threshold the variance map computed based on the irradiance variation of pixels over different exposures. Similarly, Jacobs et al. [Jacobs et al., 2008] applied a threshold on the entropy map, while Grosch [Grosch, 2006] applied a threshold on the error map estimated from the input exposures. Besides, some researchers [Khan et al., 2006; Pedone and Heikkilä, 2008] proposed to use the kernel density estimator to iteratively determine a probability that a pixel belongs to a moving object. Gallo et al. [Gallo et al., 2009] and Eden et al. [Eden et al., 2006] proposed to composite the desirable radiance with the guidance of a reference view preselected automatically or manually.

In summary, all above work was presented in the radiance domain fully or partially. Hence, there are two common limitations. First, the performance highly relies on the success of radiometric calibration of CRF which is sensitive to image noise, illumination change and misalignment error. Second, they normally have complex working pipelines and require tone mapping for HDR reproduction. The above problems make these kinds of methods tend to be computationally expensive and restrict their applications in practice.

1.3 Thesis Outline

This thesis focuses on enhancing the visual quality of an image captured with a conventional camera on three aspects: spatial resolution, focus setting and dynamic range. This thesis is divided into six chapters.

Chapter 1 gives an introduction about the thesis, including the motivation, objectives, related work and thesis organization.

Chapter 2 and Chapter 3 address the challenging single image SR problem, which is to recover a HR image from a single LR input. Chapter 2 presents a learning-based framework which aims at face image SR task from the perspective of DCT domain.

Chapter 3 describes an efficient two-step scheme which aims at super-resolving generic image by exploiting the salient edges in the LR input.

Chapter 4 describes a new and complete focus editing system that is able to handle the tasks of focus map estimation, image refocusing and defocusing. Given an image with a mixture of focused and defocused objects, we first detect the edges and then estimate the focus map based on the edge blurriness which is depicted explicitly by a parametric model. Then, by means of refocusing and defocusing, we seek to overcome the optical limitations and create novel images with different styles of focus effects.

Chapter 5 describes a simple but effective approach that is able to bypass the typical HDR process and directly yield a well-exposed image in both static and dynamic scenes by compositing multi-exposure images with the guidance of image quality assessment. A novel quality assessment system is developed by taking advantage of the gradient change information in differently exposed images. Compared to conventional HDR work, the proposed approach is quite appealing in practice since it is computationally efficient, easy to use and frees users from the tedious radiometric calibration and tone mapping steps.

Chapter 6 closes the thesis with a summary of the main contributions and several directions for further work.

Chapter 2

Super-Resolution for Face Image – Face Hallucination

2.1 Introduction

As an active research field in image processing and computer vision, super-resolution (SR) is to produce a high-resolution image (HRI) or a sequence of HRIs from a low-resolution image (LRI) or a sequence of LRIs. Recently, *face hallucination*, an interesting topic within SR, has aroused much attention. This term, firstly introduced by Baker and Kanade [Baker and Kanade, 2000], is about the generation of a high-resolution (HR) face image from low-resolution (LR) input. Face hallucination can be applied in many fields ranging from image compression to face identification. For example, in video surveillance, the ability to generate a higher resolution face image with detailed facial features from low resolution face images can raise the system performance.

In this chapter, we propose a novel learning-based face hallucination framework built in the Discrete Cosine Transform (DCT) domain as shown in Figure 2.1. Instead of estimating pixel intensities directly as the traditional learning-based algorithms, we concern ourselves with inferring the DCT coefficients, which contains two parts: DC coefficient estimation and AC coefficient inference. DC coefficients, which represent the average pixel intensity of the target blocks, can be estimated fairly accurately by interpolation methods such as Bilinear and Cubic B-Spline. AC coefficients, which contain the information of local features such as edges and corners around eyes, mouth of face image, cannot be estimated well by interpolation. Therefore, a simple but effective learning-based inference model is proposed to tackle this challenging problem in this work. The basic idea of the proposed method is that we are interested in learning the local facial features embodied in AC coefficients only, so that a more specific and efficient training set for AC coefficients can be built and used. Without considering

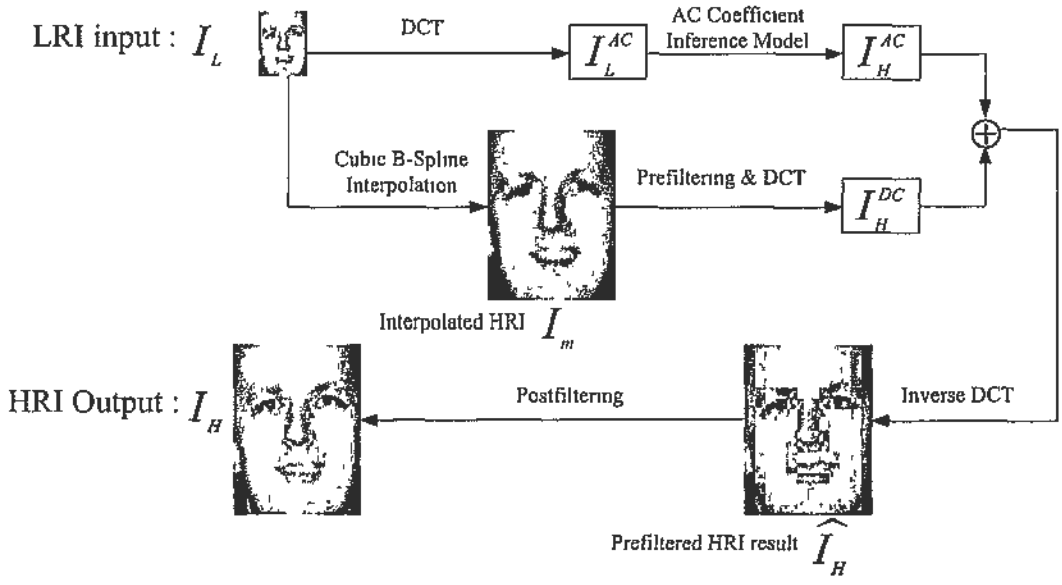

Figure 2.1: *The proposed face hallucination framework.*

DC coefficients, the proposed learning process will be more robust since it is much less influenced by image illumination. Moreover, in order to reduce the redundancy of the training set, a compact block dictionary is built by a clustering-based training scheme as stated in Section 2.6.

Furthermore, the intermediate hallucinated result $\widehat{I_H}$ in Figure 2.1 is an image preprocessed by a prefiltering scheme [Tran et al., 2003; Yu and Tran, 2002] which processes the block boundaries to remove the correlation of neighboring blocks. Therefore we can assume that each HRI block in the proposed AC coefficient inference model is independent of its adjacent HRI blocks. This significantly simplifies the inference model. The final output I_H can be obtained from $\widehat{I_H}$ by postfiltering. Another important benefit of combining the filtering scheme into the face hallucination process is that the blocking artifacts which often occur in block or patch based algorithms are greatly reduced. Besides, unlike conventional SR work such as [Freeman et al., 2002], a more general way of utilizing training priors - k-pass criterion, is adopted in the proposed learning process. In detail, each target HRI block in the proposed inference model is derived from multiple training samples instead of only one.

The rest of this chapter is organized as follows. In Section 2.2, we briefly review existing relevant work. Section 2.3 formulates the problem and gives an overview of

the proposed method. The simplified AC coefficient inference model is introduced in Section 2.4. The reconstruction of the target HRI is given in Section 2.5. Section 2.6 introduces the clustering-based training scheme and shows some experimental results. Section 2.7 draws some concluding remarks.

2.2 Related Work

Face hallucination from a single LR face image which is also referred as single-image SR problem has received a lot of attention in recent years. A number of related SR and face hallucination algorithms have been proposed, which can be grouped into three types. Interpolation-based algorithms (e.g., Bilinear, Cubic B-Spline) suffer from severe blurring problem especially when the resolution of the input is very low. Reconstruction-based methods [Morse and Schwartzwald, 2001; Lin and Shum, 2004], which try to model the process of image formation to build the relationship between LRI and HRI based on reconstruction constraints and smoothness constraints, are quite limited by the number of input LRIs and usually cannot work well in single-image SR problem.

Recently, learning-based methods become very popular. Usually, the unknown HRI is inferred by making use of some training set directly or indirectly. In comparison with other methods, learning-based method can achieve higher magnification factor and better visual quality especially for single-image SR problem [Lin et al., 2008]. Baker and Kanade [Baker and Kanade, 2000; Baker and Kanade, 2002] presented a pioneering work on hallucinating face image based on a Bayesian formulation. The target HRI is inferred by resorting to a training set. Capel and Zisserman [Capel and Zisserman, 2001] extracted eigenfaces from a collection of training face images as a prior model to constrain and super-resolve LR face images. Freeman et al. [Freeman et al., 2000] proposed a well-known parametric MRF (markov random field) based inference model to learn the statistics between the underlying *scene* and the observed *image* data. This framework was applied to the SR problem as well as other typical low-level vision problems. Based on such framework, Liu et al. [Liu et al., 2007] developed a two-step statistical modeling approach for face hallucination which integrates a global parametric model and a local nonparametric model. Besides, Muresan and Parks [Muresan and Parks, 2002] presented a learning-based face hallucination method from an adaptive optimal recovery point of view. Liu et al. [Liu et al., 2005] proposed a TensorPatch model

face hallucination and devised a residue compensation step to enhance the hallucination result. All above mentioned learning-based methods are built in spatial domain for the inference of pixel intensities of the target HRI, and differ with each other on the learning manner from the training set. A major problem of these methods is the high computation requirement due to the complex learning process. Especially when the MRF based inference model is used and the training set is very large, rather taxing computation and heavy memory load are required.

Some SR algorithms have been proposed to tackle the problem in transform domain which is normally the DCT domain. It is because the DCT has high energy packing ability and is adopted in most image and video coding standards. Ni and Nguyen [Ni and Nguyen, 2007] used SVR (Support Vector Regression) and utilized the DCT structural properties to solve their proposed regression structure. Patti and Altunbasak [Patti and Altunbasak, 1999] proposed a POCS solution that directly incorporates the transform domain quantization information by working with the compressed bit stream. Park et al. [Park et al., 2004] presented a HR reconstruction method for DCT-based compressed images that simultaneously estimates the quantization noise modeled as a correlated Gaussian process in spatial domain. Pham et al. [Pham et al., 2006] implemented the prevalent learning-based method [Freeman et al., 2002] in the DCT domain for fast super-resolving the compressed video. However their results suffer from severe blocking artifacts, even with a strict constraint that the HR priors are limited to use the same scene as that of the LR video.

Recently, some work was proposed to allow face hallucination technology to handle faces with different poses or expressions. For example, Li and Lin [Li and Lin, 2004] proposed to tackle the pose variation problem by estimating the pose of the profile input face image based on SVM (support vector machine) classifier. The corresponding frontal face image is synthesized and then super-resolved into a HR frontal one. A more generalized approach was proposed by Jia and Gong [Jia and Gong, 2008] to super-resolve LR face images with variations in facial expression and pose based on a hierarchical tensor space representation.

HRIs inferred using interpolation-based methods suffer from blurring problem which is especially severe in high activity regions containing edges and corners. For example in Figure 2.2, Cubic B-Spline interpolation is used to enlarge a 24×32 LR face image to

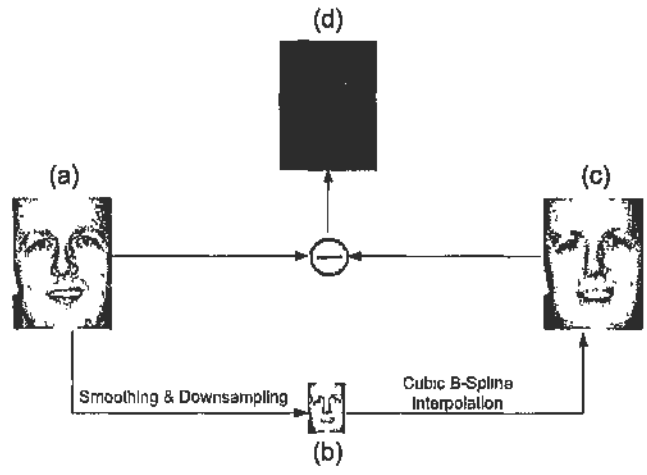

Figure 2.2: *Face hallucination using Cubic B-Spline interpolation. (a) the original HRI (96 × 128); (b) the synthesized LRI (24 × 32); (c) the interpolated HRI using Cubic B-Spline; (d) the difference image*

a 96×128 HRI. The difference image shown in Figure 2 2(d) shows that Cubic B-Spline works well in the smooth parts of face, but introduces large distortion in high activity regions such as eyes, mouth and nose. This is because the higher frequency components which contain the information of local details are missing. Interpolation-based methods do not introduce new high frequency components required by the inferred HRI at high activity regions. Learning-based methods solve the above problems by creating the required high frequency components from a training set. However, the training set for faces with a particular pose and expression is only applicable to the hallucination of faces under similar conditions. Our experimental results show that visual quality deteriorates quickly when difference in pose is larger than 10 degrees. This problem can be tackled by methods like [Li and Lin, 2004; Jia and Gong, 2008]. Alternatively, one may first detect the pose of a face and then perform face hallucination using the corresponding training set. This method can produce better results but requires many training set and so heavier memory load and more computation. Therefore, it is important to develop a simple algorithm that can perform face hallucination effectively.

2.3 Problem Formulation and Overview of the Proposed Framework

2.3.1 Problem Formulation after Transform Face Image by the DCT

As a popular transform in image processing, the DCT [Wang et al., 2002a] refers to a separable orthogonal linear mapping of blocks of image pixels into blocks of transform coefficients. Similar to Discrete Fourier Transform (DFT), it transforms a signal or an image from spatial domain to frequency domain.

The m_{th} element of the u_{th} basis vector of the 1-D N-point DCT is defined as:

$$V(u,m) = \begin{cases} \sqrt{1/N} & , u = 0,\ 0 \leq m \leq N-1 \\ \sqrt{2/N} cos\frac{\pi(2m+1)u}{2N} & , 1 \leq u \leq N-1,\ 0 \leq m \leq N-1. \end{cases}$$

Also $V^{\top} = V^{-1}$ since the DCT is a real and orthogonal transform. To obtain the 2-D DCT of an $N \times N$ image block, one can first apply the 1-D DCT to each row of the block and then to each column of the row transformed block, i.e., $C = VTV^{\top}$ and $T = V^{\top}CV$ where T denotes an image block, C denotes the block of the DCT coefficients. Also, the image block can be regarded as the sum of N^2 basis images $B(u,v)$ weighted by $C(u,v)$ as follows.

$$T = \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} C(u,v)B(u,v). \tag{2.1}$$

Note that $B(u,v)$ is constructed by the outer product of the u_{th} and v_{th} basis vectors. The DCT coefficient $C(u,v)$ specifies the contribution of the basis image $B(u,v)$ to T. For example, the DC coefficient, $C(0,0)$, denotes DC level and the average pixel intensity of the target block. The other coefficients, known as AC coefficients, are associated with higher frequencies.

In this work, an image is divided into 8×8 non-overlapped blocks and the hallucination is performed block by block. The block size is chosen to be 8×8 which is informative enough to represent the target scene. Another reason is that the 2-D 8-point DCT is widely adopted in image and video coding. Figure 2.3 shows the basis images $B(u,v)$ of the 8×8 DCT. The frequency of the basis image increases from left to right and top to bottom. In the proposed method, the DCT coefficients are divided into three groups based on the zig-zag scanning order, which are DC coefficients, low frequency and high frequency coefficients as shown in Figure 2.4.

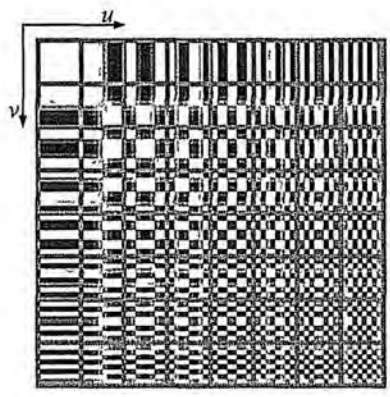

Figure 2.3· *Graphical illustration of the 8 × 8 DCT basis images The frequencies u and v of the basis image increase from left to right and top to bottom*

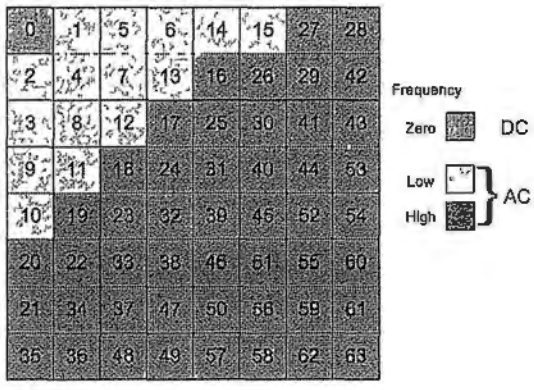

Figure 2.4· *The 8 × 8 DCT coefficients are categorized as zero (DC), low and high frequency coefficients based on the zig-zag scanning order*

The reason that the DCT is well suited for image compression is that an image block can often be represented by a few low frequency DCT coefficients [Wang et al., 2002a]. This is because natural images are often smooth and significant high frequency components exist only occasionally around edges. Hence, much of the energy lies at low frequency coefficients. High frequency coefficients are often small and can be discarded with little visible distortion.

This is also true for face images which are smooth and contain few high frequency components. Figure 2.5(c) shows the energy distribution (i.e. the variance) of the 8×8 DCT coefficients of a face image. The energy of the DCT coefficients drops quickly as frequency index increases. Figure 2.5(a) and (b) show the images reconstructed with different numbers of the DCT coefficients. We can see that with only 16 out of 64

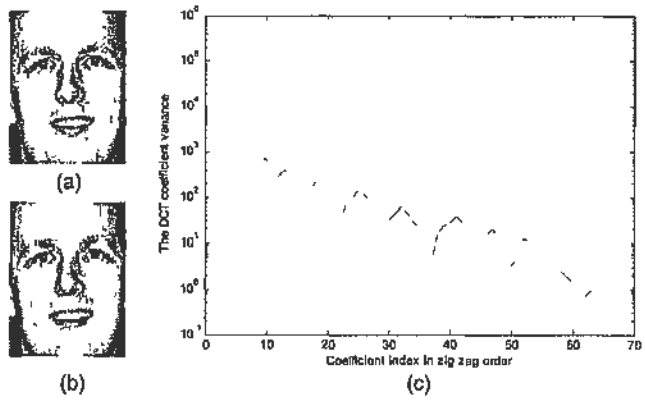

Figure 2.5: *Face image coded by the* 8×8 *DCT. (a) the original image reconstructed with all 64 DCT coefficients (zero+low+high frequencies) per block. (b) the reconstructed image with the first 16 coefficients (zero+low frequencies). (c) the energy distribution of the* 8×8 *DCT coefficients of the face image.*

coefficients per block, the target image is already well represented.

This implies that it is not necessary to infer all coefficients. Instead, we only need to focus on inferring the coefficients that are vital to the visual quality. Hence, the intent of this work is to infer the DC and the low frequency coefficients in each 8×8 block of the target HRI. High frequency coefficients are excluded due to their weak energy in the face image.

2.3.2 Advantages of Face Hallucination in the DCT Domain

In the proposed method, face hallucination is tackled by inferring the DC and the 15 low frequency AC coefficients for each block of the target image in the DCT domain. Such formulation will benefit us in several aspects:

1. As shown in Figure 2.6, the DC coefficient which represents the average pixel intensity of a target block, can be estimated fairly accurately by a simple interpolation-based method such as Cubic B-Spline.

2. We only need to focus on building a specific learning-based inference model for these low frequency AC coefficients which correspond to the local details of face image such as the edges, corners around eyes.

3. A simplified learning-based inference model can be developed to infer the AC coefficients efficiently based on a reasonable assumption that blocks of the prefiltered

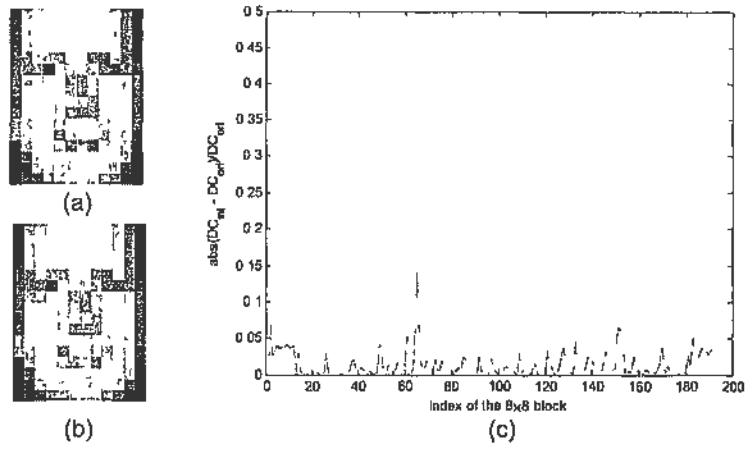

Figure 2.6: *DC coefficient estimation by Cubic B-Spline interpolation. (a) the reconstructed image only with the original DC coefficients (refer to Figure 2.2(a)); (b) the reconstructed image only with the DC coefficients which are estimated by Cubic B-Spline interpolation (refer to Figure 2.2(c)). There is litter noticeable difference between (a) and (b). (c) shows each block's relative estimation error of the estimated DC coefficient (DC_{int}) to the original DC coefficient (DC_{ori}). It is evident that the errors are all very small.*

HRI built in the DCT domain are independent with each other.

4. The data dimension of training and testing set can be reduced significantly. As 15 AC coefficients in an 8×8 block are enough to produce a satisfying result with detailed local features as shown in Figure 2.5, the dimension of HRI block can be reduced from 64 in spatial domain to 15 in the DCT domain in this case [1].

In summary, as shown in Figure 2.1, the proposed framework can be divided into two steps. Firstly, the prefiltered HRI I_H is inferred in the DCT domain, which includes AC coefficient inference by learning and DC coefficient estimation by interpolation. Secondly, the final hallucinated result I_H is reconstructed from the prefiltered result $\widehat{I_H}$ by postfiltering.

[1] Note that this conclusion is made based on the fact that the texture (skin) of face image is generally smooth. Hence, high frequency DCT coefficients of face images have very small magnitude and a small number of low frequency coefficients only are good enough to produce satisfying result with enough details. For those general images which contain large high frequency components, more AC coefficients need to be used to preserve the finest details of the output image.

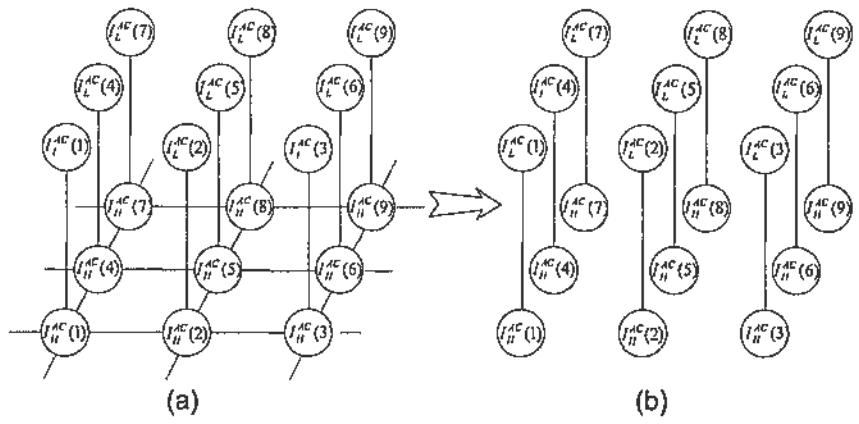

Figure 2.7: *Graphical models for AC coefficient inference. (a) the prevalent MRF based inference model [Freeman et al., 2000]; (b) the proposed simplified inference model.*

2.4 Learning-based AC Coefficient Inference Model

The inference of the AC coefficients I_H^{AC} for the target HRI can be formulated as a *maximum a posteriori* (MAP) problem:

$$I_H^{AC\,*} = \arg\max_{I_H^{AC}} p(I_H^{AC}|I_L^{AC}). \tag{2.2}$$

As shown in Figure 2.7(a), a typical MRF inference model [Freeman et al., 2000] used in the low-level vision tasks can be employed to address this problem. Node $I_H^{AC}(i)$ and node $I_L^{AC}(i)$ are used to represent unknown i_{th} HR block of HRI and the observed i_{th} LR block of LRI respectively. The links between nodes indicate statistical dependencies which as given by the MRF model in Figure 2.7(a) have two implications: 1) HRI block $I_H^{AC}(i)$ provides all the information about the observed LRI block $I_L^{AC}(i)$ as the only link to $I_L^{AC}(i)$ is from $I_H^{AC}(i)$; 2) HRI block $I_H^{AC}(i)$ gives information about the adjacent HRI blocks by the links from $I_H^{AC}(i)$ to its adjacent HRI blocks.

Since $p(I_H^{AC}|I_L^{AC}) = \frac{p(I_H^{AC},I_L^{AC})}{p(I_L^{AC})}$ and $p(I_L^{AC})$ is constant over I_H^{AC}, (2.2) can be rewritten as

$$I_H^{AC\,*} = \arg\max_{I_H^{AC}} p(I_H^{AC}, I_L^{AC}). \tag{2.3}$$

According to the MRF model in Figure 2.7(a), the joint probability of I_L^{AC} and I_H^{AC}

can be written as:

$$p(I_H^{AC}, I_L^{AC}) = p(I_H^{AC}(1), ..., I_H^{AC}(n), I_L^{AC}(1), ..., I_L^{AC}(n))$$
$$= \frac{1}{Z} \prod_{(i,j)} \psi(I_H^{AC}(i), I_H^{AC}(j)) \prod_i \phi(I_H^{AC}(i), I_L^{AC}(i)) \quad (2.4)$$

where Z is a normalization factor, n denotes the number of block pairs, (i, j) indicates neighboring blocks. Both ψ and ϕ are pairwise compatibility functions which can be learned from the training set. ψ is used to model the spatial smoothness between the neighboring HRI blocks. ϕ is used to model the dependency between the corresponding LRI and HRI blocks.

Now the problem in (2.2) becomes:

$$I_H^{AC\,*} = \arg\max_{I_H^{AC}} \prod_{(i,j)} \psi(I_H^{AC}(i), I_H^{AC}(j)) \prod_i \phi(I_H^{AC}(i), I_L^{AC}(i)). \quad (2.5)$$

Hence, the target I_H^{AC} can be inferred from a training set with the loopy Belief Propagation (BP) algorithm [Freeman et al., 2000]. However, finding a global optimum for (2.5) is difficult and certainly time consuming. Fortunately, the inference model as well as the optimization can be made more tractable. Next, we shall first analyze the correlation among AC coefficients and then derive a simplified AC coefficient inference model.

2.4.1 Analysis of the AC Coefficient Correlation

Given a $LN \times MN$ image, the $N \times N$ DCT will map it into a $L \times M$ grid of $N \times N$ coefficient blocks and $C(u, v; l, m)$ can be introduced to index the DCT coefficients, where (u, v) $(0 \leq u, v < N)$ denotes the frequency index. (l, m) $(0 \leq l < L, 0 \leq m < M)$ is the block index. Figure 2.8 shows an example when $N = 4, L = M = 2$. In Figure 2.8(a), block (l, m) contains coefficients computed using pixels in block (l, m). The coefficients such as (u, v) represent different frequency components of a local spatial region. In Figure 2.8(b), subband (u, v) contains coefficients (u, v) collected from each block. Hence, the DCT coefficient $C(u, v; l, m)$ in each block has two kinds of neighbors: spatial neighbors and subband neighbors.

Accordingly, there are also two kinds of correlation for each AC coefficient. The correlation between AC coefficient $C(u, v; l, m)$ and its spatial neighbors such as $C(u+$

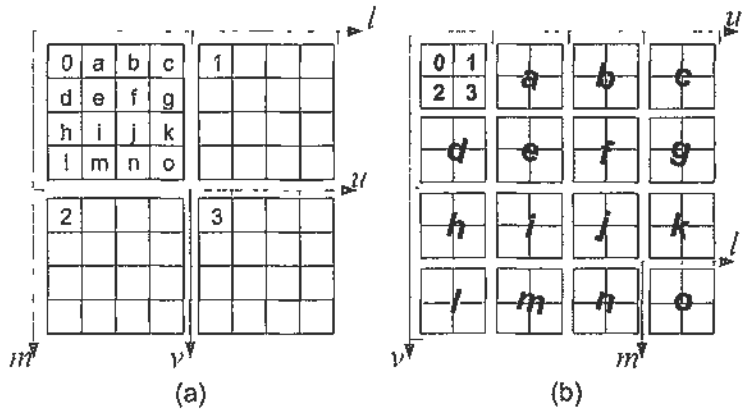

Figure 2.8: *Block representation (a) and subband representation (b) for the 4 × 4 DCT coefficients.*

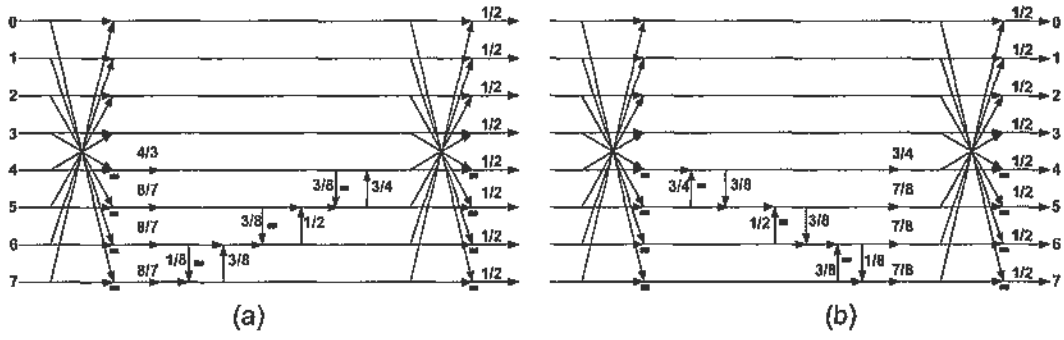

Figure 2.9: *Prefilter P (a) and postfilter P^{-1} (b) for 8 × 8 block processing [Tran et al., 2003; Tu and Tran, 2002].*

$1, v; l, m)$ and $C(u, v+1; l, m)$ is very weak and can be ignored due to the excellent decorrelation capability of the DCT. The correlation between AC coefficient $C(u, v; l, m)$ and its subband neighbors such as $C(u, v; l + 1, m)$ and $C(u, v; l, m + 1)$ is stronger than that between the spatial neighbors. This correlation referred as interblock correlation, is exhibited by the smoothness of neighboring blocks in spatial domain. Inspired by the recent work in image compression and coding [Tran et al., 2003; Tu and Tran, 2002], we adopted a filtering method to process the boundaries of neighboring blocks for exploiting their interblock correlation. A pair of filters, prefilter P and postfilter P^{-1} shown in Figure 2.9 are used in the proposed method. The prefilter P, depicted in Figure 2.9(a), is performed in a separable fashion to remove the 8 × 8 interblock correlation of the image. Postfilter P^{-1} depicted in Figure 2.9(b), is to reconstruct

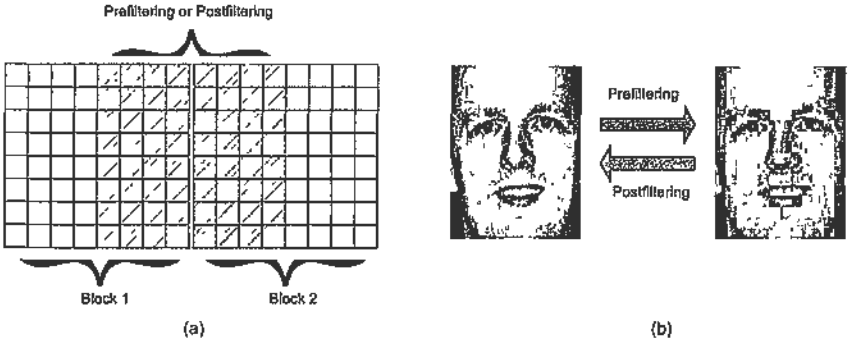

Figure 2.10: *Prefiltering and postfiltering on a face image. (a) the working mode of the prefilter and postfilter shown in Figure 2.9. Block 1 and Block 2 are two 8 × 8 adjacent blocks in horizontal direction, the prefilter and postfilter will work on their neighboring boundaries. (b) left is the original image, right is the prefiltered image.*

the original smooth image from the prefiltered image by recovering the interblock correlation. The effects of prefiltering and postfiltering are illustrated in Figure 2.10. It is found that the prefiltered image becomes blocky and obvious discontinuity exists in both horizontal and vertical block boundaries since the prefilter has taken away their interblock correlation. Postfilter is able to recover the original smooth image.

In the proposed framework, the filtering scheme will work together with the DCT as shown in Figure 2.11. The prefiltering scheme is used to remove the interblock correlation of the training HRI samples. The intermediate HRI result I_H as shown in Figure 2.1 is a prefiltered image and the final HRI result I_H can be obtained by postfiltering. In summary, we can assume that each AC coefficient of the prefiltered image is neither correlated with its spatial neighbors nor correlated with its subband neighbors. As a result, a reasonable assumption can be made that each block $I_H^{AC}(i)$ in the target prefiltered HRI is independent with its adjacent HRI blocks.

2.4.2 A Simplified AC Coefficient Inference Model

Now the MRF model can be simplified a lot by eliminating all links among HRI blocks as shown in Figure 2.7. Hence, (2.5) becomes:

$$I_H^{AC\,*} = \arg\max_{I_H^{AC}} \prod_i \phi(I_H^{AC}(i), I_L^{AC}(i)). \tag{2.6}$$

The next problem is to build a reasonable compatibility function $\phi(I_H^{AC}(i), I_L^{AC}(i))$ for the proposed inference model.

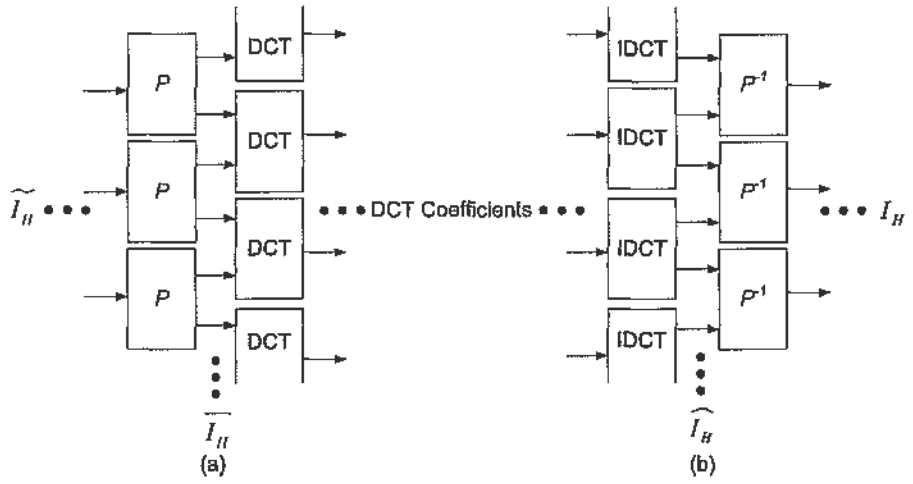

Figure 2.11: *Prefiltering (a) and postfiltering (b) are performed along the block boundaries block-wise locally similar to the DCT. (a) prefilter P is adopted to preprocess the training HRI samples $\widetilde{I_H}$ as prefiltered HRI samples $\overline{I_H}$. (b) postfilter P^{-1} is adopted to reconstruct the final hallucinated result I_H from the intermediate result $\widehat{I_H}$ which is a prefiltered HRI.*

K-pass Algorithm

In previous SR work (e.g., [Freeman et al., 2002; Sun et al., 2003; Bishop et al., 2003]), single-pass criterion is often used to select the best fitting sample from the training set. While in this chapter, a more general k-pass criterion is adopted as that k candidate blocks are selected to construct the target HRI block only based on the compatibility function ϕ. Intuitively k-pass is more effective than single-pass in producing a target block with high fidelity because the linear combination of k blocks is more informative than a single one. Recently, locally linear embedding (LLE) [Roweis and Saul, 2000] was presented to map high-dimensional data into low-dimensional space by preserving the neighborhood relationship. Inspired by this idea, we make an assumption similar to those of [Chang et al., 2004] and [Chang et al., 2006] as follows. For each pair of corresponding LRI and HRI blocks, their local neighborhoods on some proper manifolds are assumed similar. In detail, it is assumed that each LRI block $I_L^{AC}(i)$ and its nearest neighbors in low dimension lie on or close to a locally-linear structure. Hence, $I_L^{AC}(i)$ can be linearly approximated by its k nearest neighbors $\overline{I_L^{AC}}(j)$ selected from the LR training samples with the weighting coefficients $W_i(j)$. On the other hand, when $I_L^{AC}(i)$ and its neighbors $\overline{I_L^{AC}}(j)$ are fixed, $W_i(j)$ can be obtained easily by minimizing the reconstruction error in (2.7) subject to $\sum_{j=1}^{k} W_i(j) = 1$. Besides, since the LR and

HR blocks are assumed have similar linear structure, the weights W_i minimizing the reconstruction error on the LRI blocks should also yield a small value when the data is replaced with the HRI blocks. In this work, the reconstruction errors on LRI blocks and on HRI blocks are modeled with zero mean Gaussian distributions of variance σ_L^2 and σ_H^2 respectively.

Now we can describe the two local structures by:

$$I_L^{AC}(i) = \sum_{j=1}^{k} W_i(j)\overline{I_L^{AC}}(j) + \mathcal{N}(0, \sigma_L^2) \tag{2.7}$$

$$I_H^{AC}(i) = \sum_{j=1}^{k} W_i(j)\overline{I_H^{AC}}(j) + \mathcal{N}(0, \sigma_H^2) \tag{2.8}$$

where $\overline{I_L^{AC}}(j)$ and $\overline{I_H^{AC}}(j)$ are the training samples selected from the training set $\Phi = \{\overline{I_L^{AC}}\ \overline{I_H^{AC}}\}$. Hence, each HRI block is generated using several candidates instead of one. The compatibility function $\phi(I_H^{AC}(i), I_L^{AC}(i))$ in (2.6) is defined as:

$$\begin{aligned}&\phi(I_H^{AC}(i), I_L^{AC}(i), W_i) = \\ &exp\{-(I_L^{AC}(i) - \sum_{j=1}^{k} W_i(j)\overline{I_L^{AC}}(j))^2/2\sigma_L^2\} \times exp\{-(I_H^{AC}(i) - \sum_{j=1}^{k} W_i(j)\overline{I_H^{AC}}(j))^2/2\sigma_H^2\}\end{aligned} \tag{2.9}$$

whose value is between [0 1]. After introducing $\lambda = \sigma_L^2/\sigma_H^2$, we can define an error function by applying the negative logarithms to (2.9):

$$E(I_H^{AC}(i), I_L^{AC}(i), W_i) = E_1(I_L^{AC}(i), W_i) + \lambda E_2(I_H^{AC}(i), W_i) \tag{2.10}$$

where

$$E_1(I_L^{AC}(i), W_i) = (I_L^{AC}(i) - \sum_{j=1}^{k} W_i(j)\overline{I_L^{AC}}(j))^2 \tag{2.11}$$

$$E_2(I_H^{AC}(i), W_i) = (I_H^{AC}(i) - \sum_{j=1}^{k} W_i(j)\overline{I_H^{AC}}(j))^2. \tag{2.12}$$

Thus, the optimization of (2.6) can be solved by minimizing $E(I_H^{AC}, I_L^{AC}, W)$ over W

Algorithm 1 AC coefficient inference

Given: training set $\Phi = \{\overline{I_L^{AC}}, \overline{I_H^{AC}}\}$, number of nearest neighbors k.

Input: I_L^{AC}

Loop: *for each LRI block* $I_L^{AC}(i)$

1. Find its k nearest low-resolution neighbors to constitute $\bigcup_{j=1}^{k} \overline{I_L^{AC}}(j)$
2. According to step 1, take the corresponding k high-resolution blocks for $I_H^{AC}(i)$ to constitute $\bigcup_{j=1}^{k} \overline{I_H^{AC}}(j)$
3. Calculate W_i by minimizing (2.11) as a constrained least squares problem
4. Given W_i, set $I_H^{AC}(i) = \sum_{j=1}^{k} W_i(j)\overline{I_H^{AC}}(j)$

Output: I_H^{AC}

and I_H^{AC} alternatively.

$$E(I_H^{AC}, I_L^{AC}, W) = \sum_i E(I_H^{AC}(i), I_L^{AC}(i), W_i) \tag{2.13}$$

Minimization of the Error Function

Given the training set, the minimization of (2.10) can be solved easily because the error function $E(I_H^{AC}(i), I_L^{AC}(i), W_i)$ is quadratic to $I_H^{AC}(i)$ and W_i respectively. For the i_{th} block, the optimal $I_H^{AC}(i)$ can be obtained as $I_H^{AC}(i) = \sum_{j=1}^{k} W_i(j)\overline{I_H^{AC}}(j)$ by setting the derivative of $E(I_H^{AC}(i), I_L^{AC}(i), W_i)$ w.r.t $I_H^{AC}(i)$ to zero. Substituting the optimal $I_H^{AC}(i)$ into (2.10), we can obtain the optimal W_i by simply minimizing $E_1(I_L^{AC}(i), W_i)$ in (2.13) with $\sum_{j=1}^{k} W_i(j) = 1$ as a constrained least squares problem [Roweis and Saul, 2000]. Therefore the whole optimization can be done efficiently without iteration. Details of the AC coefficient inference algorithm are summarized in Algorithm 1 .

2.5 HRI Reconstruction by the Inverse DCT and Postfiltering

As shown in Figure 2.1, the DCT coefficients of each block in the prefiltered HRI I_H can be recovered in two steps: 1) the low frequency AC coefficients which constitute $I_H^{AC}(i)$ are estimated by the aforementioned AC coefficient inference model and other high frequency AC coefficients are set to zero; 2) interpolate the HRI I_h from the input LRI by the Cubic B-Spline method and then apply prefiltering on it. The target DC coefficients are estimated from the corresponding blocks of the prefiltered HRI. From all these estimated DC and AC coefficients, the target prefiltered HRI I_H is reconstructed by the inverse DCT. Then the final HRI result I_H is derived from $\widehat{I_H}$ by the postfiltering scheme as shown in Figure 2.1(b).

2.6 Experimental Results

2.6.1 Learning Block Dictionary Φ by Clustering

As discussed in [Liu et al., 2008], the performance of learning-based method often depends on how well the input LRI matches the samples in the training set. Obviously, the more training samples are collected, the more robust the learning-based algorithm is. However, a huge training set requires taxing computation and heavy memory load. Fortunately, the blocks cropped from the face images do not have much variation since human facial features are similar. This is also true in our case because our training set contains only local facial features represented by AC coefficients. Hence, the raw training set should have much redundancy and it is possible to learn those most representative blocks and build a compact block dictionary by clustering.

In our method, all collected training images are firstly aligned by affine transform based on three marked points: the centers of the two eyes and the center of the mouth. Then each image is cropped to a canonical 96×128 image as the HRI. Its corresponding 24×32 LRI can be obtained by smoothing and downsampling. After being preprocessed by the above prefiltering scheme, all HRIs are transformed from spatial domain to frequency domain by the 8×8 DCT. The HRI blocks $\overline{I_H^{AC}}$ of the training data are these non-overlapped 8×8 blocks and represented by only using the low frequency AC coefficients. Since the LRIs will be initially enlarged via Cubic B-Spline interpolation, AC coefficients of the corresponding LRI blocks $\overline{I_L^{AC}}$ are also obtained by performing the 8×8 DCT. Finally, the redundancy of the raw training samples will be reduced through affinity propagation clustering [Frey and Dueck, 2007].

2.6.2 Comparison

This experiment was conducted with a large number of frontal face images from the Facial Recognition Technology (FERET) database [Phillips et al., 1998; Phillips et al., 2000] and other collections, which consist of many different races, illuminations and types of face images. Among these samples, about 1600 images were selected as training data and the remaining images were for testing. In our experiments, the number of nearest neighbors k was set to 7. Please also visit `http://www.ee.cuhk.edu.hk/~zhangwei/HalluciFace.html` to see the results.

Figure 2.12: *Face hallucination results of frontal face images. (a) the input LRIs (24 × 32); (b) our intermediate prefiltered results $\widehat{I_H}$; (c) our final results I_H; (d) Cubic B-Spline Interpolation; (e) Freeman et al. [Freeman et al., 2000]; (f) Baker et al. [Baker and Kanade, 2002]; (g) Liu et al. [Liu et al., 2007]; (h) the original HRIs (96 × 128).*

To make a fair evaluation, the proposed approach is compared to some of the existing learning-based methods using the same training samples. The experimental results are shown in Figure 2.12. We can observe that Cubic B-Spline interpolation suffers from severe blurring problem. Freeman et al.'s results are much better but still have outliers. Baker et al.'s method produces noisy results in facial features. Liu et al.'s results have satisfactory visual quality but some subtle characteristics cannot be generated well, especially the details around eyes. While, the HRIs reconstructed by the proposed method have the finest facial details. It is noted that a rigid 87 feature point system [Chen et al., 2001] is adopted in Liu et al.'s method to regularize the face shape in order to improve the performance of face hallucination. Obviously such

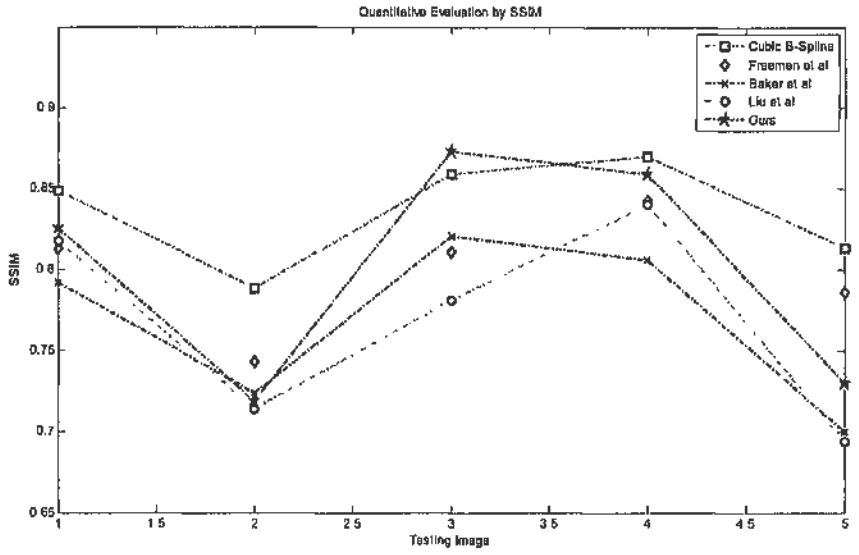

Figure 2.13: *Quantitative evaluation of the hallucinated results in Figure 2.12. The testing images from top to bottom in Figure 2.12 are indexed from number 1 to number 5.*

complicated alignment is time consuming. While in this work, the algorithms were tested in a more general scenario similar to previous work [Baker and Kanade, 2000; Baker and Kanade, 2002; Liu et al., 2005]. Specifically, the training samples used here were generated roughly as described in the last section. Only three points were used for the face alignment. If more rigid technique is implemented to regularize the face images, the performance of the learning-based method should be better.

We also compared the methods quantitatively as shown in Figure 2.13, where a recently developed measure called SSIM (structural similarity) [Wang et al., 2004] is used to assess the similarity of a reconstructed face image and the original HRI. The results show that faces hallucinated by the proposed method have high SSIM. The mean square error (MSE) is not adopted here due to its bad matching with the perceived visual quality [Girod, 1993; Wang et al., 2002b]. However, it is found that SSIM also has limitation. The Cubic B-Spline interpolated results mostly have the highest SSIM (except number 3 result) in the above comparisons, which is apparently not consistent with the perceptual quality. For face hallucination, the task is recovering the lost frequencies and enhancing the visual quality of the LR input to make people see or recognize the target face clearly. Therefore, so far the most effective way of measuring image quality is through subjective evaluation. More results can be found in Figure 2.14.

Figure 2.14: *More face hallucination results. In each triple, from left to right are the input LRI (24 × 32), the hallucinated result by the proposed method and the original HRI (96 × 128).*

2.6.3 Robustness to Image Illumination

As aforementioned, we only concern ourselves with learning local features embedded in AC coefficients from the training prior. It is found that without considering DC coefficients will make the learning process more robust since the matching from input to training samples is much less influenced by image illumination. An experiment as shown in Figure 2.15 is conducted to test the learning robustness of the proposed method. The five 96×128 images of the same person as shown in Figure 2.15(a) and (f) were taken at different time with different expressions and illumination conditions. Four images as shown in Figure 2.15(a) with high illumination are selected for training, Figure 2.15(f) captured under low illumination is used for testing. Given the LRI input Figure 2.15(b) derived from Figure 2.15(f) by smoothing and downsampling, Figure 2.15(c) is super-resolved by the proposed method. It is obvious that the proposed algorithm is nearly exempted from the illumination influence and capable of learning high quality local features from such a small training set. In contrast, Figure 2.15(d) which is inferred by learning the pixel intensities directly in spatial domain (i.e., considering both DC and AC coefficients) with an example-based manner, is very bad because the input LRI can not match well with the training samples due to the influence of image illumination. Although Baker et al.'s method produces a better result as shown in Figure 2.15(e), it still fails in digging out some subtle features from the training samples.

2.6.4 Test on Hallucinating Profile Face Image

The aforementioned experiments were conducted on frontal face images. In order to test the robustness of the proposed method on LR profile face images, an experiment was conducted on about 400 profile face images with similar pose as the training prior. The experimental results are shown in Figure 2.16. Apparently, the proposed method also performs well on hallucinating non-frontal face images.

2.6.5 Limitation

Besides, we would like to point out that as a typical learning-based method, our algorithm is also limited by the training set and the performance relies on the matching of the input LRI with the training samples. Therefore, the proposed method does not work well for every single face. For example, if the input is a person wearing glasses or

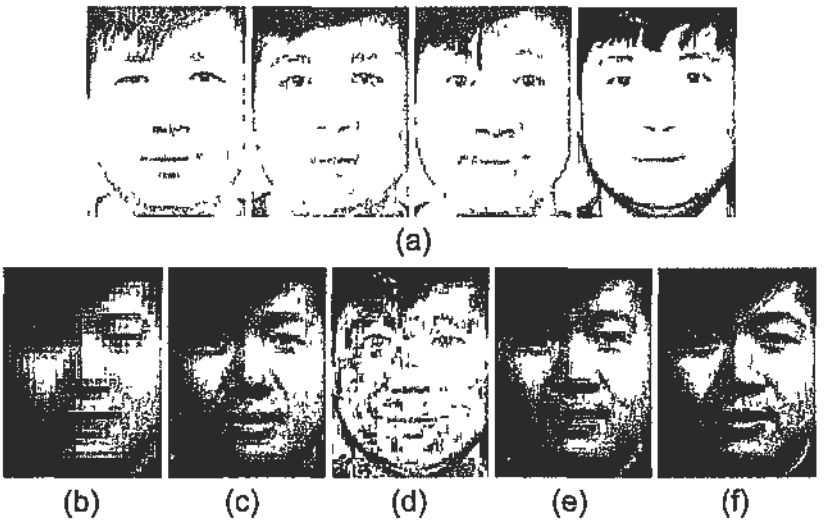

Figure 2.15 *Face hallucination with a small training set (a) the training prior HRIs (96×128); (b) the input LRI (24 × 32), (c) the proposed method, (d) learning in spatial domain; (e) Baker et al [Baker and Kanade 2002], (f) the original HRI (96 × 128)*

with closing eyes, then probably the algorithm cannot super-resolve the face accurately. This is because the training set we used does not contain sufficient samples which wear glasses or with closing eyes. We have shown some poor results in Figure 2.17. However, if more training samples are collected in the training set, the algorithm should be able to super-resolve more faces better.

2.7 Summary

In this chapter, we have presented an effective learning-based framework for face hallucination from a single LRI The problem is formulated as the DCT coefficient estimation in frequency domain, which benefits us in several aspects Firstly, it reduces the data dimension in both training set and testing set. Secondly, it reduces the complexity of the learning-based AC coefficient inference model because of the weak correlation among AC coefficients. Also, the inference model can be free of influence from the image illumination by only focusing on learning the local features embodied in AC coefficients from a collection of training samples Each block of the target HRI is generated by a linear combination of several candidate blocks selected from the training set whose redundancy has been reduced by clustering The effectiveness and robustness of the proposed approach have been demonstrated by a set of experimental results Besides, the basic idea of this work can be further extended to handle more general

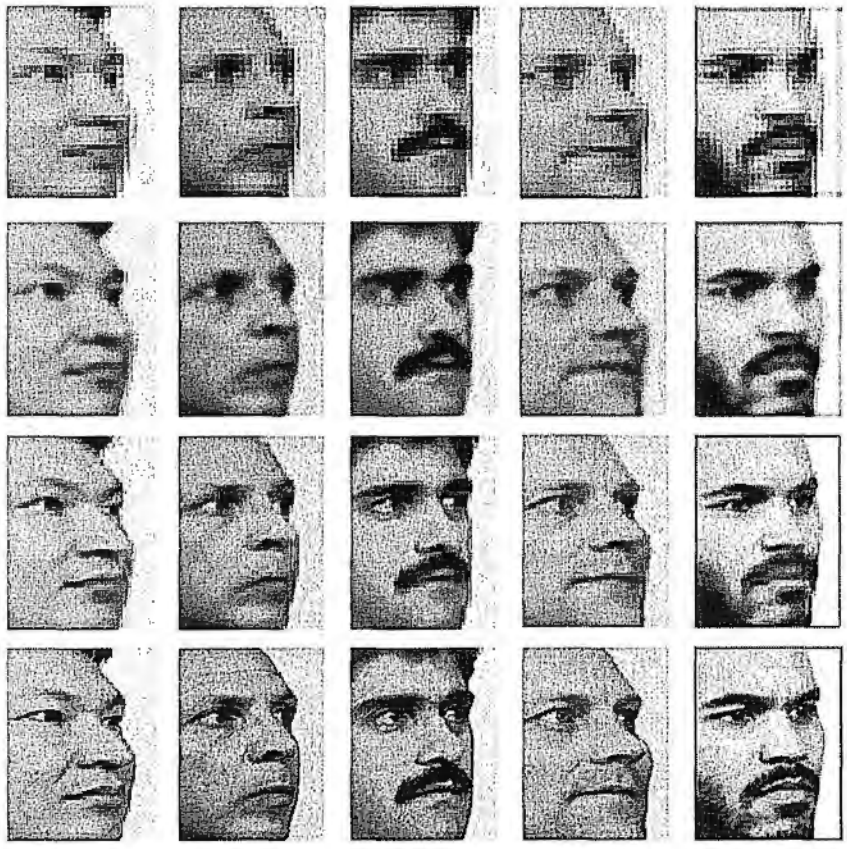

Figure 2.16: *Face hallucination results of profile face images. First row: the input LRIs (24 × 32); Second row: Cubic B-Spline Interpolation; Third row: hallucinated results by the proposed method; Forth row: the original HRIs (96 × 128).*

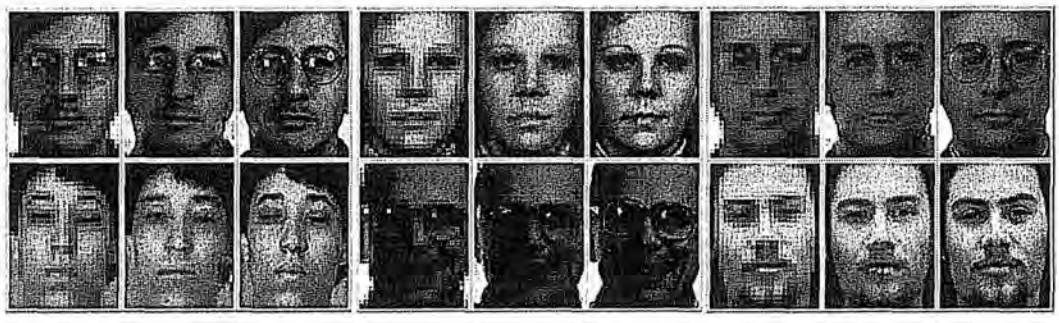

Figure 2.17: *Some examples which are not super-resolved well. In each triple, from left to right are the input LRI (24 × 32), the hallucinated result by the proposed method and the original HRI (96 × 128).*

tasks such as the resolution enhancement of general image and video. However, like other learning-based methods, the proposed algorithm is also limited by the training set and thus the performance depends on how well the LR input matches the training

samples.

Chapter 3

Super-Resolution for Generic Image

3.1 Introduction

Unlike the last chapter which concentrates on super-resolving face images, this chapter investigates the problem of generic image super-resolution (SR), which is more demanding nowadays due to the increasing popularity of High Definition Television (HDTV), webcam, camera phones and low-bandwidth video streaming.

It is acknowledged that edges are presumably the most important features in natural images. Therefore, for a super-resolved image, sharpness and freedom from artifacts on edges are the two critical factors for its perceptual quality. However, conventional SR techniques are usually susceptible to artifacts such as jaggies and blurring as shown in Figure 3.1. The perceived quality of the super-resolved image is unsatisfactory due to the presence of jagged, twisted or blurred contours.

The objective of this chapter is to seek an efficient but effective method that is capable of producing a high quality artifact-free high-resolution (HR) image from a single low-resolution (LR) input. Specifically, the single image SR is divided straightforwardly into two consecutive steps: magnification and deblurring, which is the inverse of the image acquisition pipeline. Magnification is to interpolate the image to the desired spatial resolution. Apart from blurring problem, the current prevalent scene-independent interpolators like pixel replication or bicubic interpolation fail to preserve the edge structures and thus suffer from severe jaggy artifacts (see Figure 3.2). Hence, we propose to accomplish the magnification in a structure adaptive manner. A recently developed adaptive interpolator called soft-decision adaptive interpolation (SAI) [Zhang and Wu, 2008] is adopted in the chapter due to its good performance on suppressing jaggy artifacts. However, as shown in Figure 3.1(b), the interpolated image is still far from satisfactory due to the blurring problem. Consequently, a deblurring step is introduced

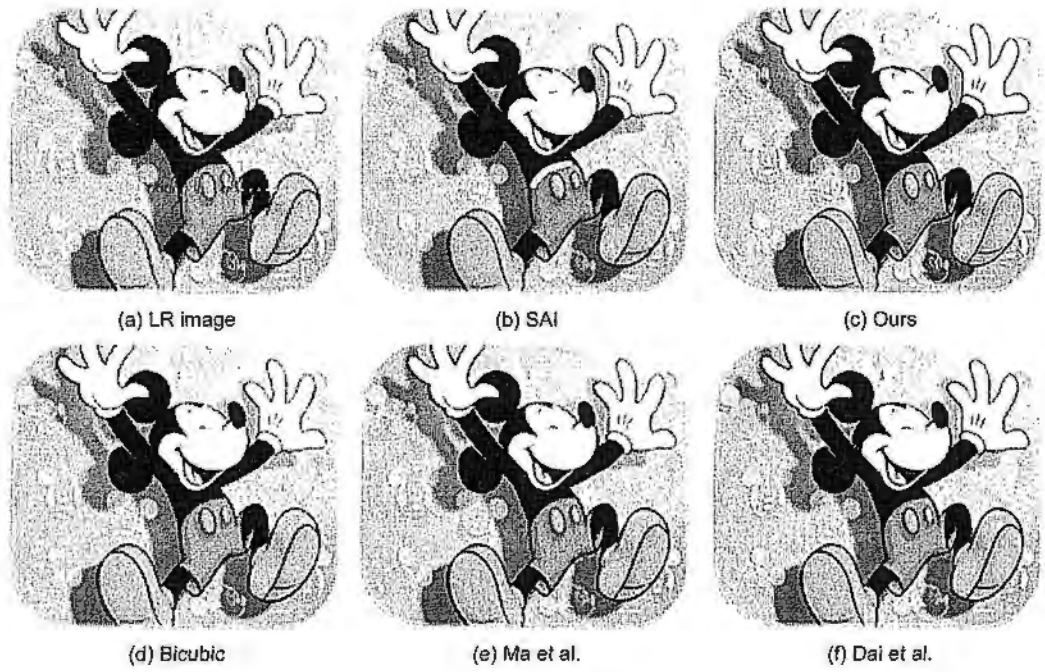

Figure 3.1: *3X SR on* Mickey. *(a) LR image. (b) Intermediate result obtained by SAI [Zhang and Wu, 2008], where the blue stroke outlines the salient edge used in the deblurring process. (c) Our final result. (d) Bicubic interpolation result. (e) Ma et al.'s result [Ma et al., 2008]. (f) Dai et al.'s result [Dai et al., 2009].*

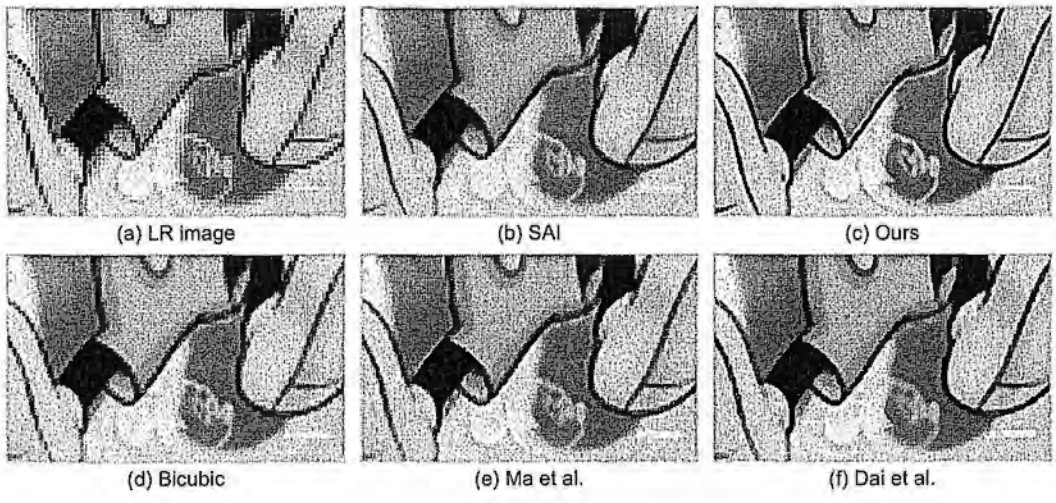

Figure 3.2: *Close-up comparison of results in Figure 3.1. Apparently, our result is free of the annoying visual defects such as jaggies and blurring associated with those of the other algorithms.*

to seek an appropriate blurring kernel to recover the sharpness of the interpolated result. Apparently, the latter step is a challenging blind deconvolution process which aims at inferring the sharp image as well as the blurring kernel simultaneously from a degraded blurry input.

Fortunately, edges can reveal the blurring information of the interpolated image. In particular, we advocate using a parametric edge model to extract the blurring kernel from the *salient* edges which refers to the pixels appear at the boundary area containing two adjacent parts with distinct colors. The *salient* edges are used because they are predictable and expected to be sharp in the output HR image, and we prefer the users to select them manually through the manner of user-drawn stroke. As illustrated in Figure 3.1(b), a single stoke is normally good enough to produce a desirable result. It is unnecessary to pick all salient edges and so the user intervention required in this work is little.

It is worth noting that the deblurring algorithms in [Joshi et al., 2008] and [Jia, 2007] share similar spirit with ours in blur kernel estimation. However, their performance relies on the success of some in-between results which is instable in tough conditions. For example, Joshi et al. [Joshi et al., 2008] needs to create a sharp edge by roughly changing the edge profile, while transparency estimation is required in Jia [Jia, 2007]. The blurring kernel is estimated from these intermediate results with a maximum a posteriori (MAP) estimator. In contrast, our method is quite computationally efficient and the blurring kernel is calculated directly from the stable edges (salient) and in closed form. When the blurring kernel is fixed, the sharp HR image is recovered efficiently with a MAP framework.

The rest of this chapter is organized as follows. Section 3.2 gives a brief review of the related work. The proposed SR framework is described in Section 3.3. Experimental results are presented in Section 3.4. Section 3.5 draws some concluding remarks.

3.2 Related Work

There are a large number of algorithms to address the single-image SR problem in the past years. Conventional interpolation methods like Bilinear or Bicubic yield jaggied and blurred edges, degrading image details. The structure adaptive methods such as [Li and Orchard, 2001] and [Zhang and Wu, 2008] work better on eliminating jaggies but

still suffer from blurring problem.

Learning-based methods such as [Freeman et al., 2000; Sun et al., 2003; Wang et al., 2005; Ma et al., 2008] can output sharp HR image even with high magnification factor by making use of additional training HR and LR image pairs directly or indirectly. However, as mentioned in Chapter 1.2.3, their performance replies on how well the input LR image matches the training samples. Therefore sufficient number of appropriate training samples are required to guarantee the SR performance. Reconstruction-based methods like [Zomet et al., 2001; Shan et al., 2008b] are built based on a generative imaging model which simulates how the HR scene is transformed, filtered and sampled to give rise to LR images. However, the reconstruction of HR image from LR input is a typical ill-posed inverse problem. Therefore, to regularize the ill-posedness, image priors are introduced to impose additional constraints in the SR process. Studies on image statistics [Roth and Black, 2009] show that sparse prior is a sound choice due to the heavy tailed property of image response to a collection of convolution filters. Recently, edge-based priors [Fattal, 2007; Sun et al., 2008; Dai et al., 2009] are developed to further preserve the edge sharpness. More recently, Glasner et al. [Glasner et al., 2009] combined the reconstruction-based method and learning-based method and presented a unified framework that can be applied to single image SR without any additional external data.

Nevertheless, the above methods have the following problems that limits their applicability in practice. One problem is that these methods often have complex working pipeline where the performance replies on a number of factors such as parameter tweaking, iteration number or training set quality. In addition, the blurring process of the optic is assumed to be known as a priori in these methods. That is, the camera's point spread function (PSF) kernel can be fixed in advance. Even the learning-based methods also require the knowledge of PSF to make sure that the training and testing data are degraded in the same way. However, this assumption does not hold in real scenarios because of the unpredictable behavior of PSF. Different images may be formed with different PSFs due to the influence of camera lens, focusing condition and so on. Although some methods like [Jia and Gondi, 2005; Yang et al., 2008a] were presented to iteratively estimate the PSF and the latent HR image, their applicability is limited due the high computational complexity and multiple LR input requirement. Finally, most

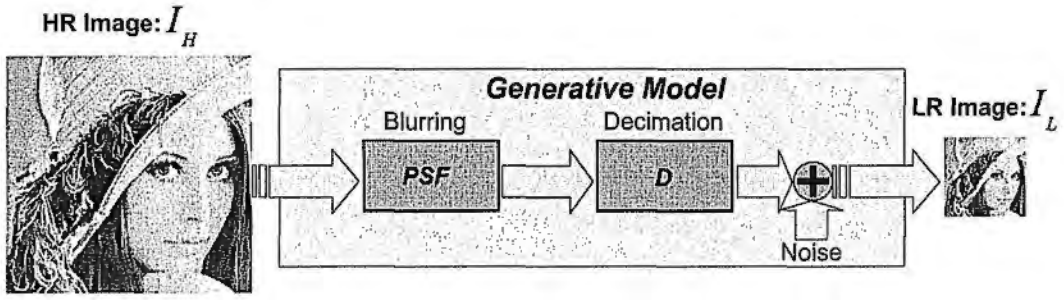

Figure 3.3: *The imaging model in single image SR.*

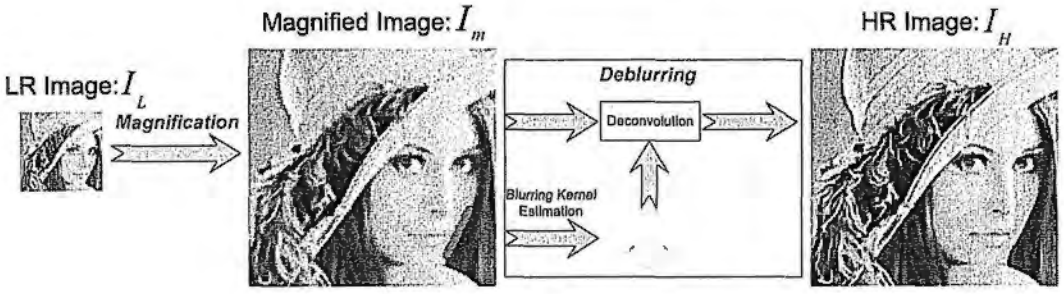

Figure 3.4: *The proposed SR scheme.*

existing methods ignored the structure coherence of HR image and LR image, and thus are susceptible to artifacts like jaggies.

3.3 Problem Formulation and The Proposed Algorithm

The generative imaging model in single image SR is shown in Figure 3.3. The observed LR image can be produced by blurring and downsampling a HR image. Namely, $I_L = D \downarrow (PSF \otimes I_H) + noise$, where $\otimes$ denotes the convolution operation. PSF is the camera's PSF which is normally assumed to be a Gaussian filter. The validity of Gaussian PSF assumption has been proved in [Capel, 2004]. SR is an inverse process which is to recover the HR image from a LR input.

3.3.1 Structure Adaptive Interpolation

In this work, we propose to divide the SR straightforwardly into two consecutive steps as illustrated in Figure 3.4. A LR image I_L is first magnified to the desired resolution using interpolation. As aforementioned, conventional scene-independent interpolators

like pixel replication or bicubic interpolation fail to preserve the edge structures and thus suffer from severe jaggy artifacts (see Figure 3.2). To overcome this problem and achieve structure adaptive interpolation, some algorithms such as [Li and Orchard, 2001; Zhang and Wu, 2008] were presented by utilizing a piecewise autoregressive (PAR) progress to model a natural image. That is, each pixel $I_m(x, y)$ can be approximated by the linear combination of its neighboring pixels $I_m(x+i, y+j)$.

$$I_m(x, y) \approx \sum_{(i,j)} \Omega(i, j) I_m(x+i, y+j), \tag{3.1}$$

where (i, j) defines the spatial neighborhood. The weights $\Omega(i, j)$ imply the the local structure around the current pixel and can be assumed to be locally stationary. Especially, Zhang and Wu [Zhang and Wu, 2008] proposed a soft-decision estimation interpolation framework based on the PAR model that achieves superior performance than the other work. To better preserve the local structure, the missing pixels are jointly estimated in SAI by enforcing the local relation not only between known pixels and missing pixels but also between missing pixels themselves. Besides, SAI operates on blocks of pixels and thus runs efficiently. In this work, we adopt the SAI to accomplish the image magnification.

As shown in Figure 3.4, although SAI has impressive performance on eliminating jaggy artifacts, the magnified image is still undesirable due to the severe blurring problem. As depicted in (3.2), the blurry magnified image I_m is regarded as the convolution result of a sharp image I_H with a blurring kernel f.

$$I_m = f(\sigma_f) \otimes I_H + n, \tag{3.2}$$

where $f(\sigma_f)$ is a smoothing function and depends on some kind of smoothing parameter $\sigma_f > 0$. n is an additive noise and normally assumed to be Gaussian. Similar to the camera's PSF, the blurring kernel f can be reasonably assumed to be a 2-D Gaussian filter g with standard deviation σ. Hence, $f(\sigma_f) = g(\sigma)$ with $\sigma_f = \sigma$. Note that the blurring kernel f normally has larger scale than the latent camera's PSF, as interpolation incurs additional blurriness.

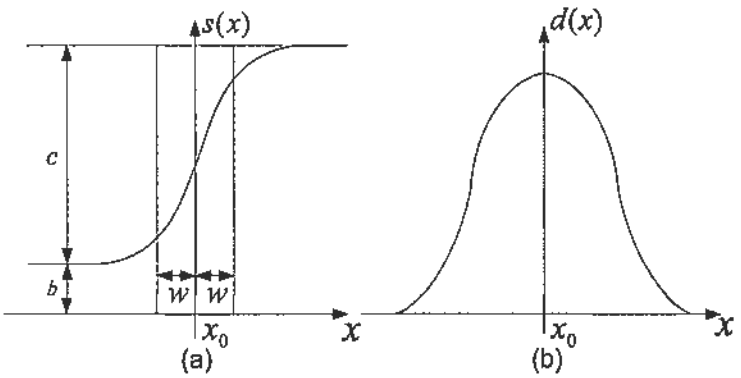

Figure 3.5: *(a) 1-D parametric edge model. (b) Response of convolving an edge with the derivative of a Gaussian filter.*

3.3.2 Deblurring from Salient Edge

The goal of this section is to seek a suitable deblurring scheme which can recover the sharpness of the magnified image result with appropriate blurring kernel. As shown in Figure 3.4, the deblurring process has two tasks: blurring kernel estimation and sharp image recovery. They will be addressed one by one in the following sections.

Blurring Kernel Estimation

As shown in (3.2), edges in I_m can be obtained by Gaussian blurring the corresponding edges in I_H with $f(\sigma_f)$. Next, we adopted a parametric edge model to depict edges motivated by [van Beck, 1995; Fan and Cham, 2000]. Without loss of generality, we take the 1-D form to explain the edge model, since edges in a 2-D image can be characterized by sharp intensity changes in one direction. Mathematically, a step edge at x_0 can be depicted as $e(x; b, c, x_0) = cU(x - x_0) + b$ where $U(\cdot)$ is the unit step function, b denotes the edge basis and c represents the edge contrast. A typical edge $s(x; b, c, w, x_0)$ can be regarded as a smoothed step edge which is obtained by convolving $e(x; b, c, x_0)$ with a 1-D Gaussian filter $g(\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-x^2}{2\sigma^2})$ and so

$$s(x; b, c, w, x_0) = b + \frac{c}{2}\left(1 + erf(\frac{x - x_0}{w\sqrt{2}})\right), \tag{3.3}$$

where $erf(\cdot)$ is the error function. As shown in Figure 3.5(a), w is equal to σ and determines edge blurriness and can also be referred as edge width. The larger w is, the blurrier the edge is. Roughly speaking, all edges can be depicted parametrically

by fitting (3.3) on them. Such fitting process includes two steps: edge detection and parameter estimation.

Similar to that of Canny detection [Canny, 1986], edge is detected by convolving $s(x; b, c, w, x_0)$ with the derivative of a predefined Gaussian filter $g'_d(\sigma_d)$. The response is:

$$d(x; c, w, \sigma_d, x_0) = \frac{c}{\sqrt{2\pi(w^2 + \sigma_d^2)}} \exp\left(\frac{-(x - x_0)^2}{2(w^2 + \sigma_d^2)}\right). \tag{3.4}$$

The peak of the response can be used to locate the edge as illustrated in Figure 3.5(b). Also, the standard deviation w of the blurring filter g as well as the other parameters of (3.3) can be estimated as (3.5)-(3.8) based on three measurements which are selected by sampling the response $d(x; c, w, \sigma_d, x_0)$ at $x = 0, a, -a$. They are: $d_1 = d(0; c, w, \sigma_d, x_0)$, $d_2 = d(a; c, w, \sigma_d, x_0)$ and $d_3 = d(-a; c, w, \sigma_d, x_0)$. a is normally set to 1.

$$w = \sqrt{a^2/ln(l_1) - \sigma_d^2}, \tag{3.5}$$

$$x_0 = 0.5 \cdot a \cdot ln(l_2)/ln(l_1), \tag{3.6}$$

$$c = d_1 \cdot \sqrt{2\pi a^2/ln(l_1)} \cdot l_2^{\frac{1}{4a}}, \tag{3.7}$$

$$b = s(x_0) - c/2 \tag{3.8}$$

where $l_1 = \frac{d_1^2}{d_2 d_3}$ and $l_2 = d_2/d_3$. The above analysis can be extended to the 2-D case directly except that an extra parameter θ is required to represent the edge direction. Please refer to [van Beek, 1995] for more details. Therefore, we can similarly recover the 2-D Gaussian blurring filter $f(\sigma_f)$ of (3.2) by measuring the edge blurriness in I_m based on (3.5).

However, it is noted that only the salient edges of I_m can be used in the blurring kernel estimation due to the following reasons: First and most importantly, their corresponding edges in latent HR image I_H are very sharp and have rapid transition similar to step edges. Thus the estimated blurriness (i.e. w) can fully reflect the true blurring difference between I_H and I_m. Secondly, the salient edges are stable indicators of the blurring kernel and quite detectable even if the feature strength of the interpolated image I_m is weakened considerable. As exemplified in Figure 3.6, to avoid the influence

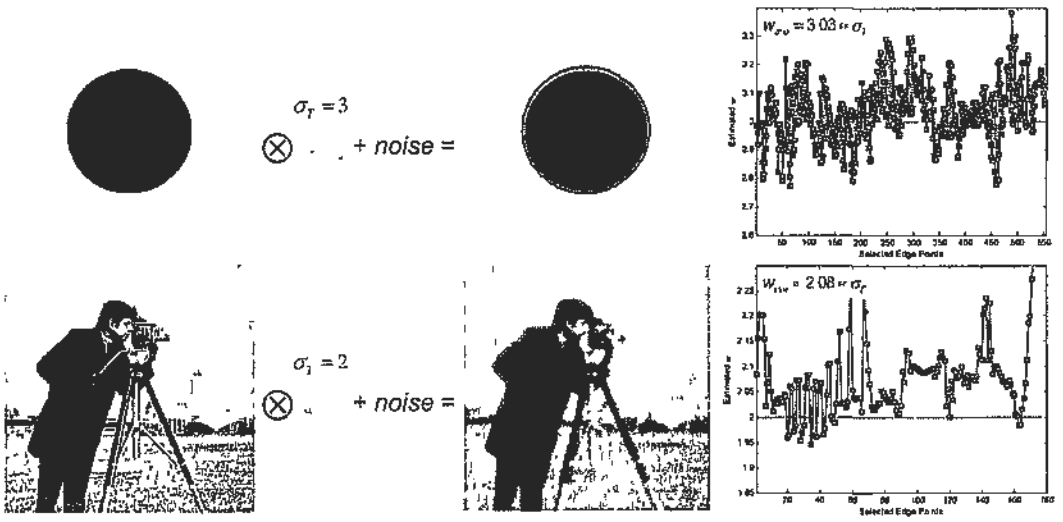

Figure 3.6: *Testing on kernel estimation. The selected distinct edges are outlined with red Red squares in the right figures show the estimated w of all the points in the selected edges and* w_{aver} *denotes the average estimation. Blue lines show the ground truth standard deviation* σ_T

of neighboring edges and make blurring kernel estimation more stable, the salient edge locating at boundary of two adjacent regions with distinct colors is favored for blurring kernel estimation.

Two simulated examples are given in Figure 3 6 to test the proposed kernel estimation scheme. We first blur the original image by Gaussian filter with σ_T and then try to recover this ground truth filter by measuring the blurriness of the selected distinct edges in the degraded image. The results show that all of the estimated standard deviations are close to the ground truth σ_T. To make the estimation more robust, we use their average w_{aver} for deblurring in the following experiments.

Sharp Image Recovery

Once $f(\sigma_f)$ is fixed ($\sigma_f = w_{aver}$), recovering the target sharp image I_H from the blurred I_m becomes a typical non-blind deconvolution task. This problem is still challenging and effective image prior is required to regularize its ill-posedness. As shown in (3 9), the task is tackled with a MAP estimator.

$$I_H^* = \arg\max_{I_H} \; p(I_m|I_H, f(\sigma_f))p(I_H). \tag{3.9}$$

$p(I_H)$ can be defined with a sparse derivative prior to favor sharp image result. Hence, (3.9) is rewritten as:

$$I_H^* = \arg\min_{I_H} \alpha \|f(\sigma_f) \otimes I_H - I_m\|_2^2 + \sum_{i=1}^{2} \|(t_i \otimes I_H)\|_2 , \tag{3.10}$$

where α is a weighting factor. t_i is simply defined with the first order derivative filter as: $t_1 = [1 \ -1]$ and $t_2 = [1 \ -1]^T$.

Apparently, directly optimizing (3.10) is difficult and computationally demanding. Inspired by [Wang et al., 2007], a variable-splitting and penalty technique is employed to render the optimization more tractable. In brief, an auxiliary variable ξ_i is introduced to transfer $t_i \otimes I_H$ out of the non-differentiable term $\|\cdot\|_2$ and the difference between them is penalized with a quadratic term. Thus, (3.10) turns to be:

$$\begin{aligned} I_H^* = \arg\min_{I_H} \ & \alpha \|f(\sigma_f) \otimes I_H - I_m\|_2^2 + \sum_{i=1}^{2} \|\xi_i\|_2 \\ & + \beta \sum_{i=1}^{2} \|(\xi_i - t_i \otimes I_H)\|_2^2 . \end{aligned} \tag{3.11}$$

The penalty factor β increases by 2 times after each iteration and the iteration will be stopped once the stopping criterion is satisfied. The solution of (3.11) converges to that of (3.10) as β becomes very large. Please refer to [Wang et al., 2007] for more details. α depends on the noise level and normally ranges from 500 to 1000. It is worth noting that the optimization of (3.11) can be solved efficiently, since when one of the two variables I_H and σ_f is fixed, minimizing the function w.r.t the other has a closed-form solution. Moreover, the solver can be accelerated greatly by performing Fast Fourier Transform (FFT) to avoid the computational complexity caused by convolution.

3.4 Experimental Results

In this section, we first generated some synthesized examples to test the effectiveness of the proposed algorithm and evaluated its performance quantitatively. As shown in Figure 3.7, the original HR images in (d) are firstly blurred and then downsampled by a factor of 3 to yield the LR images shown (a). Subjectively, our results (c) are more visually pleasant than (b) generated by SAI [Zhang and Wu, 2008] due to the significant improvement on sharpness. We further assessed the sharpness improvement

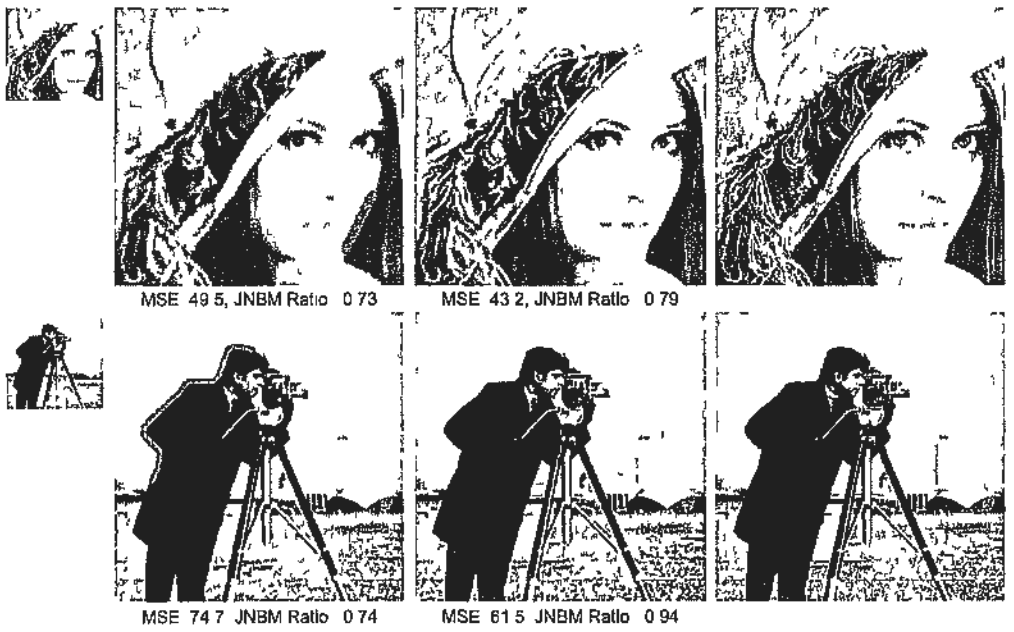

Figure 3.7 *Testing on synthesized examples In each row, images from left to right are the input LR image, the intermediate result interpolated using SAI, our final SR result and the original HR image The selected distinct edges are outlined with red in the intermediate magnified image using SAI*

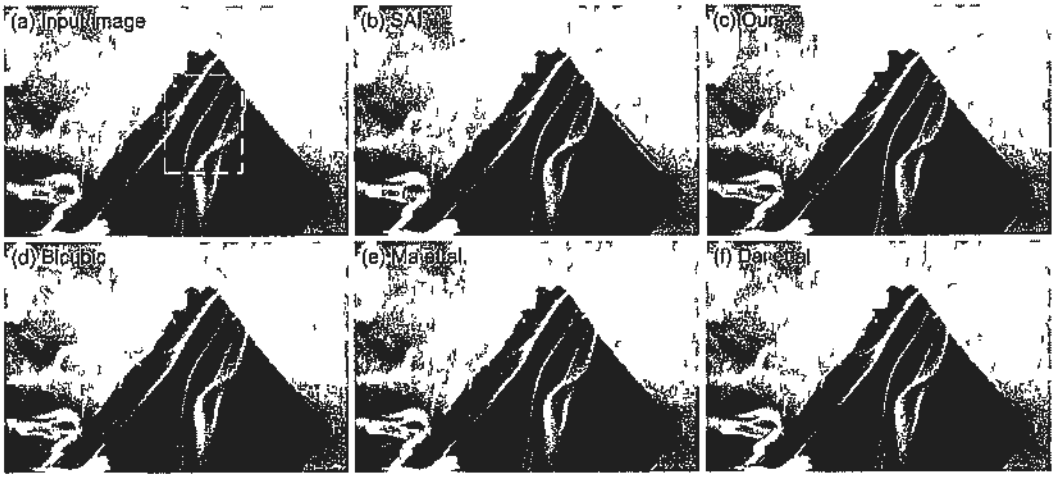

Figure 3.8 *SR on* Fire *with a magnification factor of 3 (a) LR image (b) Intermediate result obtained by SAI [Zhang and Wu, 2008], where the red stroke outlines the salient edge used in the deblurring process (c) Our final result (d) Bicubic interpolation result (e) Ma et al 's result [Ma et al 2008] (f) Dai et al 's result [Dai et al , 2009]*

quantitatively with Just Noticeable Blur Metric (JNBM) [Ferzli and Karam 2009], which is a perceptual-based no-reference sharpness metric and can predict the relative

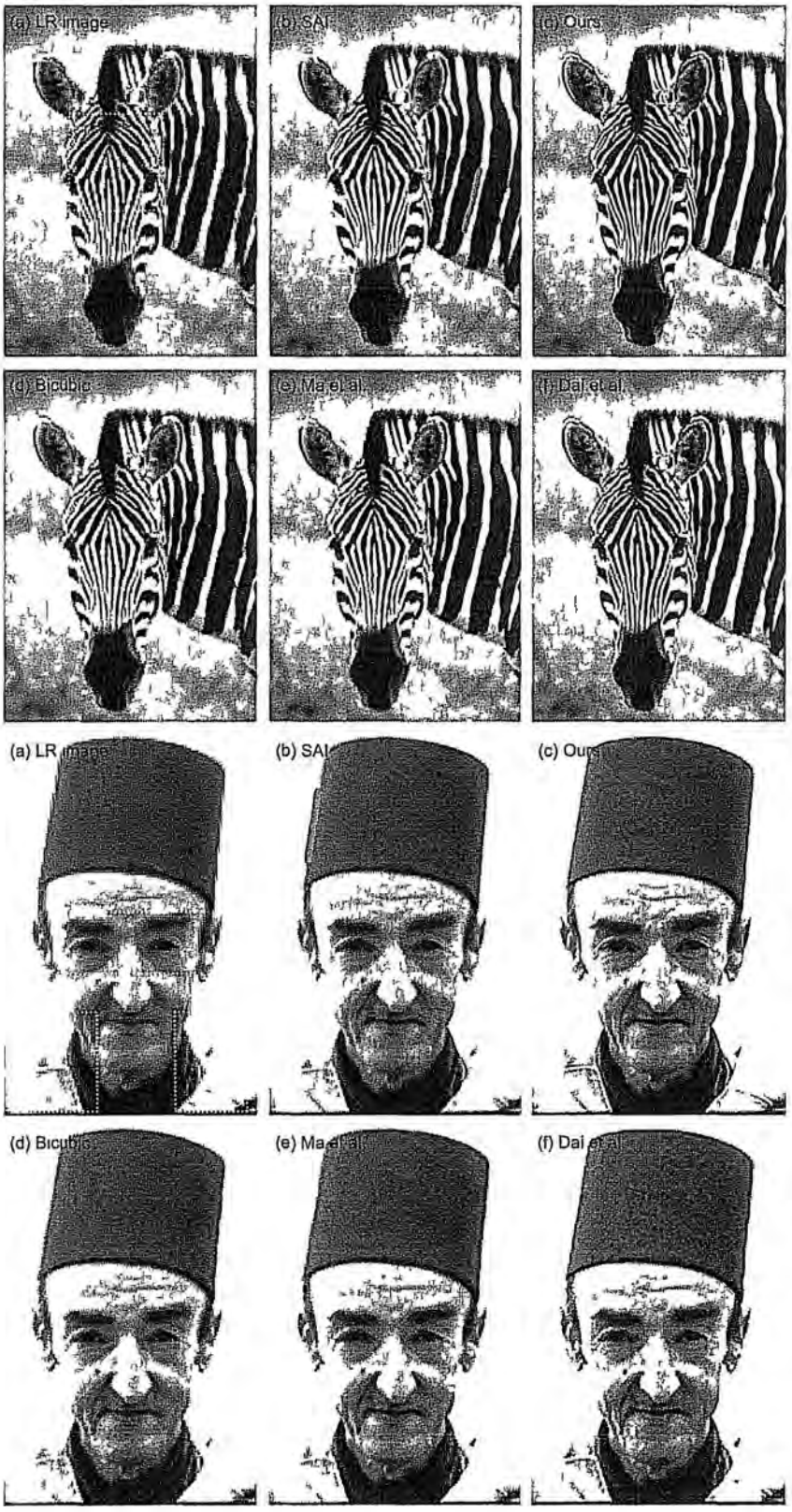

Figure 3.9 *SR on Zebra and Man with a magnification factor of 3 (a) LR image (b) Intermediate result obtained by SAI [Zhang and Wu 2008], where the red strokes outline the salient edge used in the deblurring process (c) Our final result (d) Bicubic interpolation result (e) Ma et al 's result [Ma et al 2008] (f) Dai et al 's result [Dai et al 2009]*

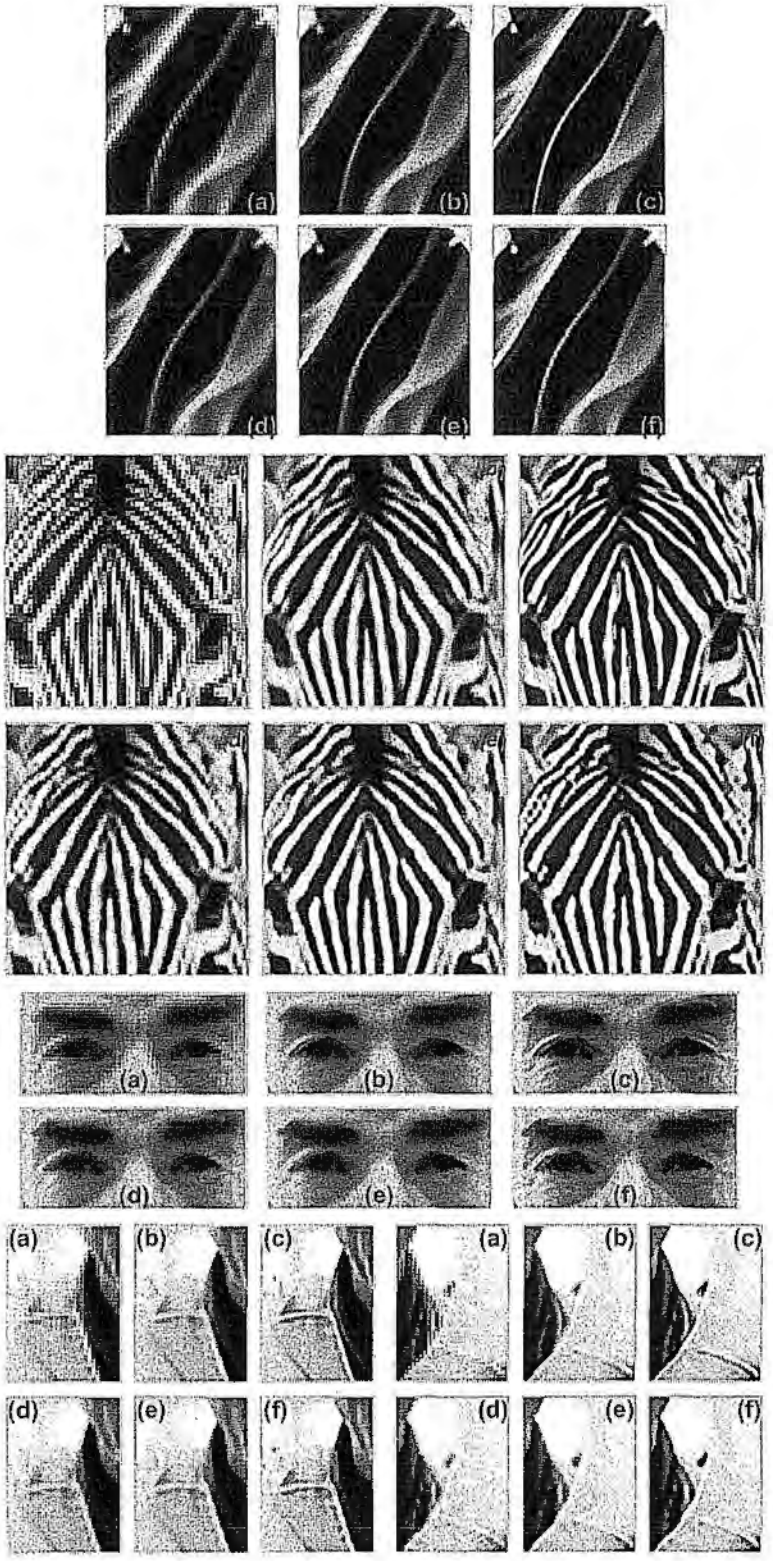

Figure 3.10: *Close-up comparison. The patches (a)(b)(c)(d)(e)(f) are cropped from LR image, interpolation result using SAI, our result, Bicubic interpolation result, Ma et al.'s result, Dai et al.'s result.*

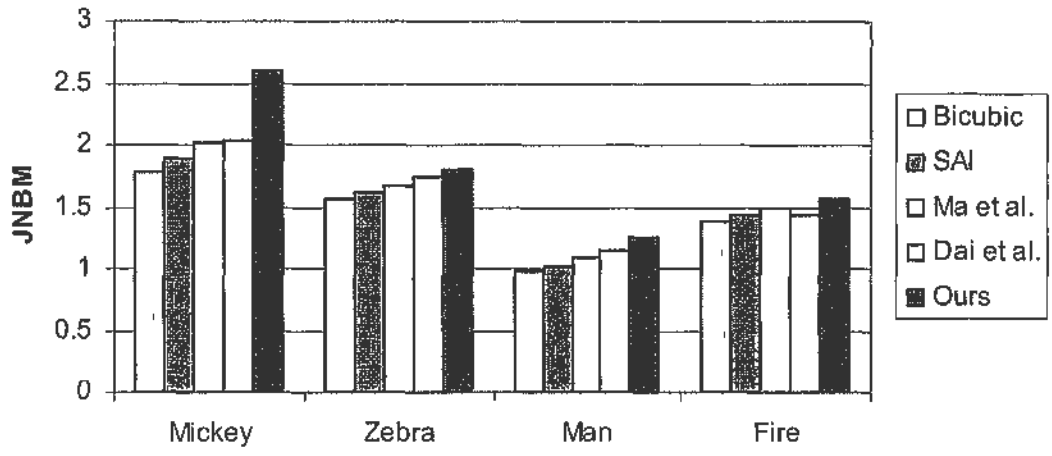

Figure 3.11: *Quantitative evaluation on sharpness with JNBM. High JNBM value represents the image has good sharpness.*

amount of blurriness in images with high accuracy. Since the original image is known, we use the ratio (i.e., $JNBM(I_{sr})/JNBM(I_{ori})$) between the JNBM value of the super-resolved image and that of the original image to show the sharpness performance. Thus, the larger the ratio is, the sharper the super-resolved image is and when the ratio equals to 1, the super-resolved image has the same sharpness as that of the original image. Besides, we also assessed the improvement in terms of the mean square error (MSE). The above two quantitative measures both prove that proper deblurring improves interpolation performance and gives more faithful SR results to the original images. As shown in Figure 3.1 and Figure 3.10, we reproduced some results published before and made comparisons with the existing work. In summary, the results obtained by Ma et al. [Ma et al., 2008] suffer from blurring and jaggy artifacts. The overall visual quality of Dai et al.'s results [Dai et al., 2009] and ours are comparable. However, as illustrated in the close-up comparison, Dai et al.'s results are less detailed and prone to jaggy artifacts especially around the long edge area which substantially degrade the perceptual quality. While, our results are not only free of artifacts but also exhibit the best sharpness as shown in Figure 3.11. Note that the above results of [Ma et al., 2008] and [Dai et al., 2009] are produced by the authors. Please also visit `http://www.ee.cuhk.edu.hk/~zhangwei/HighQualitySR.html` to see the results.

It is worth noting that the choice of salient edge may not be unique in practice. For most images, there are several edges that can serve in the blurring kernel estimationas.

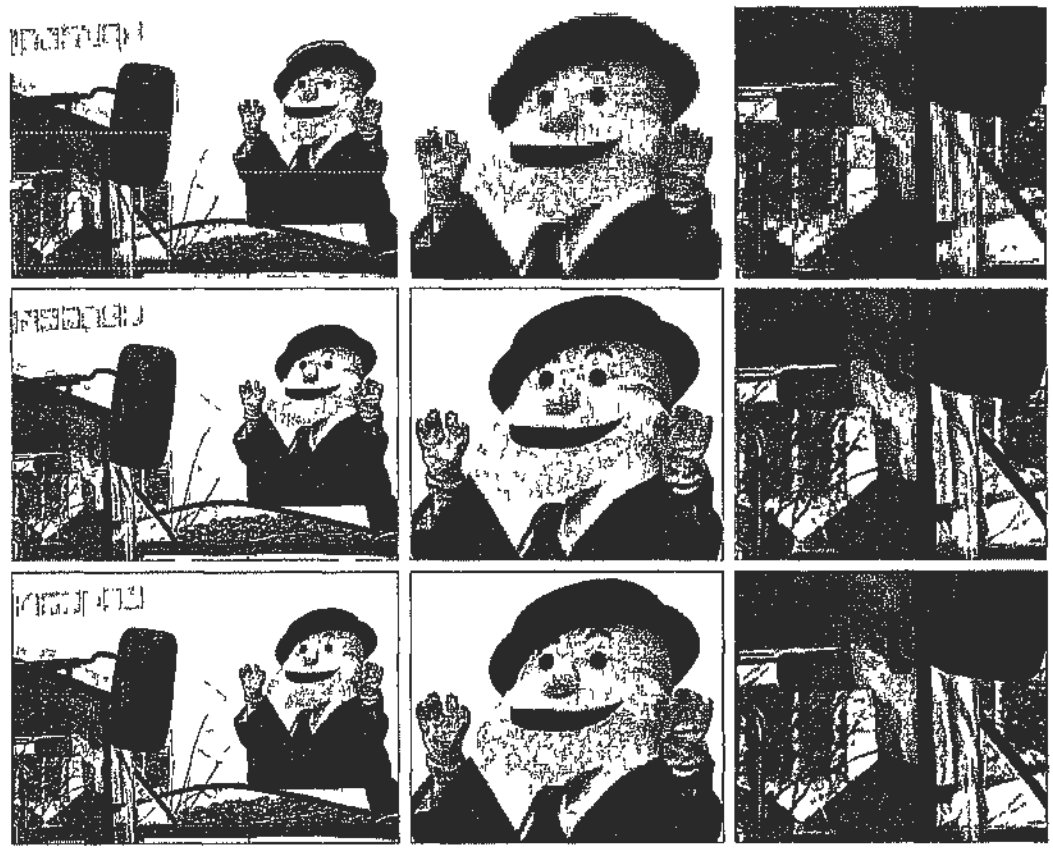

Figure 3.12 *SR with different salient edges Top LR image Middle and bottom show the 3 times HR results super-resolved with the salient edges outlined by blue and red strokes, respectively*

As illustrated in Figure 3 12, two salient edges (drawn by blue and red stokes) can both be used in the deblurring process and result in quite similar standard derivations 1 35 (blue) and 1 32 (red) Hence, the deblurred HR images are similar as well

Apart from the good performance, our method is appealing due to its low computational complexity To super-resolve a 352 × 288 image, the current non-optimized Matlab implementation takes less than 20 seconds There is still much room to improve its efficiency by optimized C++ or GPU implementation It is interesting to note that most of the computation cost comes from the sharp image reconvery (non-blind deconvolution) The blurring kernel estimation can be finished very rapidly (normally less than 2 second), since only the pixels of the salient edge (instead of the entire image) are needed to be taken into computation In summary, the key idea we advocate here is that it may not be necessary to make the SR that complex Alternatively, a simple

framework plus little user assistance can also achieve impressive SR performance.

3.5 Summary

In this chapter, a simple but effective algorithm is presented to address the challenging single image SR problem. To create a pleasant artifact-free HR result, we first magnified the LR image to the desired resolution through structure adaptive interpolation and then introduce a salient edge directed deblurring scheme for sharpness recovery. Unlike most existing work, the camera's PSF is not assumed to be known in this work. Experiments demonstrate that the proposed approach produced high quality results both perceptually and quantitatively. Nevertheless, since the underlying principle is to take advantage of the salient edge to seek suitable deblurring, the standard derivation σ_f of the blurring kernel cannot be estimated accurately if no salient edges can be found in the magnified image. In this case, we have to resort to parameter tweaking to find the optimal σ_f.

Chapter 4

Single Image Focus Editing

4.1 Introduction

Single image refocusing and defocusing is an interesting research topic and has received a lot of attention recently. Two tasks are mainly involved in this topic. One is image refocusing which is to recover the sharpness of the blurry defocused objects in an input image and generate a virtual all-focused image. The other is defocusing which is to blur an image and create defocus effects. In some photography such as portrait, shallow depth of field (DOF) is preferred so as to highlight the foreground subject with a defocused blurry background. But due to the limitations of the lens and sensors, some cameras such as point-and-shoot cameras cannot produce enough defocus effects.

In this chapter, we present a novel method which is able to handle the tasks of focus map estimation, image refocusing and defocusing. One example is shown in Figure 4.1, where an input image (a) contains focused foreground object and defocused background which includes the girl. Firstly, if we find the defocus effect on the background is not adequate, then the proposed method can be used to increase the defocus effect of the background while keep the foreground unchanged. Hence, a portrait-like image similar to that formed by using a shallower DOF is produced as shown in (b). Secondly, the proposed method can be used to refocus the defocused background to generate a plausible all-focused image as shown in (c). Besides, as shown in (d), the highlight of the input image can also be changed after defocusing the original focused foreground on the synthesized all-focused result. The comparison in (f) indicates that the proposed refocusing method outperforms the lens deblurring of *Photoshop* significantly.

The proposed method first estimates a focus map and then use it to separate the focused and defocused objects as shown in Figure 4.1(e). The focus map estimation is based on the assumption that blurring of edges is only due to the defocus effect and so

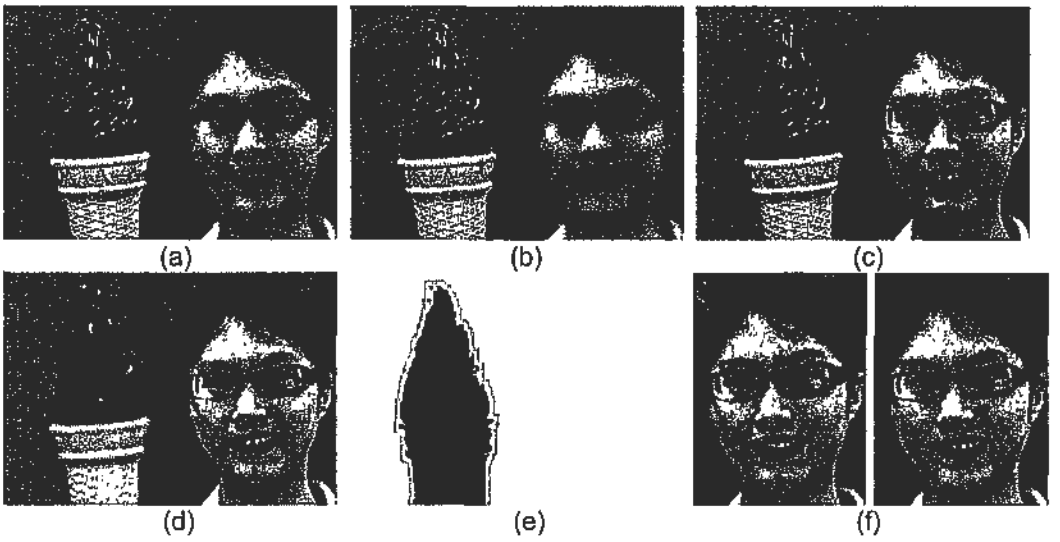

Figure 4.1: *(a) Input narrow aperture image focusing on the foreground object. (b) Synthesized image with shallower DOF. (c) Synthesized all-focused image (d) Synthesized image focusing on the background. (e) The detected focus mask (white: defocused regions, black: focused regions, gray: focus boundaries). (f) Close-up comparison. Left: removing the lens blur using the lens delurring in smart sharpen of* Photoshop. *Right: our refocused result.*

the focus information can be indicated by edge blurriness. A parametric edge model based scheme is presented to generate the focus map automatically. More specifically, we will first measure the blurriness for edge pixels and then propagate it from the edge pixels to their neighboring non-edge pixels based on the similarities of intensity and position.

In this work, refocusing is formulated as a single-image blind deconvolution (SBD) problem based on the fact that the defocused image can be regarded as a result of convolving the focused image with a point spread function (PSF). Therefore, the challenge is to infer the sharp focused image as well as the PSF simultaneously from a degraded blurry image. To regularize this unconstrained problem effectively, two additional local prior models are introduced in the proposed SBD framework besides a global image prior. One novel sharp prior is adopted to ensure the sharpness of the refocused image. Another local smooth prior is to constrain the low-contrast regions unchanged for suppressing the ring artifacts. Our study shows that their combination, named as Sharp-and-Smooth prior, provides an effective regularization for ensuring image sharpness and suppressing ring artifacts. Extensive experiments on synthesized and real images were performed to test the proposed SBD method. Defocusing in this

work is handled by Gaussian blurring as stated in Sec 4.3.1.

The rest of this chapter is organized as follows. Section 4.2 gives a brief review of the related work. Some foundational knowledge about image formation and edge model is introduced in Section 4.3. Section 4.4 introduces the edge model based method for focus map generation. A new SBD approach is proposed in Section 4.5 for image refocusing. Experimental results on refocusing and defocusing are shown in Section 4.6. Section 4.7 draws some concluding remarks.

4.2 Related Work

Focus and defocus cues are popular for the recovery of depth map in the study of depth from focus and defocus [Schechner and Kiryati, 2000; Rajagopalan et al., 2004], where multiple images with different focus settings are required to estimate a depth map for the latent scene. For example, Rajagopalan et al. [Rajagopalan et al., 2004] proposed a depth estimation method by combining the defocus and stereo cues. While in this work, we are concerned with extracting the focus information instead of accurate depth from a single image.

As mentioned in Section 1.2.2, lots of efforts have been made to address image refocusing and defocusing. The hardware solutions (e.g. [Ng et al., 2005; Levin et al., 2007; Moreno-Noguer et al., 2007]) requires additional optical elements or devices to help the camera capture more information about the target scene. Some postprocessing-based methods [Kubota et al., 2004; Kubota and Aizawa, 2005; Hasinoff and Kutulakos, 2007; Yang and Schonfeld, 2010] were presented based on multiple images of the same scene. This chapter concentrates on achieving image refocusing and defocusing from a single input without changing the camera, but with only image processing. Here, we just refer some methods which are the most similar to ours. Bae and Durand [Bae and Durand, 2007] contributed at proposing an automatic focus map estimation method by estimating the edge blurriness with a brute-force fitting strategy. The defocusing there is handled with the aid of the lens blur tool in *Photoshop*. In this chapter, a simple and well-parameterized multi-point scheme is adopted to measure the edge blurriness. Besides defocusing, we also address the more challenging refocusing problem with a blind deconvolution framework. The edge information is exploited not only in focus detection but also in image refocusing in this chapter. Yan et al. [Yan et al.,

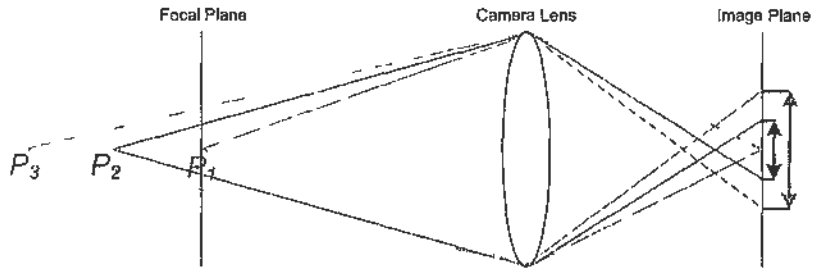

Figure 4.2· *Geometry of the imaging model* P_1, P_2 *and* P_3 *represent the scene points at different depth*

2009] developed an interactive defocusing system, where user intervention is required to obtain the depth information of an input image. Similarly, Bando and Nishita [Bando and Nishita, 2007] presented an interactive method to address single image refocusing, where lots of user intervention is needed to determine the blur kernel from a number of predefined candidates. While in this work, focus map, blur kernel and refocused image are all obtained automatically.

4.3 Background and Problem Formulation

4.3.1 Imaging Model

As shown in Figure 4.2, the rays originating from a scene point P_1 on the focal plane can converge to a point on the image plane. However, when the scene point moves away from the focal plane, the rays will give rise to a blur circle on the image plane and the image is regarded as defocused. When the point moves farther, a blurrier defocused image is produced. Such blurring process is often modeled as the convolution of a focused image I_F with a PSF h, i.e

$$I_D = h \otimes I_F + n, \tag{4.1}$$

where I_D denotes the defocused image and n is the noise term. Due to the diffraction and aberration of the camera lens, the PSF is normally approximated by a 2-D Gaussian filter [Lin and Chang, 2006; Favaro and Soatto, 2005] given by $g(x, y; \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(x^2+y^2)}{2\sigma^2})$. The spread parameter σ which is related to the distance of the object to the focal plane determines the blurriness of the captured image. In this chapter, (4.1) and Gaussian PSF are used to model the defocusing process.

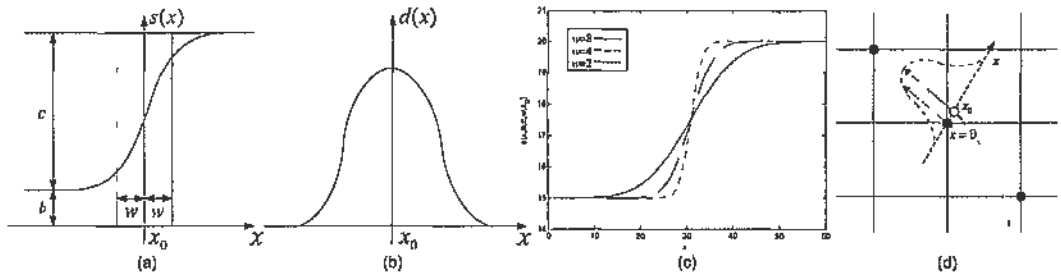

Figure 4.3: *(a) 1-D parametric edge model. (b) Response of convolving an edge with the derivative of a Gaussian filter. (c) Effect of decreasing w. (d) Edge detection in an image, where the gray line outlines the contour of an edge, the solid dots are the detected peak positions located at the grid points, the circle is one of the true edge positions.*

It is worth noting that unlike the multi-image based approach [Hasinoff and Kutulakos, 2007], the occlusion problem is not formulated directly into the imaging model in this chapter, because the single-image based work itself is already highly unconstrained and adding more unknowns will make the whole framework intractable. However, owing to the estimated focus map, we can locate the possible occlusion regions such as the gray regions in Figure 4.1(e) along the layer boundaries and then use alpha blending to synthesize these regions to avoid artifacts. The results show that this is a visually realistic way to handle the occlusion problem.

4.3.2 Edge Modeling

Focus map estimation on a single input is challenging. Fortunately, edges in an image carry important information which may hint how the image is formed. Similar to Section 3.3.2, a parametric edge model [van Beek, 1995; Fan and Cham, 2000] is adopted for edge description. As shown in Figure 4.3(a), a typical edge $s(x; b, c, w, x_0)$ can be represented mathematically as:

$$s(x; b, c, w, x_0) = b + \frac{c}{2}\left(1 + erf(\frac{x - x_0}{w\sqrt{2}})\right), \tag{4.2}$$

where $erf(\cdot)$ is the error function. b denotes the edge basis. c represents the edge contrast. w is referred as the edge width parameter. As shown in Figure 4.3(c), the edge is sharper when w becomes smaller. x_0 is a real number which can represent the edge location continuously. In practice, the position of the detected peak in a 2-D image is constrained to a grid point location which is represented by an integer and thus may

not coincide with the truth. As the example shown in Figure 4.3(d), the detected peak locates at $x = 0$ instead of the true edge position x_0. As stated in Section 3.3.2, all parameters of (4.2) can be estimated as follows:

$$x_0 = 0.5 \cdot a \cdot ln(l_2)/ln(l_1), \tag{4.3}$$

$$w = \sqrt{a^2/ln(l_1) - \sigma_d^2}, \tag{4.4}$$

$$c = d_1 \cdot \sqrt{2\pi a^2/ln(l_1)} \cdot l_2^{\frac{1}{4a}}, \tag{4.5}$$

$$b = s(x_0) - c/2 \tag{4.6}$$

where $l_1 = \frac{d_1^2}{d_2 d_3}$ and $l_2 = d_2/d_3$, and d_1, d_2 and d_3 are three sample measurements at $x = 0, a, -a$ of $d(x; c, w, \sigma_d)$, which is the response of convolving $s(x; b, c, \sigma, x_0)$ with the derivative of a predefined Gaussian filter $g'_d(x; \sigma_d)$. Value of a can be chosen freely and normally $a = 1$. Please see Section 3.3.2 for more details.

The above derivation can be referred as a multi-point estimation method. With this parametric model, the edge can be changed easily by controlling these parameters. For example in Figure 4.3(c), decreasing w will result in a sharper edge. Hence, edge in an image can be sharpened by first detecting the edges and estimating the parameters. Then the edge is reconstructed by substituting the new w' to (4.2) and keeping the other parameters unchanged.

4.4 Edge based Focus Map Estimation

Based on the above edge model, a method is proposed in this section to estimate the focus map automatically for an image containing a mixture of focused and defocused objects. This can be done automatically because the edge blurriness carries important cue about the focus setting. Different degrees of blurriness imply different defocus scales. Hence, focus map herein also corresponds to blurriness map. It is worth noting that this conclusion is under the assumption that the occurrence of edge blurriness is only due to defocusing effect, which is prevalent in previous focus map estimation work such as [Bae and Durand, 2007]. Our proposed method is also based on this assumption and thus shares the common limitation that it cannot estimate an accurate focus map for natural blurry objects like clouds and shadows.

The proposed scheme proceeds as follows. Firstly, as stated in Sec 4.3.2, a single

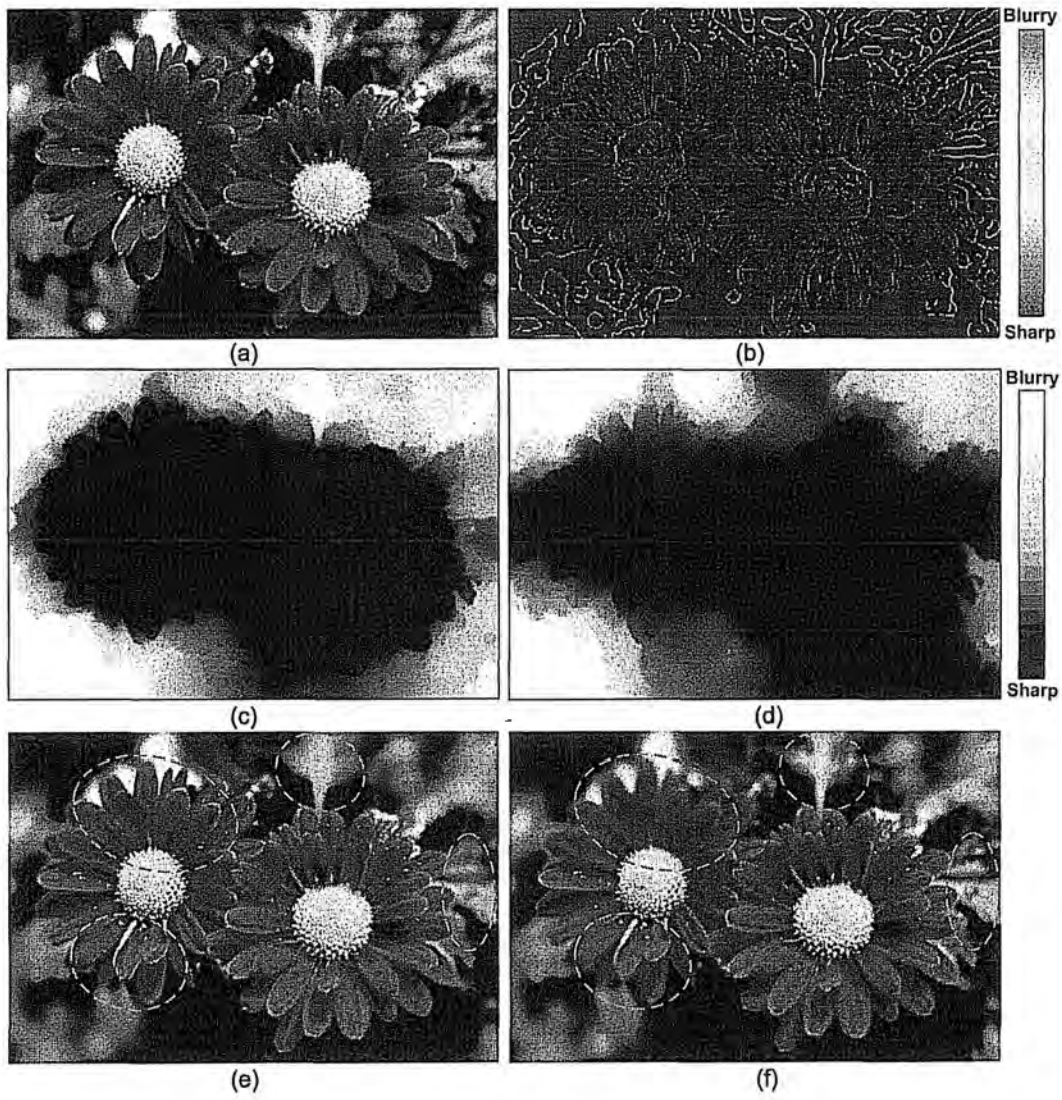

Figure 4.4: *Focus map estimation. (a) Input image. (b) Results of blurriness measurement (w) on edges. Blurriness increases gradually from blue to red. No edge is detected in the crimson regions. (c) and (d) show the focus map results of ours and Bae et al.'s respectively. Defocus increases gradually from black to white. Note that focus map here has been normalized for display purpose. (e) and (f) show the defocus magnification results of ours and Bae et al.'s respectively. The dashed ellipses outline the obvious errors occurred in Bae et al.'s result.*

scale Gaussian filter is employed to detect the edges in the image. Our study shows that [1, 3] is a reasonable range for setting σ_d. Then the edge blurriness w as well as the other parameters can be calculated directly based on (4.3)-(4.6). Next, to remove the outliers that occur in edge detection and parameter estimation, cross-bilateral filtering [Eisemann and Durand, 2004; Petschnigg et al., 2004; Paris and Durand, 2006] is conducted to refine the obtained edge blurriness results. Secondly, considering that

neighboring pixels with similar colors can be reasonably assumed having similar blurriness, we employ an additional blurriness propagation step for the non-edge pixels whose blurriness cannot be estimated by the first step. More specifically, the blurriness information at edge pixels is propagated to their neighboring non-edge pixels based on the similarities of intensity and position. According to the work on image colorization [Levin et al., 2004], such propagation can be formulated as the minimization of a quadratic cost function whose optimization can be solved efficiently within a linear system.

A similar method to ours is [Bae and Durand, 2007] which adopts a multi-scale edge detector and estimates the blurriness using a brute-force strategy. In detail, the degree of blurriness there was determined by approximately fitting the second derivative Gaussian filter response with a number of predefined candidates to a window around the edge pixel and along the gradient direction. By contrast, our proposed method is simpler and has lower computational complexity since all edge parameters are derived in closed form. Besides, a parameter x_0 is used to represent the edge position accurately in sub-pixel level. As stated in Sec 4.3.2, this representation is particularly advantageous when the actual edge center lies somewhere between two grid points.

Experiments were conducted to test the proposed method. The results in Figure 4.4(b) prove that the edge blurriness can be measured with good accuracy using our method. Compared to Bae et al.'s result in (d), our focus map result in (c) has less outliers and is more faithful to the perceived truth. Moreover, the proposed method is more efficient and only took about 31 seconds while Bae et al. [Bae and Durand, 2007] needs about one minute. Next, similar to [Bae and Durand, 2007], in order to further evaluate the accuracy of the focus map, we employed the lens blur of *Photoshop* to increase the defocus effect of (a) by inputing (c) and (d) as the alpha channels respectively. As expected, due to the influence of the outliers in (d), Bae et al.'s result in (f) is not satisfactory, where some focused regions are destroyed and some defocus regions are not blurred adequately. In contrast, our result in (e) is more visually realistic. This can also be concluded from the comparisons in Figure 4.5 and Figure 4.6, where the two input images are adopted in [Bae and Durand, 2007]. Noted that Figure 4.6 is a testing on a narrow aperture (f/8) image (a). Apparently, our defocusing result in (d) is better and visually closer to the ground truth wide aperture (f/4) image

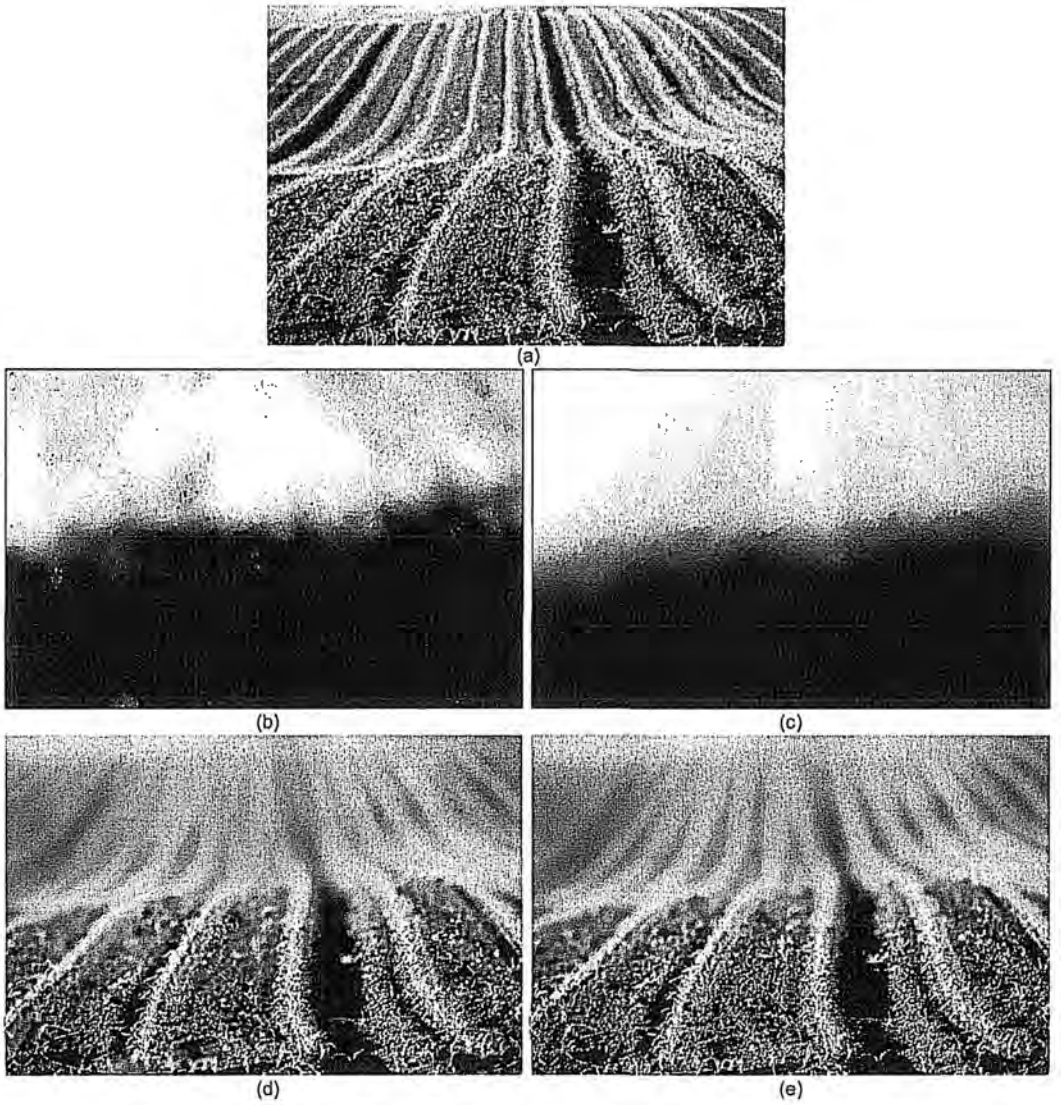

Figure 4.5: *Comparison on focus map estimation and defocus magnification. (a) Input narrow aperture image. (b) and (d) are Bae et al.'s presented results. (c) and (e) are our results. The dashed ellipses outline the obvious errors occurred in (d).*

(b) than Bae et al.'s result in (c).

4.5 Image Refocusing by Blind Deconvolution

Compared to defocusing, refocusing is more challenging. As aforementioned, image refocusing can be regarded a SBD problem whose goal is to recover the sharp image I_F as well as the PSF simultaneously from a blurry input I_D. However, the SBD is ill-posed since there are different pairs of images and PSFs that can output the same blurry

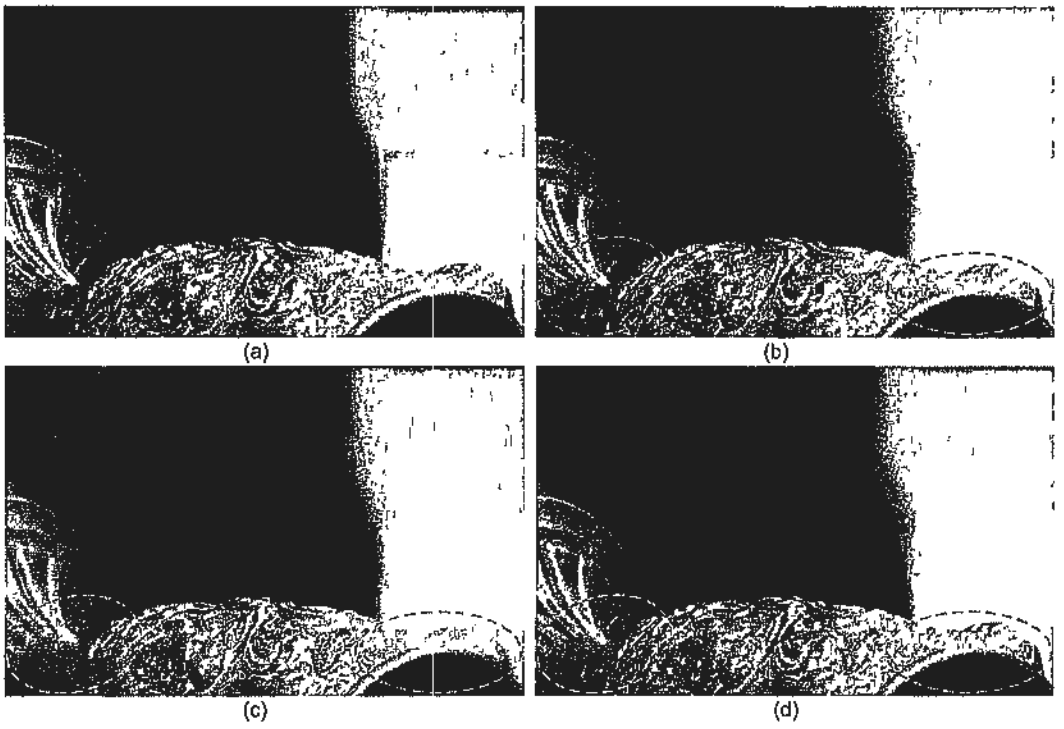

Figure 4.6: *Comparison on defocus magnification (a) Input narrow aperture (f/8) image (b) Ground truth image taken with wide aperture (f/4) (c) Bae et al 's presented result (d) Our result The dashed ellipses outline the obvious errors occurred in (c)*

image. In the past years, a variety of methods [Chan and Wong, 1998, Joshi et al 2008; Fergus et al 2006; Joshi et al , 2009, Shan et al , 2008a, Jia 2007] have been presented to tackle this challenging problem Most methods employ a simultaneous *maximum a posteriori* (MAP) estimator to infer the latent sharp image and PSF in an iterative manner As analysed in [Levin et al , 2009], such MAP estimator may not approach the desired global optimum since it favors the no-blur explanation That is, the PSF is delta kernel and the latent image is the same with observed blurry one. Besides, proper user intervention is often required at the initialization stage and poor initialization may result in undesired local convergence Although some efforts such as [Joshi et al 2008, Jia, 2007] were made to seek PSF from edges, the edge sharpness cue is not utilized adequately. In this chapter, we present a novel refocusing method that takes full advantage of the edge sharpness cue First, it is utilized for PSF estimation Then, an edge sharpness prior is developed to constrain the PSF not to blur the edges and enforce the refocusing image to agree with the precalculated sharpened

image in the vicinity of edges. Next, the proposed SBD method will be presented by assuming the PSF is spatially invariant for the sake of simplicity. Figure 4.7 shows an illustrative example to explain the proposed SBD process.

4.5.1 Expectations for the Refocused Image

Let I_F be the refocused image of a blurry image I_D. I_F is expected to satisfy two conditions. First, the edges should become sharpened in I_F. Second, the locally smooth regions in I_D should remain almost unchanged in I_F. By means of the parametric edge model introduced in Sec 4.3.2, we can formulate the first condition explicitly by ensuring a small width parameter w in (4.2) for each refocused edge. In detail, for a blurry input, we first reconstruct its predicted image I_p with sharp edges by decreasing the width parameter w like setting $w' = w/10$ as exemplified in Figure 4.3(c). Meanwhile, a binary edge mask M_e can also be determined, where white denotes the edge regions which comprise all edge pixels and their adjacent neighboring pixels and black denotes the non-edge regions. As the example shown in Figure 4.7, (c) and (d) show the corresponding edge mask and predicted image of (a) respectively. Apparently, the detected edges in (d) have been sharpened significantly compared to (a).

Second, locally smooth regions in I_D and I_F should be similar. Similar to [Shan et al., 2008a], the locally smooth region can be determined as follow. For each pixel in I_D, a window centering at it with size similar to that of the PSF is defined. If the standard deviation of pixels in this window is smaller than a threshold, this pixel will be regarded as locally smooth. As shown in Figure 4.7(b), a smooth mask M_s is obtained, where white denotes the locally smooth regions and black denotes the non-smooth regions.

4.5.2 Estimation of PSF

As shown in (4.7), the PSF h can be estimated in a MAP framework by taking advantage of the predicted sharp image I_p like Figure 4.7(d).

$$h^* = \arg\max_h \; p(h|I_D, I_p, M_e) = \arg\max_h \; p(I_D|h, I_p, M_e)p(h). \tag{4.7}$$

The likelihood term $p(I_D|h, I_p, M_e)$ can be defined based on the image formation model stated in (4.1).

$$p(I_D|h, I_p, M_e) \propto \exp\left(-\alpha_h M_e \bullet \|h \otimes I_p - I_D\|_2^2\right) \tag{4.8}$$

where $\bullet$ denotes element-wise multiplication operation. α_h acts as a weighting factor which is dependent on the noise level of the likelihood. Based on (4.7), an energy term can be defined as:

$$E_h(h) = \alpha_h M_e \bullet \|h \otimes I_p - I_D\|_2^2 + \|h\|_1 . \tag{4.9}$$

The PSF prior $p(h)$ is defined using a general l_1 norm sparse prior with non-negativity constraint. To obtain the solution efficiently, the minimization of (4.9) is recasted as an equivalent basis pursuit denoising (BPDN) problem:

$$\underset{\mathbf{x}}{\text{minimize}} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 < \eta, \tag{4.10}$$

where convolution is replaced with matrix multiplication. $\mathbf{x}$ is the vector form of h, the matrix $\mathbf{A}$ and vector $\mathbf{b}$ are derived from I_p and I_D with the guidance of M_e. Benefiting from [van den Berg and Friedlander, 2008] where a fast root-finding solver [van den Berg and Friedlander, 2007] for BPDN is presented, the latent PSF can be estimated efficiently from (4.10). Note that the resulting PSF will be rectified to ensure all elements are non-negative and the sum is one. The threshold value η should be chosen relative to the noise level and we have found empirically that [1, 15] is a practical range to produce good results.

4.5.3 Recovery of Focused Sharp Image

After h is determined, the recovery of I_F becomes a non-blind deconvolution problem as:

$$I_F^* = \arg\max_{I_F} \; p(I_F|I_D, h) = \arg\max_{I_F} \; p(I_D|I_F, h)p(I_F). \tag{4.11}$$

Similar to (4.8), the likelihood term $p(I_D|I_F, h)$ is defined based on the image formation model stated in (4.1) by assuming that I_D differs from the convolution of I_F with the PSF h by a zero mean Gaussian noise of variance $\frac{1}{2\alpha_n}$. Hence,

$$p(I_D|I_F, h) \propto \exp\left(-\alpha_n \|h \otimes I_F - I_D\|_2^2\right). \tag{4.12}$$

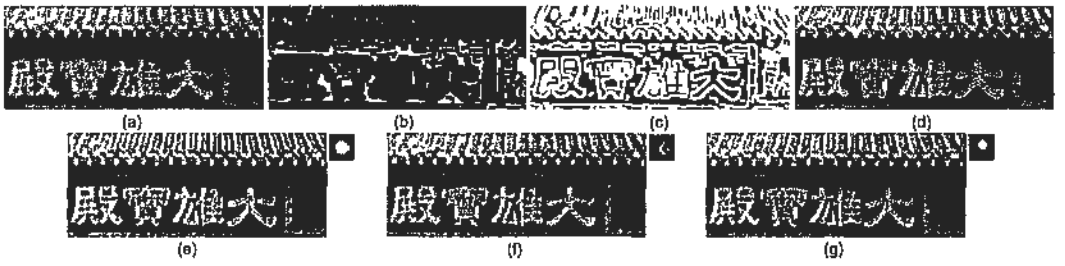

Figure 4.7: *Illustration of the proposed SBD (a) Defocus image cropped from Figure 1.1d(a). (b) Smooth mask M_s (setting the threshold to 6). (c) Edge region mask M_e. (d) Predicted image I_p obtained by sharpening the edges (decreasing w) in M_e. (e) Our results: refocused image and PSF (f) Fergus et al.'s results. (g) Shan et al.'s results.*

To impose an effective regularization, $p(I_F)$ is defined by combining three different priors as:

$$p(I_F) = p_g(I_F)p_e(I_F)p_s(I_F), \tag{4.13}$$

where $p_g(I_F)$ is a global prior, and $p_e(I_F)$ and $p_s(I_F)$ are local priors introduced based on the aforementioned expectations described in Section 4.5.1. The global prior $p_g(I_F)$ is defined by using the total variation regularizer as shown in (4.14).

$$p_g(I_F) \propto \exp\left(-\alpha_g \sum_i \|(t_i \otimes I_F)\|_2\right), \tag{4.14}$$

where t_i can be simply defined by the horizontal and vertical first order derivative filters: $t_1 = [1 \ -1]$ and $t_2 = [1 \ -1]^T$.

The sharp prior $p_e(I_F)$ is introduced based on the fact that the edge regions of the latent focused image I_F is expected to have similar sharpness with that of the predicted I_p. As shown in (4.15), the first order derivatives are utilized to measure the difference.

$$p_e(I_F) \propto \exp\left(-\alpha_e \sum_i M_e \bullet \|(t_i \otimes I_F - t_i \otimes I_p)\|_2^2\right). \tag{4.15}$$

The smooth prior $p_s(I_F)$ is introduced for suppressing the ring artifacts as in [Shan et al., 2008a]. As shown in (4.16), $p_s(I_F)$ is defined based on the expectation that the smooth regions of the defocused image I_D and the latent focused image I_F share similar first order derivatives.

$$p_s(I_F) \propto \exp\left(-\alpha_s \sum_i M_s \bullet \|(t_i \otimes I_F - t_i \otimes I_D)\|_2^2\right). \tag{4.16}$$

The maximization problem in (4.11) can be recasted as the minimization of an energy term defined based on (4.11)-(4.16).

$$E_F(I_F) = \alpha_n \|h \otimes I_F - I_D\|_2^2 + \alpha_g \sum_i \|(t_i \otimes I_F)\|_2 + \alpha_s \sum_i M_s \bullet \|(t_i \otimes I_F - t_i \otimes I_D)\|_2^2$$
$$+ \alpha_e \sum_i M_e \bullet \|(t_i \otimes I_F - t_i \otimes I_p)\|_2^2, \tag{4.17}$$

where the term 3 and term 4 on the right-hand side are for suppressing ringing artifacts and ensuring image sharpness respectively. Similar to the problem in Section 3.3.2, direct minimization of E_F is intractable since E_F is non-quadratic to the unknown I_F. Similarly, the variable-splitting and penalty scheme is adopted to tackle this optimization problem. As shown in (4.18), two variables ξ_1 and ξ_2 are introduced to replace $t_1 \otimes I_F$ and $t_2 \otimes I_F$ respectively. The discrepancy between ξ_i and $t_i \otimes I_F$ is penalized in a quadratic manner.

$$E_F(I_F) = \alpha_n \|h \otimes I_F - I_D\|_2^2 + \alpha_s \sum_i M_s \bullet \|(\xi_i - t_i \otimes I_D)\|_2^2 + \alpha_e \sum_i M_e \bullet \|(\xi_i - t_i \otimes I_p)\|_2^2$$
$$+ \alpha_g \sum_i \|\xi_i\|_2 + \beta \sum_i \|(\xi_i - t_i \otimes I_F)\|_2^2. \tag{4.18}$$

Iterative scheme is employed to update the unknown I_F and ξ_i alternatively with an increasing penalty parameter β. The solution of minimizing (4.18) will converge to that of minimizing (4.17) when β becomes large enough. At each iteration, when I_F is fixed, ξ_i $(i = 1, 2)$ is updated by minimizing $E_{F\xi}(\xi_i)$ separated from $E_F(I_F)$.

$$E_{F\xi}(\xi_i) = \alpha_s \sum_i M_s \bullet \|(\xi_i - t_i \otimes I_D)\|_2^2 + \beta \sum_i \|(\xi_i - t_i \otimes I_F)\|_2^2 + \alpha_g \sum_i \|\xi_i\|_2$$
$$+ \alpha_e \sum_i M_e \bullet \|(\xi_i - t_i \otimes I_p)\|_2^2. \tag{4.19}$$

Since $E_{F\xi}(\xi_i)$ is differential to ξ_i, a closed-form solution can be obtained. It is worth noting that ξ_i is updated in a pixel by pixel manner due to the influence of M_s and M_e. Similarly, when ξ_i is fixed, I_F is updated by minimizing $E_{FF}(I_F)$.

$$E_{FF}(I_F) = \alpha_n \|h \otimes I_F - I_D\|_2^2 + \beta \sum_i \|(\xi_i - t_i \otimes I_F)\|_2^2, \tag{4.20}$$

where $E_{FF}(I_F)$ is quadratic to I_F and its minimization is a typical least square problem

which also has a closed-form solution. To avoid the computational complexity caused by convolution, it would be better to tackle the above least square problem in the Fourier transform domain. All parameters involved in (4.18) act as weighting factors and thus are used to balance the contributions of the corresponding terms. For example, the larger is α_e, the closer is the output image to the predicted sharp image I_p. For real images, we have empirically found that α_e varies from 1 to 5. α_n is tuned between 500 and 2000. β is set equal to 1 at the beginning and then increased by 2 times after each iteration. α_g and α_s are normally fixed at 1 and 40 respectively.

4.5.4 Discussion on the SBD Results

Figure 4.7 shows a close-up of the refocused result in Figure 4.14. Our estimated sharp image I_F and PSF h are shown in Figure 4.7(e). Note that the errors that occur at the sharpened edges in the predicted image (d) are due to the influence of noise and nearby edges. The proposed method is robust to such outliers since the estimations of h and I_F are handled in two separated MAP frameworks where the predicted sharp edges are only one constraint for ensuring sharpness. The influence of these outliers can be eliminated by the other constraints such as the smoothness term. Results in Figure 4.7 show that the proposed method yields the finest details and the least artifacts in comparison to the other two algorithms. Also, the resulting PSF is closer to a typical out-of-focus blur kernel. Note that all SBD methods were tested on the same defocus layer of an image for comparison as shown in Figure 4.7 and Figure 4.11(d)-(g).

Experiments using synthesized images were also conducted to evaluate the proposed SBD method. As shown in Figure 4.8, the synthesized blurry image (b) was obtained by adding Gaussian noise to the convolution result of the original sharp image (a) and Gaussian PSF ($\sigma = 1.5$). The SBD results obtained using different methods are shown in (c), (d) and (e). Apparently, the PSF and recovered image obtained by the proposed method are closer to the ground truth compared to the results obtained by the other algorithms. To make a quantitative comparison, we use the SSD (sum of squared differences) criterion to measure the accuracy of the recovered sharp image and the estimated PSF. The comparison results on SSD are shown in Figure 4.9 for σ equal to 0.8, 1.5 and 2.5. Results obtained by the proposed method have the smallest SSD. Besides, we also adopted a recently developed measure called SSIM (structural

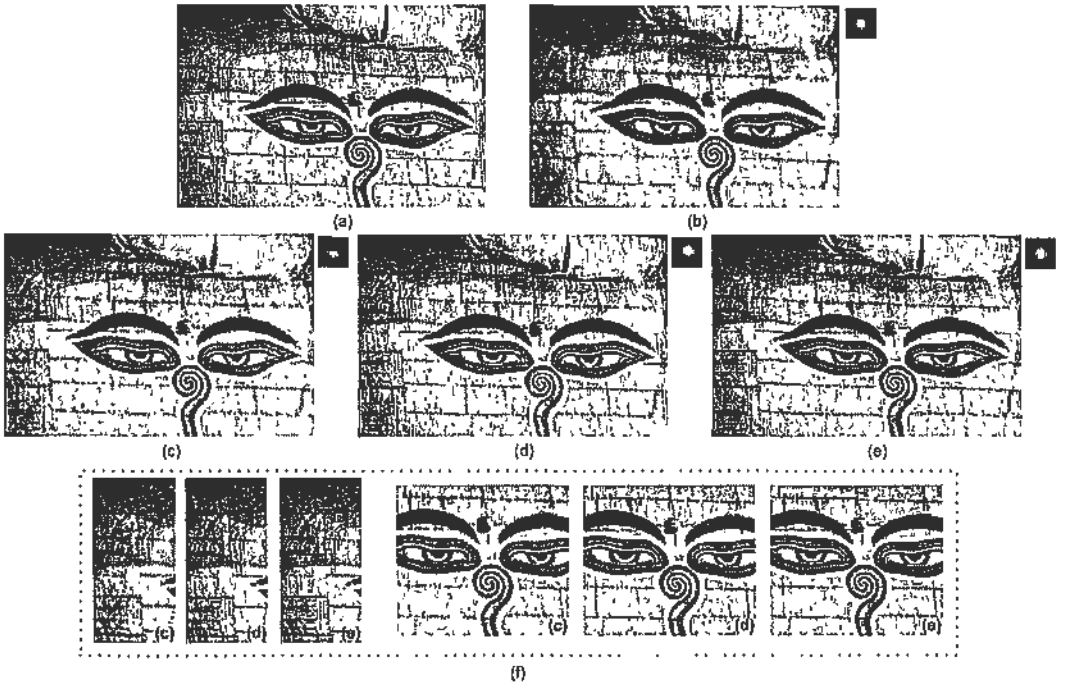

Figure 4.8: *Testing on synthesized image. (a) Original image. (b) Synthesized image blurred with Gaussian PSF ($\sigma = 1.5$) shown in the top right. (c) Fergus et al.'s results. (d) Shan et al.'s results. (e) Our results. (f) Close-up visual comparison. (The differences are better seen by zooming on a computer screen.)*

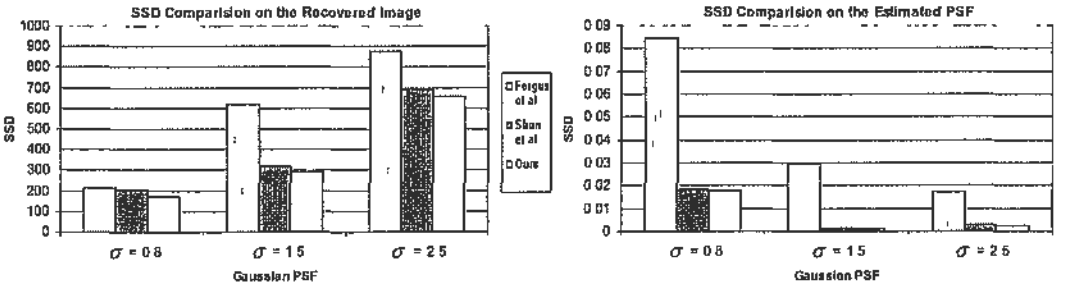

Figure 4.9: *Quantitative comparisons with SSD on the recovered images and estimated PSFs obtained with different SBD methods. Note that the small shifts of PSF center and refocused image occurred in [Fergus et al., 2006] and [Shan et al., 2008a] have been corrected for fair quantitative comparison.*

similarity) [Wang et al., 2004] to assess the similarity of the recovered sharp image and the original one. As shown in Figure 4.10, our deblurred images have the largest SSIM values. The above comparisons show that the proposed SBD method works better than the other two state-of-the-art SBD algorithms both perceptually and quantitatively. In the above experiments, the results of the methods of Fergus et al. [Fergus et al., 2006]

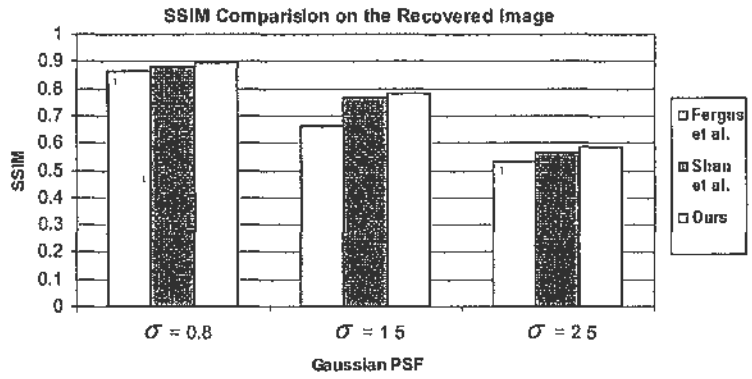

Figure 4.10: *Quantitative comparisons with SSIM on the recovered images obtained with different SBD methods.*

and Shan et al. [Shan et al., 2008a] were produced by using the original implementations with parameters adjusted based on the authors' instructions.

Moreover, since it is unnecessary to update the PSF and the latent sharp image iteratively, the proposed method has lower computational complexity. For example, to deblur a 480 × 320 image, the Matlab implementation of Fergus et al. [Fergus et al., 2006] normally runs more than 20 minutes on a PC with an Intel Core2Duo 3.0GHz CPU. The executable code from Shan et al. [Shan et al., 2008a] implemented using C runs about 2.5 minutes. The proposed algorithm which was implemented using Matlab requires comparable time as Shan et al.'s method, there is still much room to improve its efficiency by optimized C++ or GPU implementation.

4.6 Experiments and Discussions

In this section, more experiments were carried out to show that the proposed system can generate different styles of images by refocusing and defocusing. One experiment was conducted to produce results as shown in Figure 4.11. The input image (a) focused on the center of the two bottom numbers (1 and 0) was taken by using typical macro photography with shallow DOF. First, its focus map is produced as shown in Figure 4.12(a). Note that the pixels which are in the same grid affect each other significantly because of their similar colors. Since the black regions between the grids do not have much texture, their blurriness relies on that of the neighboring grid boundaries. As shown in Figure 4.12(b), the blurriness of the pixels at the dashed line drawn in (a) decreases gradually from the top to the bottom, which coincides with the focus setting

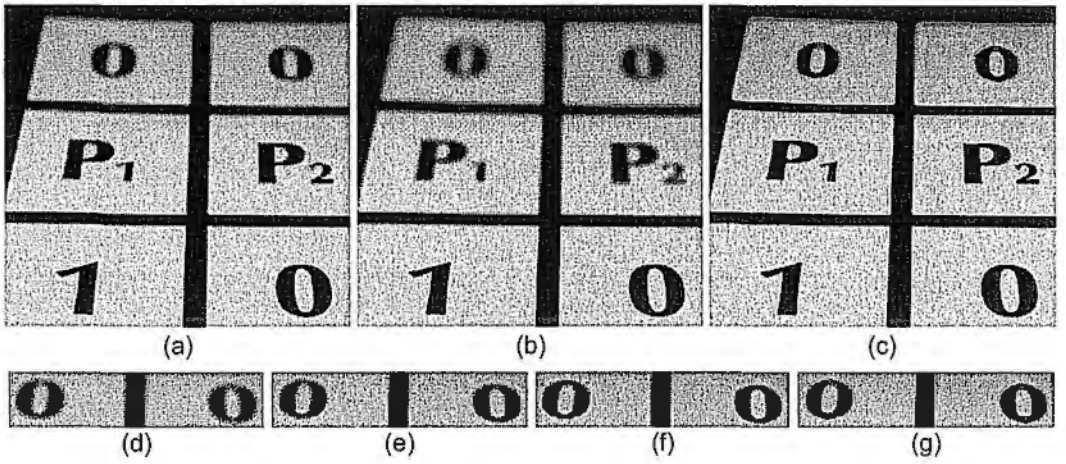

Figure 4.11: *(a) Input image. (b) Synthesized image with shallower DOF. (c) Synthesized all-focused image. (d) The defocus part cropped from (a); (e) Fergus et al.'s result. (f) Shan et al.'s result. (g) Our refocused result.*

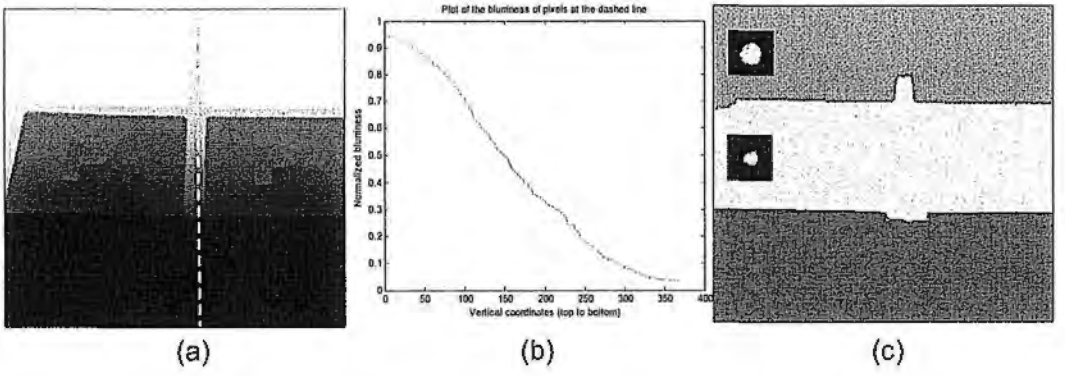

Figure 4.12: *(a) Estimated focus map. (b) Blurriness of the pixels at the dashed line of (a). (c) Segmentation result based on (a).*

of the captured image. In this example, we segmented the focus map into three layers as shown in Figure 4.12(c) by setting two thresholds as $w_{th} = 0.6$ and $w'_{th} = 2 * w_{th}$. Note that the focus threshold w_{th} is related to the image texture and can be set empirically from 0.4 to 1.2. More layers can be obtained with more thresholds. Pixels on the same segment can be reasonably assumed sharing the same blurring PSF. The bottom layer which has the smallest blurriness is regarded as focused. The proposed SBD will be implemented individually on each defocus layer for refocusing and the synthesized all-focused result is shown in Figure 4.11(c). The recovered PSFs are shown in Figure 4.12(c). Apparently the PSF of the upper layer which is more defocused has larger

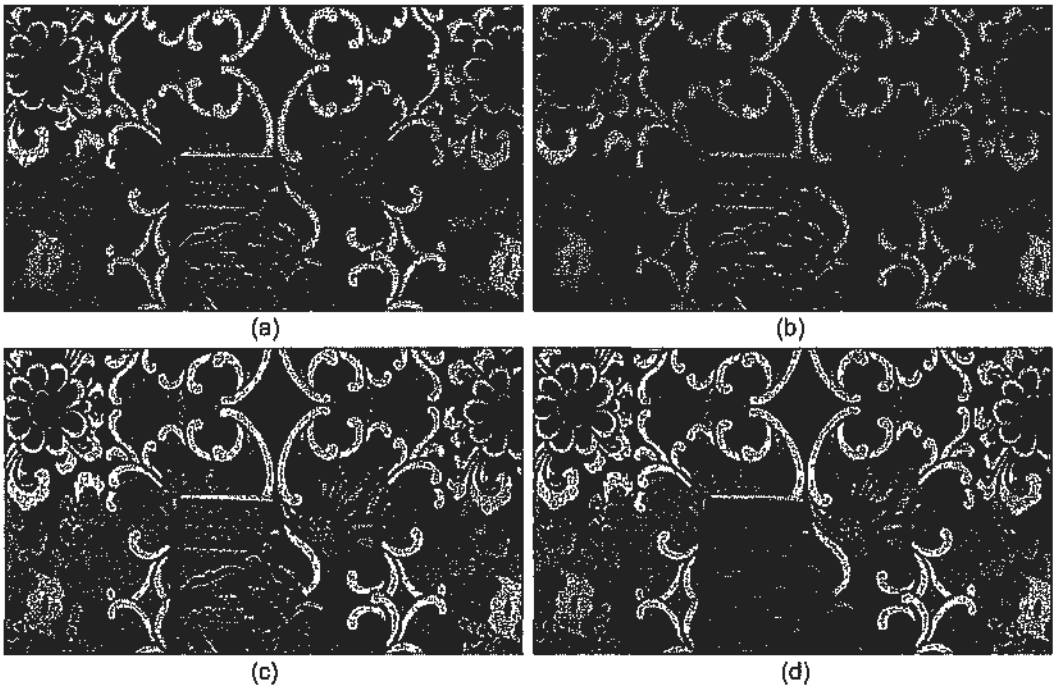

Figure 4.13: *(a) Input image. (b) Synthesized image with shallower DOF (c) Synthesized all-focused image. (d) Synthesized image focusing on the background.*

scale than that of the middle layer. To simulate the real lens defocusing effects, Figure 4.11(b) is produced by applying different Gaussian blurring on the upper and middle layers, where the ratio of the two blur scale is proportional to that of their detected blurriness.

Another example is shown in Figure 4.1, where the girl and the other background in the input image (a) can be reasonably assumed on the same defocus layer because of their similar depth. The binary mask (e) is generated with focus threshold $w_{th} = 1.1$ and divides the input image into two focus layers. Two additional examples are shown in Figure 4.13 and Figure 4.14. It is noted that the synthesis of layer boundaries such as the gray regions in Figure 4.1(e) is conducted smoothly by alpha blending to avoid generating seam artifacts. Please also visit `http://www.ee.cuhk.edu.hk/~zhangwei/FocusEditing.html` to see a demonstration video, including all above results.

The proposed method has some limitations. First of all, as aforementioned, the proposed method cannot generate desirable focus map for image that contains objects naturally blurry. Besides, since the occlusion problem is not addressed in this work, the proposed method can hardly handle the image that is composed of many focus layers

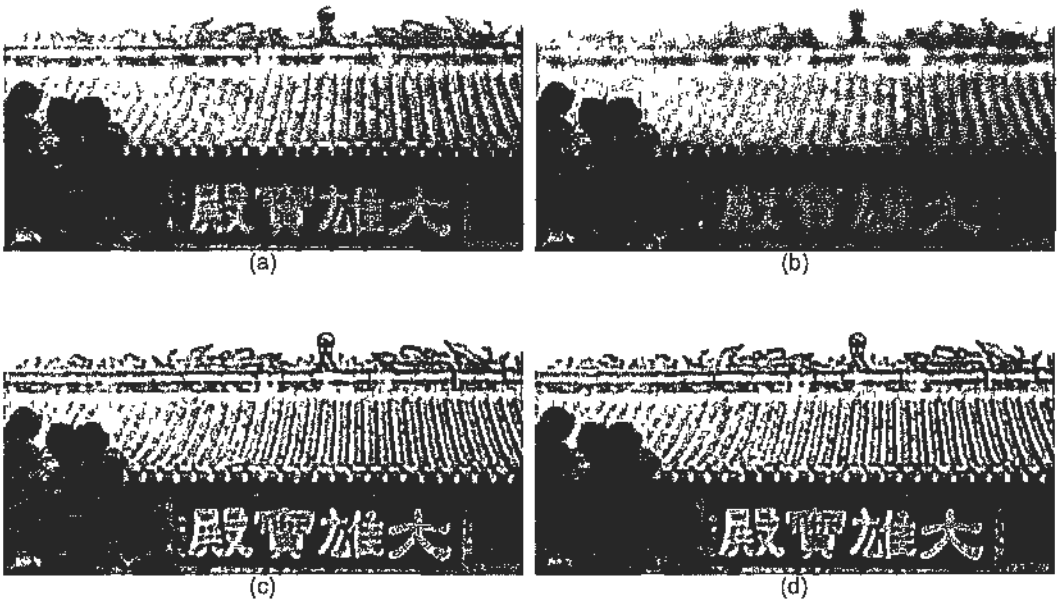

Figure 4.14: *(a) Input image. (b) Synthesized image with shallower DOF (c) Synthesized all-focused image. (d) Synthesized image focusing on the building*

with large discontinuities. Normally, we prefer segmenting the image into two or three layers. The unfocused objects similar in depth can be reasonably assumed at the same layer. This is because: first, the quality of refocused images may degrade especially when the image is segmented too much and the layer boundaries frequently appear Second, for one image, the more layers it is divided into, the less information is left at each layer and thus refocusing will become harder due to the limited amount of data available. However, one possible solution is to put in human intervention as [Yan et al., 2009; Bando and Nishita, 2007] to provide some guidance to the method especially in the aforementioned tough cases.

4.7 Summary

In this chapter, we have presented a system to handle the tasks of focus map estimation, image refocusing and defocusing. First, by means of a parametric edge model, we propose an efficient and effective focus map estimation method. Second, the challenging refocusing problem is tackled in a SBD framework which yielded visually pleasant results with the aid of the novel image sharp prior. Besides, the proposed SBD is free

of user initialization and has low computational complexity. A wide variety of images have been tested to validate the proposed algorithm.

Chapter 5

Gradient-Directed Composition of Multi-Exposure Images

5.1 Introduction

Radiance of the real world spans several orders of magnitude and its dynamic range dramatically exceeds the capability of the current electronic imaging devices. As a result, there often exist some undesirable over- or under-exposed regions in a photo when the dynamic range of the latent scene is too vast to be reproduced with a conventional camera at a single aperture and shutter speed. There exist some hardware solutions such as [Brajovic and Kanade, 1996; Nayar and Branzoi, 2003; Aggarwal and Ahuja, 2004; Tumblin et al., 2005] which aimed at extending the dynamic range of conventional cameras by including additional optical elements or devices. However, in contrast to conventional camera, high dynamic range (HDR) camera is still unavailable to consumer users due to its slow exposure speed, expensive price and high requirements on hardware. Since each exposure can be designed to capture a certain dynamic range, it is possible to capture the full dynamic range of the latent scene and create a HDR image with a conventional camera by combining a stack of images with different exposure times. Because of the popularity of consumer cameras such as single-lens reflex (SLR) cameras and point-and-shoot cameras, this kind of approach called multi-exposure technique has a greater potential to impact everyday photography.

5.1.1 Related Work

The multi-exposure technique should be discussed in two cases. First, if the stack is captured in a static scene, it is a static HDR problem whose goal is to recover the full dynamic range and make all present details visible in one image. Second, if there is any object movement in the latent scene while the exposures are being captured, the

moving objects will appear in different locations in the captured image. HDR imaging in this case is more challenging because direct combining all exposures suffering from such inconsistencies will surely cause ghosting artifacts to be visible in the resulting HDR image.

The most popular HDR tools in the current graphics software belong to this static HDR category and normally consist of two steps: (i) calibrate the camera response function (CRF) [Debevec and Malik, 1997; Grossberg and Nayar, 2003] and recover the latent radiance map (HDR image); (ii) apply tone mapping to make the HDR image displayable on the commonly used low dynamic range (LDR) monitors [Durand and Dorsey, 2002; Reinhard et al., 2002; Fattal et al., 2002; Li et al., 2005]. These tools did not consider the object movement and thus share a serious limitation that the target scene is required to be completely still throughout the image capture. As shown in Figure 5.5, any object movement in the exposure sequence can cause ghosting artifacts in the resulting image. This drawback severely affected their application in practice, since for most scenarios, it is hard to guarantee all objects involved stay stationary from one capture to the next. For instance, there often exist crowds of people moving around in tourist resorts. There are windblown trees in nature scenes.

Lots of efforts have been made to solve the ghosting problem in dynamic scene recently. The existing methods were proposed in a similar manner. They first detect the motion regions, and then produce a ghost-free HDR result by remove the contributions of these regions in the composite radiance map. For example, many different kinds of techniques such as optical flow [Kang et al., 2003], variance measurement [Reinhard et al., 2005], error map detection [Grosch, 2006], entropy calculation [Jacobs et al., 2008] and pixel's order relation detection [Sidibe et al., 2009], have been adopted to find regions where ghosting artifacts may occur due to object motion. Besides, Gallo et al. [Gallo et al., 2009] and Eden et al. [Eden et al., 2006] proposed to composite the desirable radiance with the guidance of a reference image preselected automatically or manually. Some statistical tools such as kernel density estimator were employed in [Khan et al., 2006; Pedone and Heikkilä, 2008] to iteratively determine the probability that a pixel belongs to the background.

However, all above methods were presented in the radiance domain fully or partially. Hence, they share two limitations at least. First, the performance highly relies on

the success of the radiometric calibration of camera which is sensitive to image noise, lighting change and misalignment error. Second, they normally have complex working pipelines and require tone mapping for HDR reproduction. The above problems make these kinds of methods computationally expensive and restrict their applications in practice.

5.1.2 This Work

In this chapter, we present a novel exposure composition approach that is able to bypass the typical HDR process and directly yield a tonemapped-like HDR image where all parts appear well-exposed by compositing multi-exposure images with the guidance of image quality assessment. Our algorithm shares the same spirit with the recent work [Goshtasby, 2005; Mertens et al., 2009; Shanmuganathan and Chaudhuri, 2009] for using image fusion to obtain better exposed image. But since all of them belong to the static catergory as the convention HDR work and assume no object movement in the scene, they can only deal with the images captured in static scenes and suffer from severe ghosting artifacts in dynamic scenes. Moreover, we address the multi-exposure image composition from the perspective of gradient, and develop a new quality assessment system to handle the composition in both static and dynamic scenes.

In addition, image gradient has been manipulated in several tasks such as tone mapping [Fattal et al., 2002], image editing [Pérez et al., 2003] and enhancement [Agrawal et al., 2005]. It is worth pointing out that [Fattal et al., 2002] and [Agrawal et al., 2005] differ from ours essentially. [Fattal et al., 2002] is a tone mapping method that seeks to compress the radiance map to a displayable range with a spatially varying gradient attenuation function. [Agrawal et al., 2005] aims at removing the artifacts existing in flash photography with a gradient projection scheme. Moreover, it was proposed for static scenes and thus cannot handle the dynamic scenes.

Specifically, the underlying idea of this work comes from the observations of gradient changes among differently exposed images. Firstly, gradient magnitude can imply pixel's exposure quality and will decrease gradually as the image is approaching over- or under-exposure. Consequently, it can be utilized as a measure on visibility to help preserve the details present in the exposure sequence. Secondly, it is also found that the gradient direction changes reveal object movement and thus can help account for

the ghosting problem in dynamic scenes. More detailed, if the content in some area changes among different exposures due to object movement, the gradient direction in that area will probably have significant changes as well. Consequently, exploiting the gradient direction changes leads to a consistency measure which can get rid of the influence of moving objects and preserve the desired consistent pixels in the composite image. By combining the consistency measure and visibility measure, the proposed method is still capable of compositing all exposures gracefully in dynamic scenes and producing a pleasant well-exposed image free of ghosting artifacts.

Generally speaking, there are two types of motion in a dynamic scene: (i) a moving object on a static background, e.g. moving people or cars; (ii) a moving background with dynamic objects, e.g. windblown trees or waves. Accordingly, we propose two gradient-based consistency measures to tackle the above two types of motion. One is named as accumulated consistency assessment (ACA), which which is particularly effective for removing all unwanted moving objects and producing a clean composite image. The other is named as reference view guided consistency assessment (RCA), which is particularly effective for dealing with background motion. The underlying idea of RCA is to composite all available dynamic range by taking one preselected image as a substrate. Hence, the proposed method is also able to produce a composite image with some moving object the user desired.

In summary, the proposed algorithm is designed to have the following properties: first, it is easy to use and has lower computational complexity since neither radiometric camera calibration nor tone mapping is required. Second, for dynamic scenes, the proposed approach can eliminate the ghosting artifacts automatically and efficiently without resorting to any explicit complex motion detection techniques like optical flow. Third, it allows for lighting changes and can be extended naturally to other tasks such as flash and no-flash photography.

5.2 Algorithm

5.2.1 Motivation and Overview

Since different exposures capture different dynamic range characteristics of the latent scene, taking multiple exposures and combining them together as (5.1) may create a

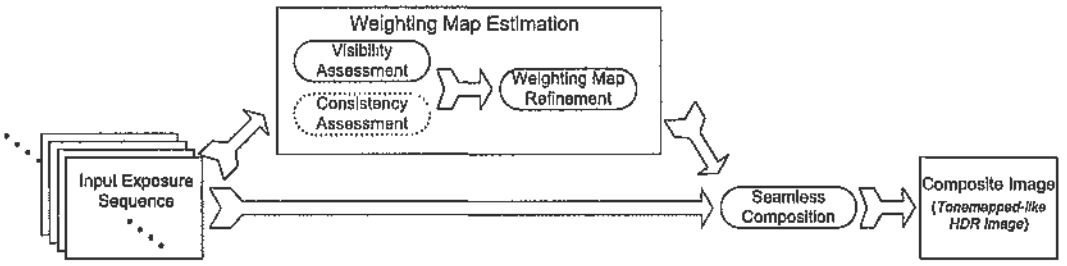

Figure 5.1: *The proposed framework. Please note that consistency assessment is unnecessary for static scenes.*

more informative image that captures all details of the scene.

$$I_C(x,y) = \sum_{i=1}^{K} W^i(x,y) I^i(x,y), \tag{5.1}$$

where K represents the number of the input exposures. $I^i(x,y)$ and $W^i(x,y)$ denote the intensity and weight of the pixel located at (x,y) in the i_{th} exposure respectively. I_C denotes the composite image to be generated. Compared to the typical HDR process, exposure composition is easier and much more efficient since neither radiometric camera calibration nor tone mapping is necessary. However, the composition performance relies on the weight term W and so it is crucial to develop an effective quality assessment system that can output the desired weights. In this chapter, we will show that the gradient information plays well in the quality assessment and makes it possible to handle the exposure composition in both static and dynamic scenes.

As illustrated in Figure 5.1, the proposed HDR process is quite simple and begins with a stack of differently exposed images. In this work, we assume all exposures are captured with the aid of a tripod or have been aligned by some registration techniques like [Ward, 2003; Brown and Lowe, 2003]. Then, the weighting map of each exposure is estimated by a gradient-based quality assessment system. For dynamic scenes, assessments on visibility and consistency are both required, while for static scenes only the former one is necessary. Besides, since every pixel is assessed independently without considering the spatial consistency within one image, some pixels may get outlier weight estimates due to the influence of image noise, inaccurate gradient detection and so on. Hence, a cross-bilateral filtering [Eisemann and Durand, 2004; Petschnigg et al.,

2004; Paris and Durand, 2006] based refinement is introduced to eliminate the outlier weights and ensure that adjacent pixels have similar weights if they share similar intensities. The standard deviations of the space and range Gaussians are normally set to 5 in the experiments. Given weighting maps, a tonemapped-like HDR image is produced eventually by compositing all exposures seamlessly with a multiresolution spline scheme [Burt and Adelson, 1983].

5.2.2 Gradient-based Image Quality Assessment

In this section, we will describe how to take advantage of the gradient information to generate weighting maps for static and dynamic scenes. Similar to Canny detection, we adopt the first derivatives of 2-D Gaussian filter $g(x, y; \sigma_d)$ in x direction and y direction to extract the gradient information in this work as follows.

$$I_x^i(x,y) = I^i(x,y) \otimes \frac{\partial}{\partial_x} g(x,y;\sigma_d), \tag{5.2}$$

$$I_y^i(x,y) = I^i(x,y) \otimes \frac{\partial}{\partial_y} g(x,y;\sigma_d), \tag{5.3}$$

where I_x^i and I_y^i are the partial derivatives of image I^i along x direction and y direction respectively. The standard deviation σ_d is set to 2 in the experiments. The gradient magnitude reflects the maximum change in pixel values while the angle points out the the direction corresponding to the maximum change. These two components are calculated in (5.4) and (5.5), respectively.

$$\nu^i(x,y) = \sqrt{\left|I_y^i(x,y)\right|^2 + \left|I_x^i(x,y)\right|^2}, \tag{5.4}$$

$$\theta^i(x,y) = \arctan \frac{I_y^i(x,y)}{I_x^i(x,y)}. \tag{5.5}$$

Visibility Assessment

As shown in Figure 5.2(a), some features that are visible in one exposure disappear in the others due to over- or under-exposure. Therefore, the basic goal of composition is to preserve all features present in the exposure sequence and make them visible in one image. Gradient is associated with image features and its magnitude is an indicator of pixel's exposure quality. As illustrated in Figure 5.2(b), gradient magnitude becomes larger when a pixel gets better exposed. It will decrease gradually as the pixel is

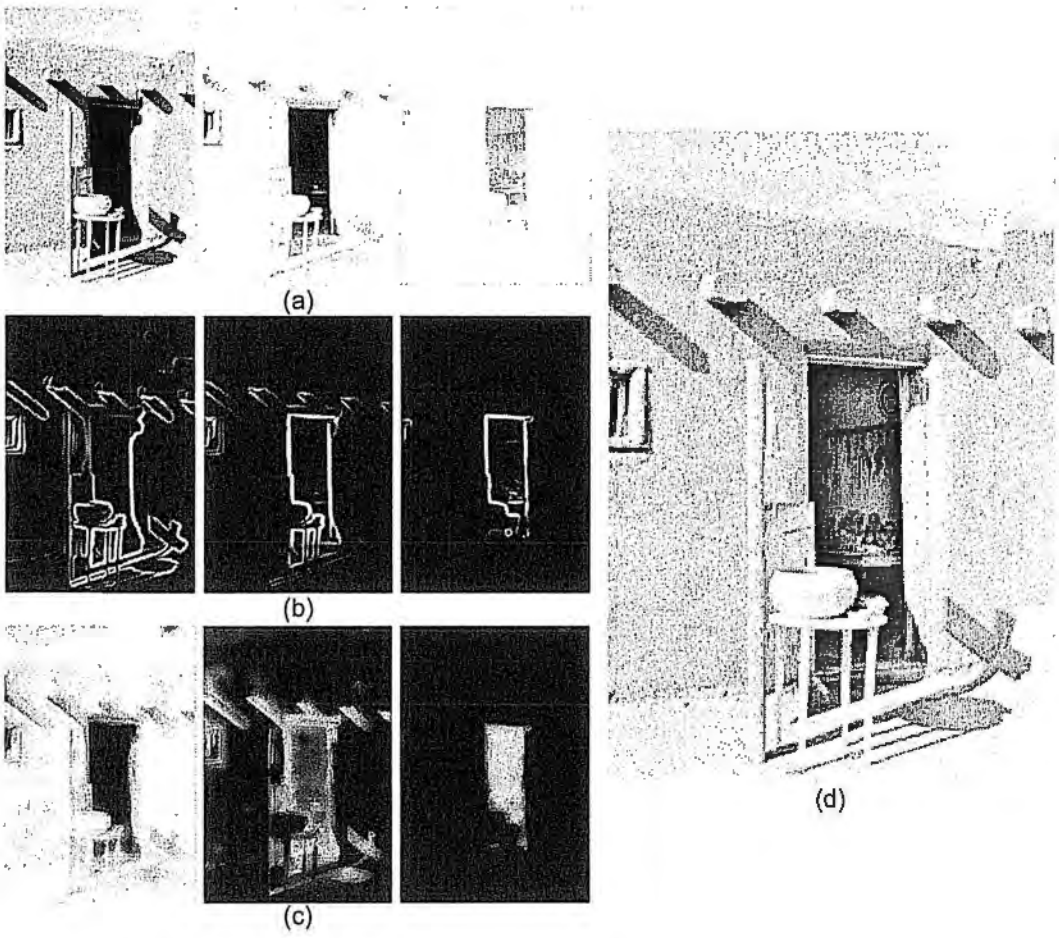

Figure 5.2: *Static example with visibility assessment. (a) Input three exposures. (b) Gradient magnitude maps. Note that each has been normalized to* [0, 1] *for display. (c) Weighting maps after refinement. (d) Composite image. Data courtesy of Shree K. Nayar.*

approaching over- or under-exposure. Therefore, a visibility measure is developed as (5.6) by exploiting the gradient magnitude information.

$$V^i(x,y) = \frac{\nu^i(x,y)}{\sum_{i=1}^{K} \nu^i(x,y) + \epsilon}, \tag{5.6}$$

where ϵ is a small value such as 10^{-25} to avoid singularity. In static scenes, the weights of (5.1) can be obtained by setting $W_V^i = V^i$. As shown in Figure 5.2, exposure composition guided by gradient magnitude in a static scene can produce a plausible result in which all visible details are preserved.

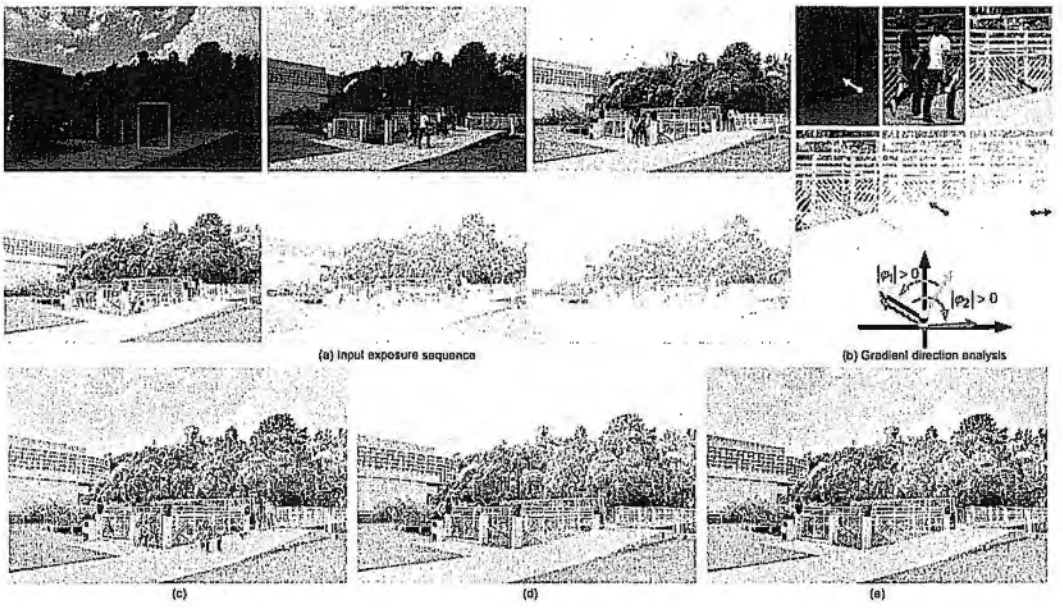

Figure 5.3: *Dynamic example. (a) Input six exposures. (b) Analysis of the gradient direction changes among differently exposed images. Please note that the arrow here is only used to indicate the gradient direction illustratively and its length is unrelated to the magnitude. (c) Composite result with only visibility assessment. (d) Composite result without exposure correction. (e) Our final composite result.*

Consistency Assessment

However, most scenes encountered in practice are non-static. It is hard to make all involved objects stay stationary while taking the exposure sequence. As shown in Figure 5.3(a), when we photography a public place, there often exist unwanted moving objects such as walking people. In this case, visibility assessment only cannot avoid compositing the inconsistent content appeared in the motion area and yields an unpleasant result ruined by ghosting artifacts as shown in Figure 5.3(c). Hence, it is necessary to seek an additional measure on consistency that can help remove the undesired moving objects and generate a ghost-free composite image.

Fortunately, we found gradient direction can serve in the consistency measure owing to its invariant property in different exposures as explained in Figure 5.3(b). In specific, it is observed that the gradient direction in the stationary region remains stable in different exposures, provided that these regions are neither under-exposed nor over-exposed. The inherent reason of this fact is that image gradients are mainly due to the local changes in 3-D geometric shape and reflectance. If the content changes due

to object movement, the gradient direction will vary accordingly (e.g. $|\varphi_1| > 0$ in Figure 5.3(b)). Therefore, we believe that the gradient direction information would be particularly effective to detect the inconsistency caused by motion. In this work, the measurement of gradient direction changes is accomplished in a window-based manner to make it more resistant to noise. Specifically, for each pixel located at (x, y) of the i_{th} image, its gradient direction change w.r.t that of the j_{th} image is calculated as follows.

$$d_{ij}(x,y) = \frac{\sum_{m=-r}^{r} |\theta^i(x+m, y+m) - \theta^j(x+m, y+m)|}{(2r+1)^2}, \tag{5.7}$$

where the size of window is $(2r+1) \times (2r+1)$ and r is normally set to 9. It is noted that $d_{ij}(x,y) = d_{ji}(x,y)$ and $d_{ij}(x,y) = 0$, when i and j are equal.

Accumulated Consistency Assessment (ACA) The first kind of consistency is developed based on the observation that many exposure sequences such as Figure 5.3(a) normally have one thing in common: the moving object is only a shot for one position and appears in a relative smaller number of images. This is because in most cases, the stationary parts of the scene that predominantly exist in the sequence are what the photographer is interested in. Consequently, a score S_A^i can be defined as (5.8) by accumulating the gradient direction changes of each exposure to reflect its consistency in the whole sequence.

$$S_A^i(x,y) = \sum_{j=1}^{K} exp(\frac{-d_{ij}(x,y)^2}{2\sigma_s^2}), \tag{5.8}$$

where σ_s is the standard deviation and fixed at 0.2 in the experiments. Apparently, a large score implies small gradient direction change and thus the content is more frequently captured in the sequence. (5.8) can favor the stationary parts of the scene under the assumption that the exposure sequence predominantly captures the stationary parts of the latent scene, which is prevalent in the previous work [Khan et al., 2006; Sidibe et al., 2009]. However, the direction changes of gradient may also be caused by over- or under-exposure (e.g. $|\varphi_2| > 0$ in Figure 5.3(b)). In this case, the score calculated based on $d_{ij}(x,y)$ is no longer desirable, since it may make the algorithm mistake the stationary visible objects for unwanted moving ones. Therefore, an additional term E^i which indicates the exposure quality of I^i, is introduced to jointly define the final

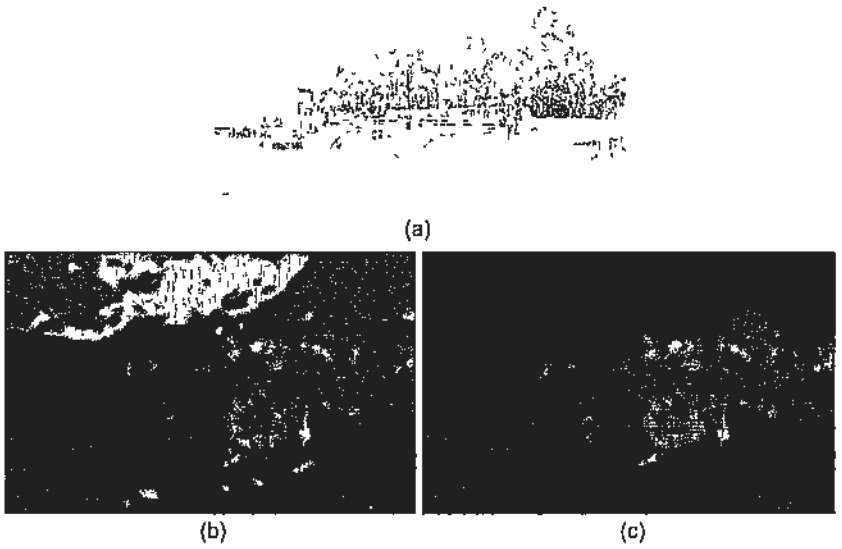

Figure 5.4· *Effect of exposure correction on weighting map estimation The left is the fifth input image of the sequence in Figure 5 3(a) The middle and right show its weighting maps before and after exposure correction respectively*

consistency measure C_A^i with S_A^i as follows.

$$C_A^i(x,y) = \frac{S_A^i(x,y) \times E^i(x,y)}{\sum_{i=1}^{K} S_A^i(x,y) \times E^i(x,y) + \epsilon}, \tag{5.9}$$

where

$$E^i(x,y) = \begin{cases} 1 & 1-\tau < I^i(x,y) < \tau \\ 0 & \text{otherwise.} \end{cases} \tag{5.10}$$

Note that E^i is used to remove the invalid scores estimated in the over- or under-exposed regions. τ defines the well-exposed range and is normally fixed at 0.9 in the experiments. The final weights in dynamic scenes are calculated by combining the visibility and consistency measures as:

$$W_A^i(x,y) = \frac{V^i(x,y) \times C_A^i(x,y)}{\sum_{i=1}^{K} V^i(x,y) \times C_A^i(x,y) + \epsilon}. \tag{5.11}$$

As shown in Figure 5 3(e), they give rise to a pleasant result where all visible details are preserved and no ghosting artifact is present.

Figure 5 1 illustrates the effect of exposure correction in the example of Figure

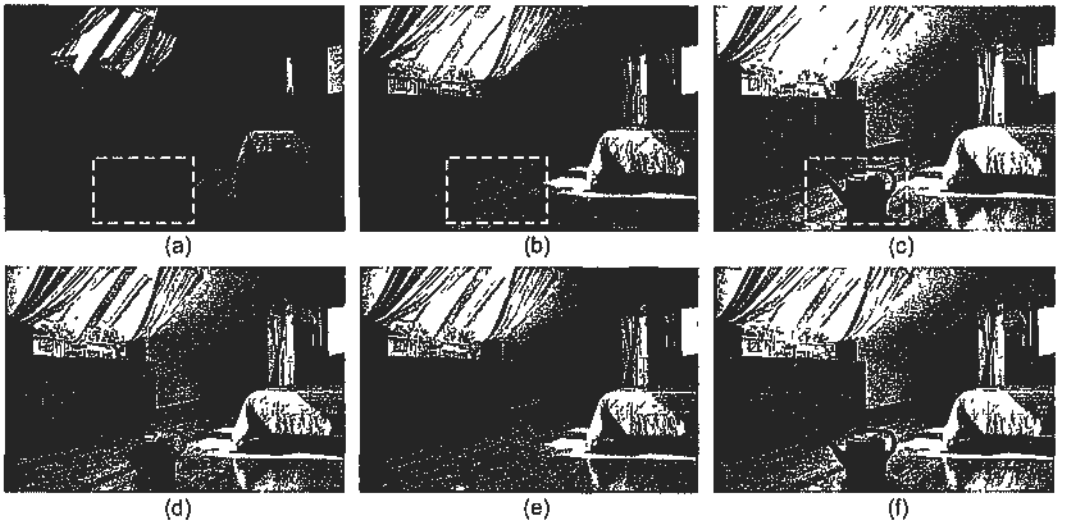

Figure 5.5: *Deghosting using ACA and RCA. Top row: input images with variable exposures, where the regions outlined by dashed rectangle are different with each other due to object movement. Bottom row: (d) shows the deghosting result using ACA. (e) and (f) are deghosting results obtained by taking image (b) and (c) as the reference view in RCA, respectively. Data courtesy of Mateusz Markowski.*

5.3. Taking the sky region for example, since pixels in this region are over-exposed in most exposures, the weights obtained without exposure correction (i.e. remove the term $E^i(x,y)$ in (5.9)) are high as shown in Figure 5.4(b). The high weights favor over-exposure and suppress the occurrence of clouds in the composite result shown in Figure 5.3(d). After exposure correction, these weights become much lower as shown in Figure 5.4(c) and thus a desirable result with clouds is obtained in Figure 5.3(e).

Reference View Guided Consistency Assessment (RCA) As aforementioned, ACA assumes that for regions corrupted by movement, the the moving object is only a shot and another object mostly on the background predominantly exists. ACA can select the predominant object for the composition image. However, if some region changes frequently, e.g. the floors of the three exposures in Figure 5.5 are different with each other due to object movement, ACA cannot work well since no object is predominant in that region. Likewise, ACA cannot remove the ghosting artifacts caused by the moving background with dynamic objects such as windblown trees and waves. To remedy this issue, we seek to develop another consistency measure by taking one image as the reference view, which is thus named as RCA. To avoid ghosting artifacts,

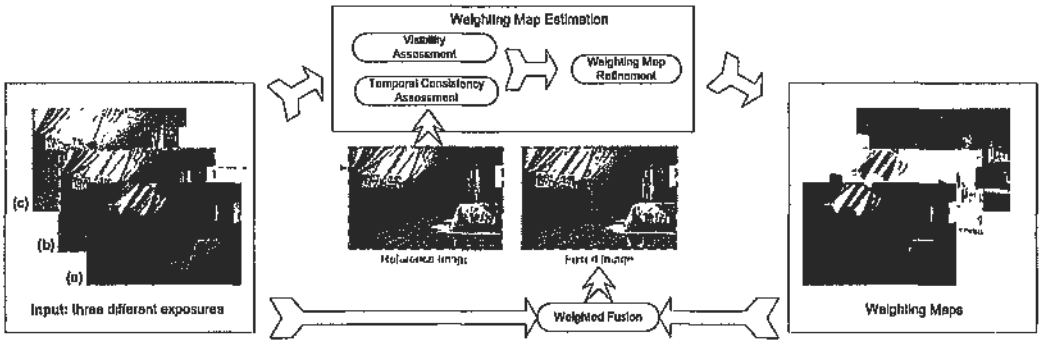

Figure 5.6: *Schematic overview of the proposed algorithm. (a), (b) and (c) are the same as those in Figure 5.5. Image (b) serves as the reference image. Note that the weighting maps shown in the right are normalized.*

additional visible details extracted from the other images will be accepted only if they are consistent with the scene defined by the reference view. Similar to [Gallo et al., 2009; Eden et al., 2006], we prefer the user to select reference image, since it will help remove the undesired moving objects and determine what the final, consistent HDR result will look like. Normally, the image whose motion area is well exposed is favored as the reference view. However, if all exposures are semantically equivalent to the users, the reference view can also be selected automatically based on which image has the least amount of saturated pixels similar to [Gallo et al., 2009].

Figure 5.6 illustrates how the proposed algorithm work with the reference view guided consistency assessment. In specific, the proposed HDR process begins with a set of differently exposed images. One image will be picked out from the stack as the reference view beforehand. Next, all images will undergo a comprehensive assessment on visibility and temporal consistency. The results are consolidated to weighting maps which will be further refined using crossbilateral filtering. Finally, the HDR result can be produced by compositing the exposures with the guidance of the weighting maps.

For each pixel of the i_{th} image, its direction change w.r.t the preselected reference image can be obtained as follows (similar to (5.7) with the j_{th} image as the reference view).

$$d_{i\rightarrow ref}(x,y) = \frac{\sum_{m=-r}^{r} \left|\theta^{i}(x+m,y+m) - \theta^{ref}(x+m,y+m)\right|}{(2r+1)^2}, \tag{5.12}$$

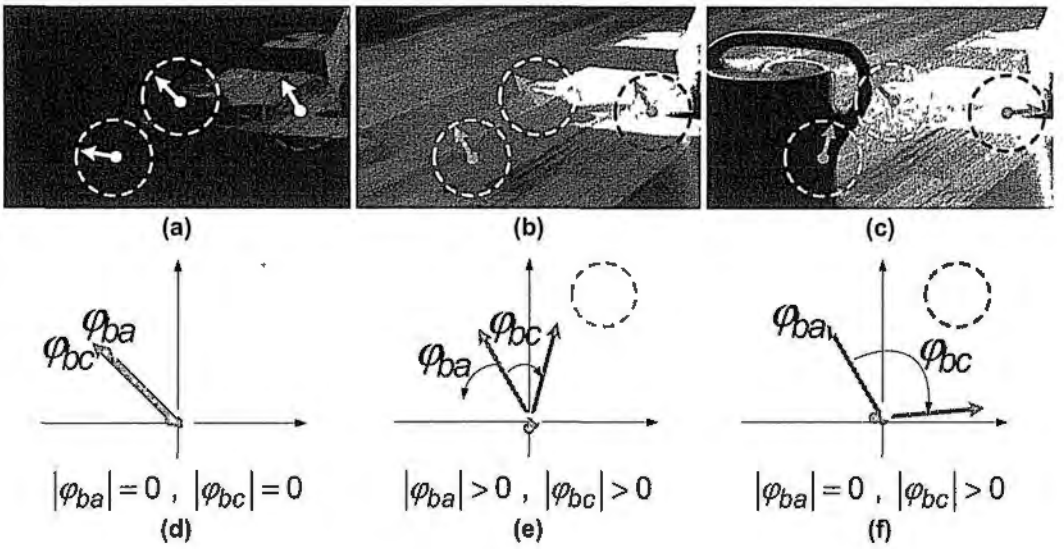

Figure 5.7: *Direction changes of image gradients in the example of Figure 5.6. Three patches (a), (b) and (c) are cropped from the corresponding input images. For illustration, we select three representative groups of gradients to explain the direction changes. Image (b) serves as the reference image, θ_{ba} and θ_{bc} are introduced to indicate the direction changes of gradients in (a) and (c). Note that the arrow is only used to indicate the gradient direction illustratively and its length does not represent the magnitude. Please enlarge to see more details.*

For the sake of simplicity, we just denote $d_{i \to ref}(x, y)$ with $d_i(x, y)$ in the rest of this chapter. Similar to (5.8), a gradient direction change based score S_R^i can be defined to reflect the consistency of each exposure w.r.t the reference image.

$$S_R^i(x, y) = exp(\frac{-d_i(x, y)^2}{2\sigma_s^2}). \tag{5.13}$$

It is noted that $S_R^i(x, y)$ always equals to 1 if i is equivalent to ref.

Likewise, the influence of over- and under-exposure to gradient direction should be considered as well. Similar to Figure 5.3(b), a brief analysis about the gradient direction change is given in Figure 5.7. Apparently, the influence of over- and under-exposure should be discussed in two cases. Case I: the pixels of the reference image are well-exposed as the example in Figure 5.7(f) where image (b) serves as the reference view. This case is quite tractable since the overexposed pixels in (c) can be suppressed by both visibility and consistency assessment. However, in case II where the pixel of the reference image is overexposed or underexposed (e.g. take image (c) as the reference view), the temporal consistency measured on this pixel may no longer be desirable, since

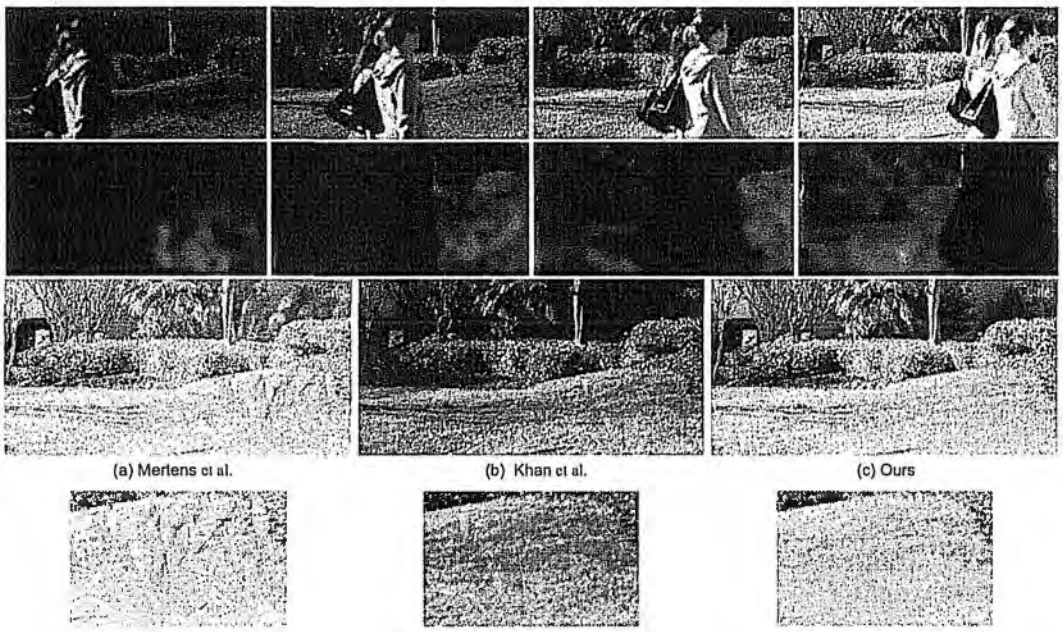

Figure 5.8: *Dynamic example with ACA. The top row shows four of the nine exposures. The second row shows their weighting maps estimated by our method. The bottom patches are cropped from (a), (b) and (c) for close-up comparison. Data courtesy of Erum Arif Khan.*

it will exclude the visible details presented in the other well-exposed images. Therefore, the final consistency $C_R^i(x,y)$ is calculated with exposure compensation as (5.9).

$$C_R^i(x,y) = \frac{S_R^i(x,y) \times E^{ref}(x,y)}{\sum_{i=1}^{K} S_R^i(x,y) \times E^{ref}(x,y) + \epsilon}, \tag{5.14}$$

where E^{ref} which indicates the exposure quality of the reference image I^{ref}, can remove the undesired gradient direction change caused by the over- or under-exposure of the reference image.

Similar to (5.11), the final weight $W_R^i(x,y)$ in this case can be obtained as:

$$W_R^i(x,y) = \frac{V^i(x,y) \times C_R^i(x,y)}{\sum_{i=1}^{K} V^i(x,y) \times C_R^i(x,y) + \epsilon}. \tag{5.15}$$

5.3 Experiments and Discussions

In this section, the proposed algorithm is tested in various static and dynamic scenes with different types of exposure sequences. Besides, we also show its potential in flash and no-flash photography. The amazing thing about the proposed method is that it

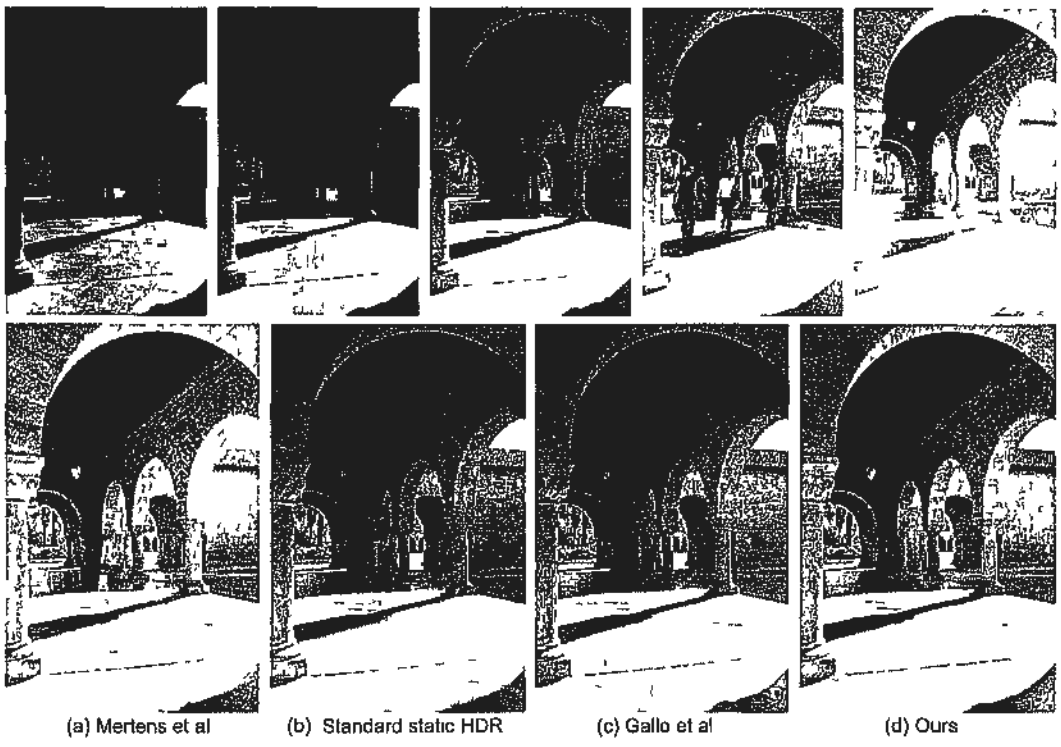

Figure 5.9 *Dynamic example with ACA The top row shows the input five exposures (a) Mertens et al 's result [Mertens et al 2009] (b) Result obtained using standard HDR (radiometric calibration and tone mapping) (c) Result presented in Gallo et al [Gallo et al 2009] (d) Our result Data courtesy of Orazio Gallo*

Figure 5.10 *Close-up comparison of (left) Gallo et al 's result in Figure 5.9(c) and (right) ours in Figure 5.9(d)*

does not require much parameter tweaking All experimental results were produced with the same parameters mentioned in the above sections For color images, gradient extraction and cross-bilateral filtering are conducted only in the luminance channel Please also visit `http //www ee cuhk edu hk/~zhangwei/GradComp html` to see the results

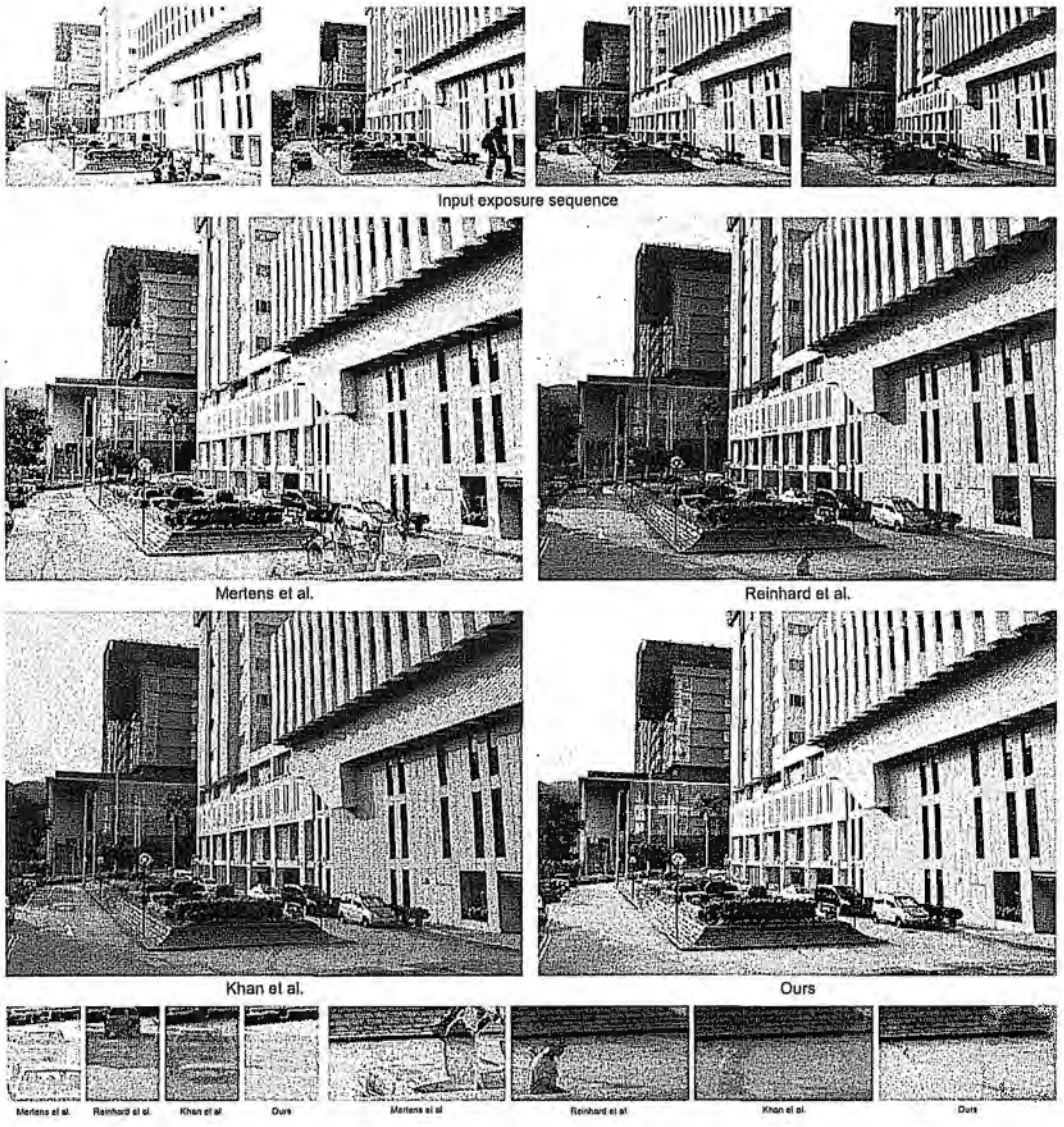

Figure 5.11: *Dynamic example with ACA. The left row shows the input three exposures. The right results are obtained by Mertens et al.'s result [Mertens et al., 2009], Reinhard et al. [Reinhard et al., 2005], Khan et al. [Khan et al., 2006] and ours, respectively. Apparently, our method gives the best result. Please enlarge to see more details.*

5.3.1 Dynamic Scene with ACA

In this part, the weights in (5.1) is set as: $W = W_A$. The following examples will prove that the proposed method not only can produce an image with extended dynamic range but also remove all unwanted moving objects. To validate the effectiveness of the proposed method in dynamic scenes, we reproduce some results published before and make comparisons with the existing work. Figure 5.8 shows a scene with people moving

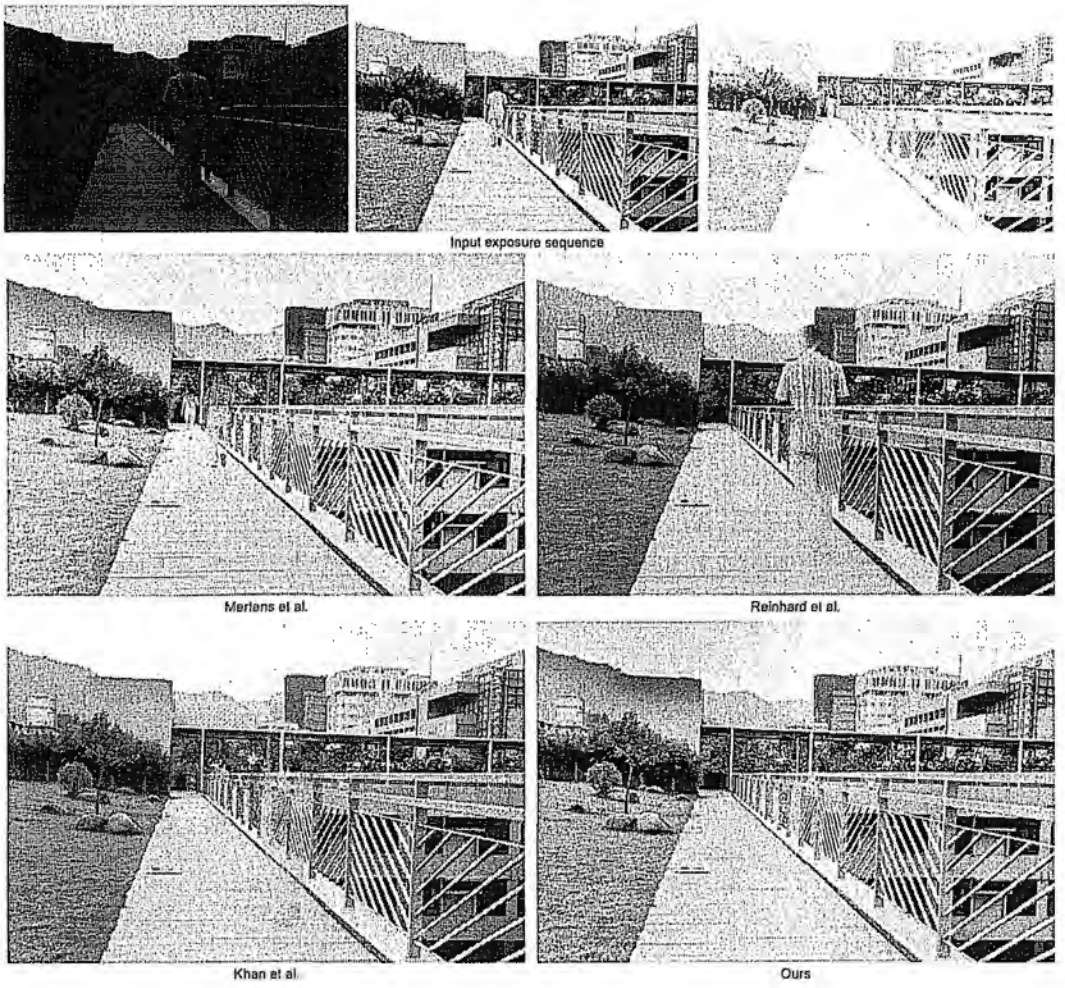

Figure 5.12: *Dynamic example with ACA. The top row shows the input three exposures. The bottom results are obtained by Mertens et al.'s result [Mertens et al., 2009], Reinhard et al. [Reinhard et al., 2005], Khan et al. [Khan et al., 2006] and ours, respectively. Apparently, our method gives the best result. Please enlarge to see more details.*

from left to right. Apparently, the result (a) generated by Mertens et al. [Mertens et al., 2009] suffers from severe ghosting artifacts. Although the deghosting result (b) presented by Khan et al. [Khan et al., 2006] is much better, some faint ghosting artifacts are still visible. Besides, the entire image (b) looks under-exposed especially around the trees and walls of the background. In contrast, our method generated a pleasant result (c) that is more informative and completely ghost-free. Figure 5.9 shows five differently exposed images with variable walking people. Similarly, the static methods like Mertens et al. [Mertens et al., 2009] and standard HDR produced noticeable ghosting artifacts as shown in (a) and (b). Our result (d) is comparable to the result (c) presented in

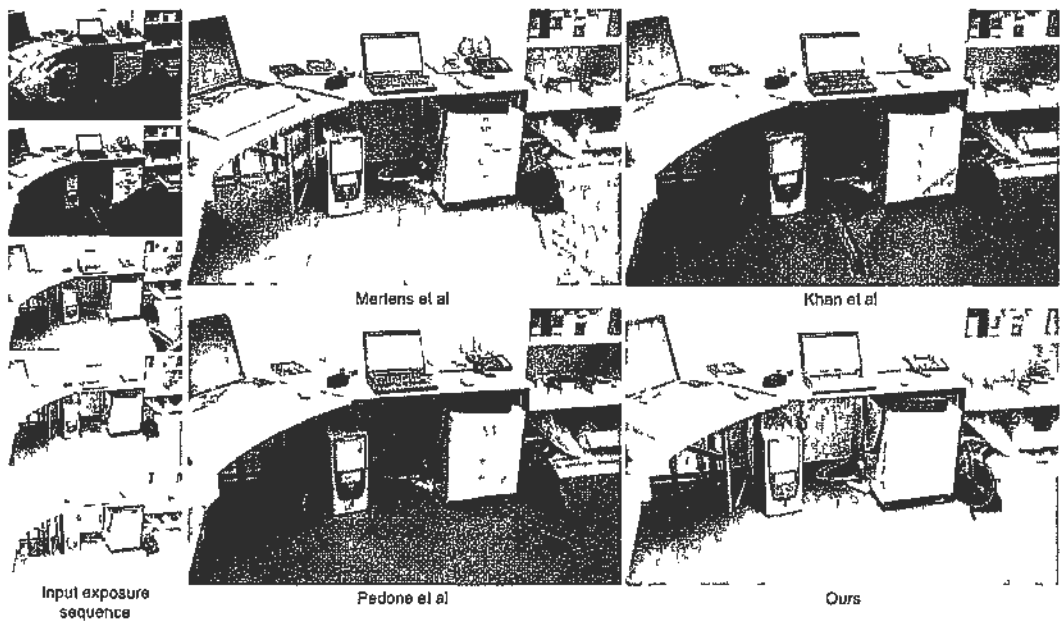

Figure 5.13 *Dynamic sample with RCA Left input sequence Right the result of Mertens et al [Mertens et al 2009], the result of Khan et al [Khan et al 2006], the result presented in Pedone et al [Pedone and Heikkil 2008] and ours (the forth image serves as the reference) Data courtesy of Matteo Pedone*

Gallo et al [Gallo et al 2009] in terms of ghost removal However, as shown in Figure 5 10, our result is less noisy and exhibits more details than Gallo et al 's It is worth noting that the performance of Gallo et al 's method relies on the quality of the selected reference image It cannot be used to remove the unwanted moving objects if they are present in all exposures as those in Figure 5 3, since no image is suitable for reference Sometimes, human intervention is required to help select a image as the reference view for deghosting For example, the fifth exposure is selected as reference manually in Figure 5 9 In contrast, our approach is fully automatic As shown in Figure 5 11 and Figure 5 12, we also compare our method to [Reinhard et al 2005] which achieves deghosting based on variance measurement

5.3.2 Dynamic Scene with RCA

In this part, the weights in (5 1) is set as $W = W_R$ Some of the following examples will prove that the proposed method is also able to produce a HDR image with some moving objects that the user desired Particular, one interesting example has been shown in Figure 5 5 In addition to the desirable performance on ghosting removal, our method

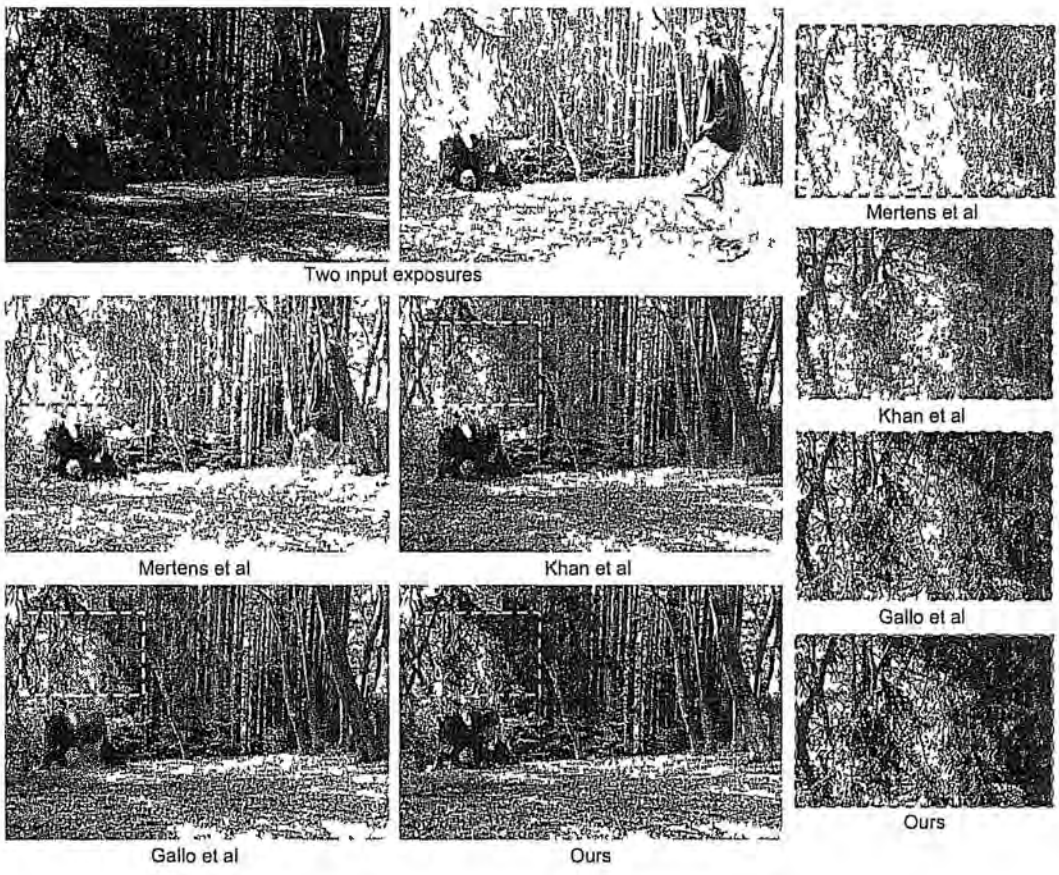

Figure 5.14 *Dynamic sample with RCA. Top row: input sequence. Bottom two rows: the results obtained by Mertens et al.'s exposure fusion [Mertens et al. 2009], Khan et al. [Khan et al. 2006], Gallo et al. [Gallo et al., 2009] and our method. Data courtesy of Orazio Gallo.*

is also promising in that it can produce different types of HDR photos by changing the reference view. The example in Figure 5.13 gives five differently exposed images which capture an indoor scene with moving objects (e.g. hand, chair). Not surprisingly, due to the lack of ghost removal, conventional exposure fusion method [Mertens et al. 2009] produced an unpleasant result with severe ghosting artifacts. Our method yielded a plausible ghost-free result, and outperformed the other deghosting methods [Khan et al., 2006, Pedone and Heikkila, 2008] significantly. Figure 5.14 shows another type of scene with chaotic motions. In addition to the moving people, the branches and leaves also tremble in the wind. Likewise, the result produced by conventional exposure fusion method [Mertens et al. 2009] suffers from severe ghosting artifacts caused by the moving people and windblown trees. The deghosting result produced by Khan et al. [Khan et al. 2006] is still unsatisfying, and some ghosts survive. Especially, it

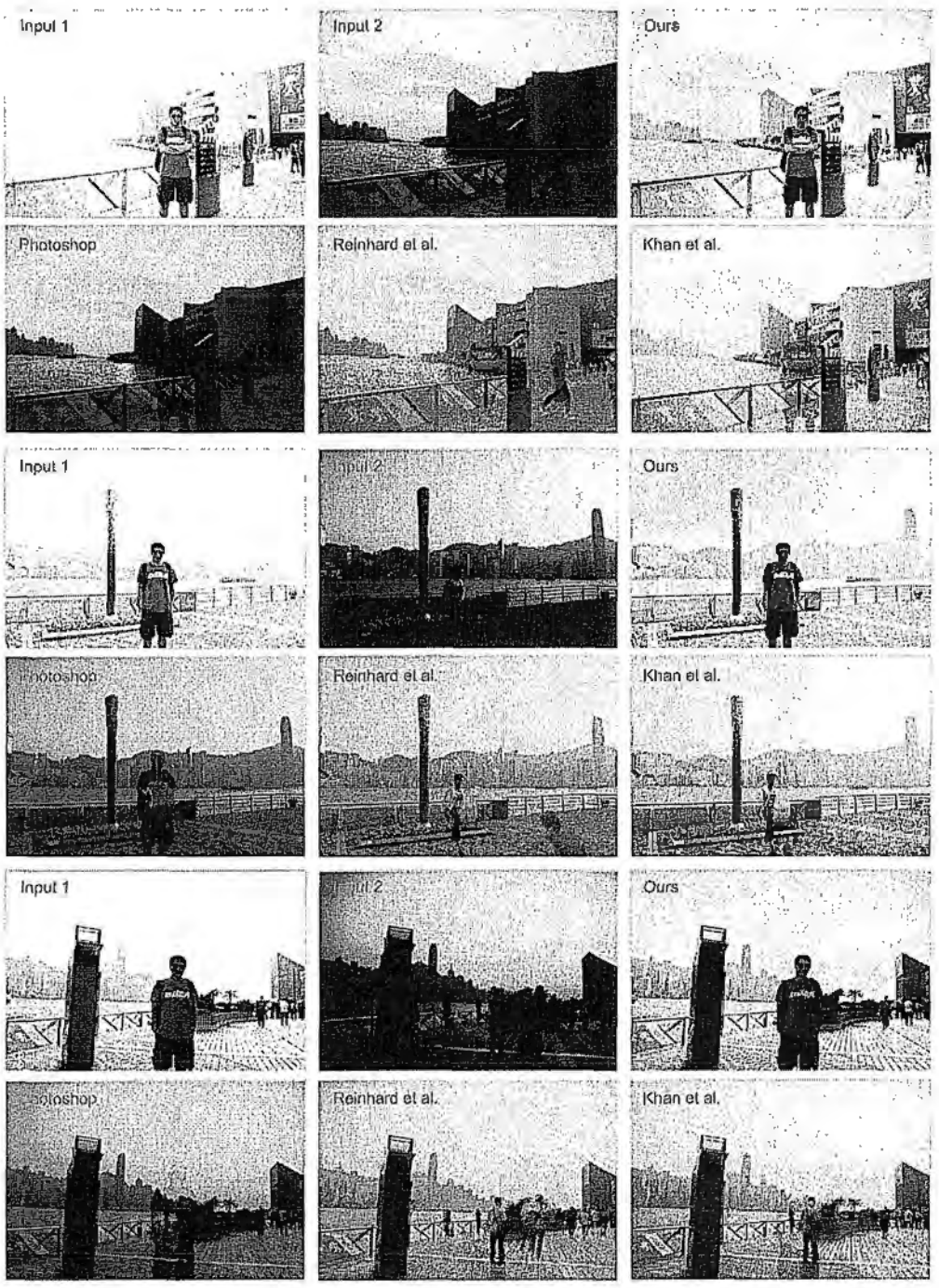

Figure 5.15: *Dual-photography with RCA. Input 1 and input 2 are manipulated to capture the man and the sunset respectively. The bottom two rows show the results obtained by* Photoshop *(no deghosting), Khan et al. [Khan et al., 2006], Reinhard et al. [Reinhard et al., 2005] and our method (input 1 serves as the reference).*

cannot remove the ghosts caused by the background blowing trees. By contrast, our method produced a pleasant result completely free of ghosts, and the performance is

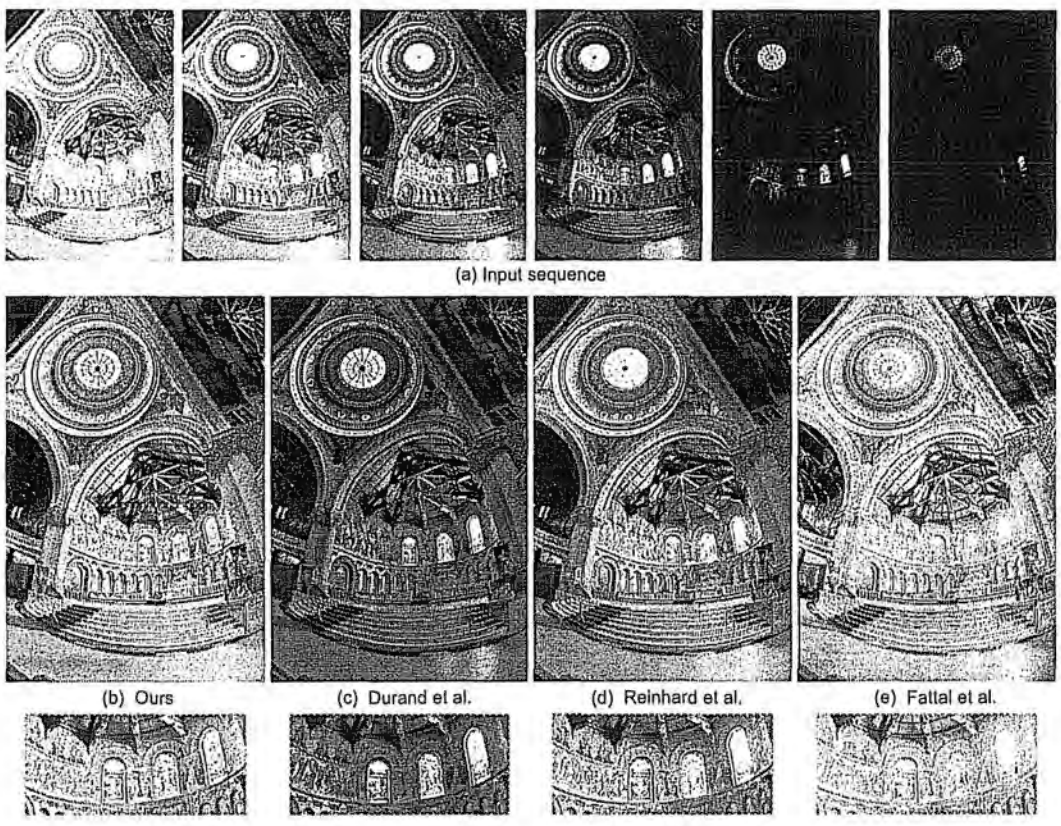

Figure 5.16: *Comparison with tone mapping operators. (a) Input six exposures. (b) Our composite result. (c) Durand et al. [Durand and Dorsey, 2002]. (d) Reinhard et al. [Reinhard et al., 2002]. (e) Fattal et al. [Fattal et al., 2002]. Data courtesy of Paul Debevec.*

comparable to that of Gallo et al.'s method [Gallo et al., 2009]. Note that in order to make a fair comparison, we use the same image as that in Gallo et al.'s work as the reference view.

Next, we employ the proposed algorithm to deal with a trouble problem people often meet when traveling. It is acknowledged that taking a good stack in a busy tourist resort is not easy because of the moving objects (e.g. people, cars and boats). As shown in Figure 5.15, our method can ease this trouble situation. In most cases, people only need to take two photographs which capture the subject and the ambient scene respectively with appropriate exposure times. In fact, this kind of exposure stack can even be obtained without helpers if using self-timer shooting and tripod. Apparently, our method yielded a desirable result where the man and sunset are both captured, whereas the results obtained by standard HDR of *Photoshop* and deghosting methods [Khan et al., 2006; Reinhard et al., 2005] are disappointing.

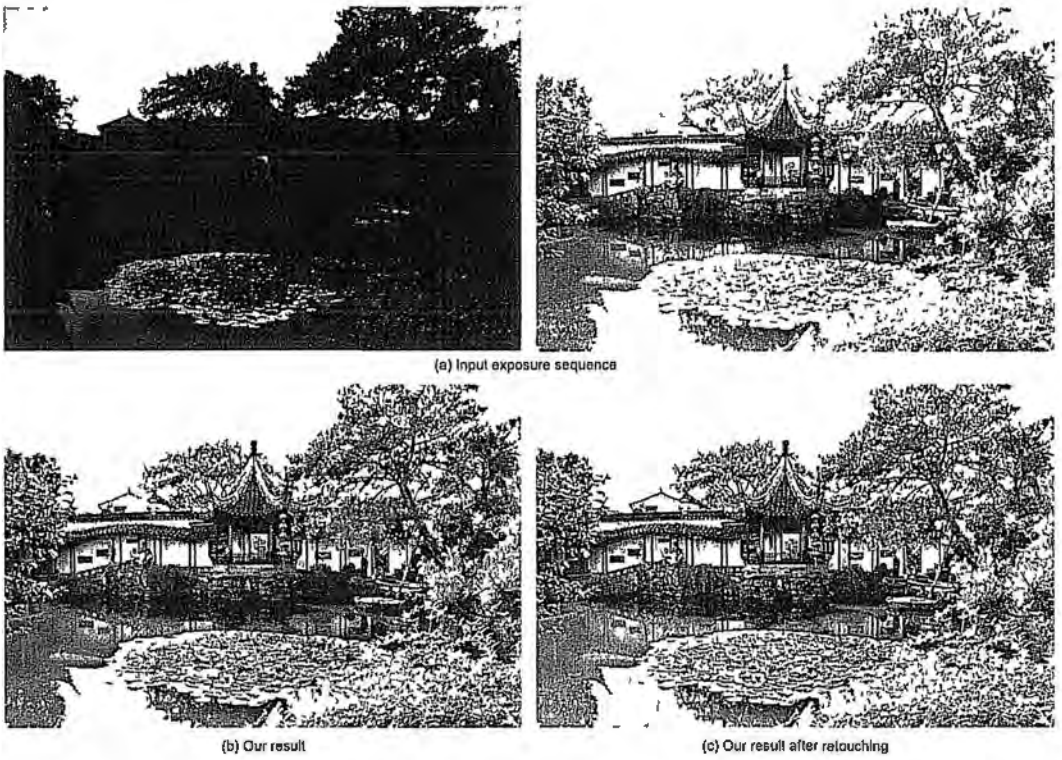

Figure 5.17 *Static example. (a) shows the input exposure sequence. (b) is our original result (c) is the enhanced version of (b) generated using some retouching techniques of Photoshop (e g contrast enhancement, saturation adjustment)*

5.3.3 Static Scene

In this part, the weights in (5 1) is set as: $W = W_V$. Figure 5 16 shows the comparison between our method and some prevalent tone mapping operators in a static scene. Note that our result is visually pleasing although it is generated with only six exposures of the original sequence. The overall quality is comparable to that of the tone mapping results produced from the whole radiance map. More static results are given in Figure 5.17 and Figure 5 18. Similarly, our method produced promising result where the details in both bright and dark regions are preserved. Besides, it is observed that retouching make the results more impressive Therefore, exposure composition + retouching is a promising alternative to conventional HDR technique (radiometric calibration + tone mapping)

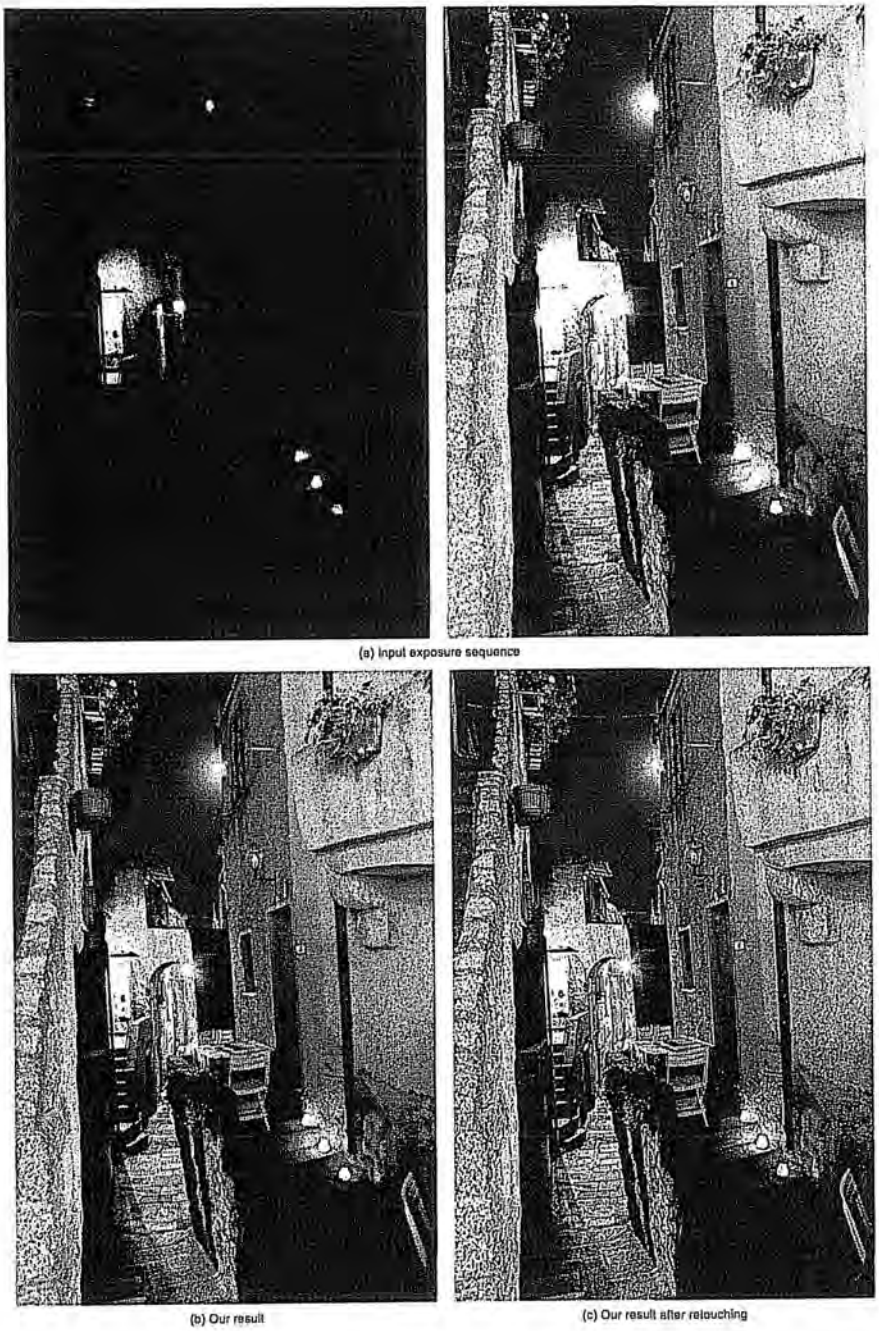

Figure 5.18: *Static example. (a) show the input exposure sequence. (b) is our original result. (c) is the enhanced version of (b) generated using some retouching techniques of Photoshop (e.g. contrast enhancement, saturation adjustment).*

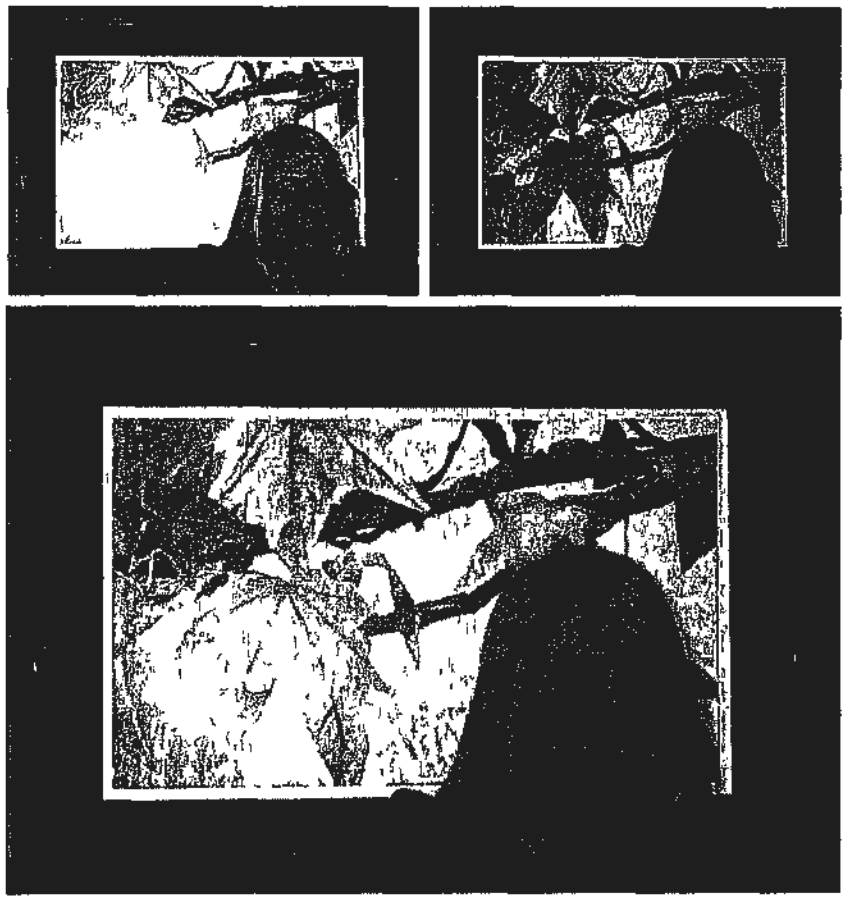

Figure 5.19 *Flash hot spot removal using visibility assessment Top flash and no-flash images Bottom our composite result Data courtesy of Amit Agrawal*

5.3.4 Computational Efficiency

As aforementioned, conventional deghosting methods such as Khan et al [Khan et al 2006], Gallo et al [Gallo et al 2009] and Reinhard et al [Reinhard et al 2005] are computationally expensive, since camera calibration and tone mapping are both required Moreover, Khan et al 's method [Khan et al 2006] works in an iterative manner In contrast, our method is quite simple and non-iterative For deghosting (ACA and RCA) in a dynamic case, the current non-optimized Matlab implementation takes about 25 – 35 seconds to process four 1 megapixel images on a PC with an Intel Core2Duo 3 0GHz CPU Note that it is hard to give the exact running time of the work [Khan et al 2006, Gallo et al 2009, Reinhard et al 2005] due to their complex pipelines, and that user intervention is usually required in the tone mapping

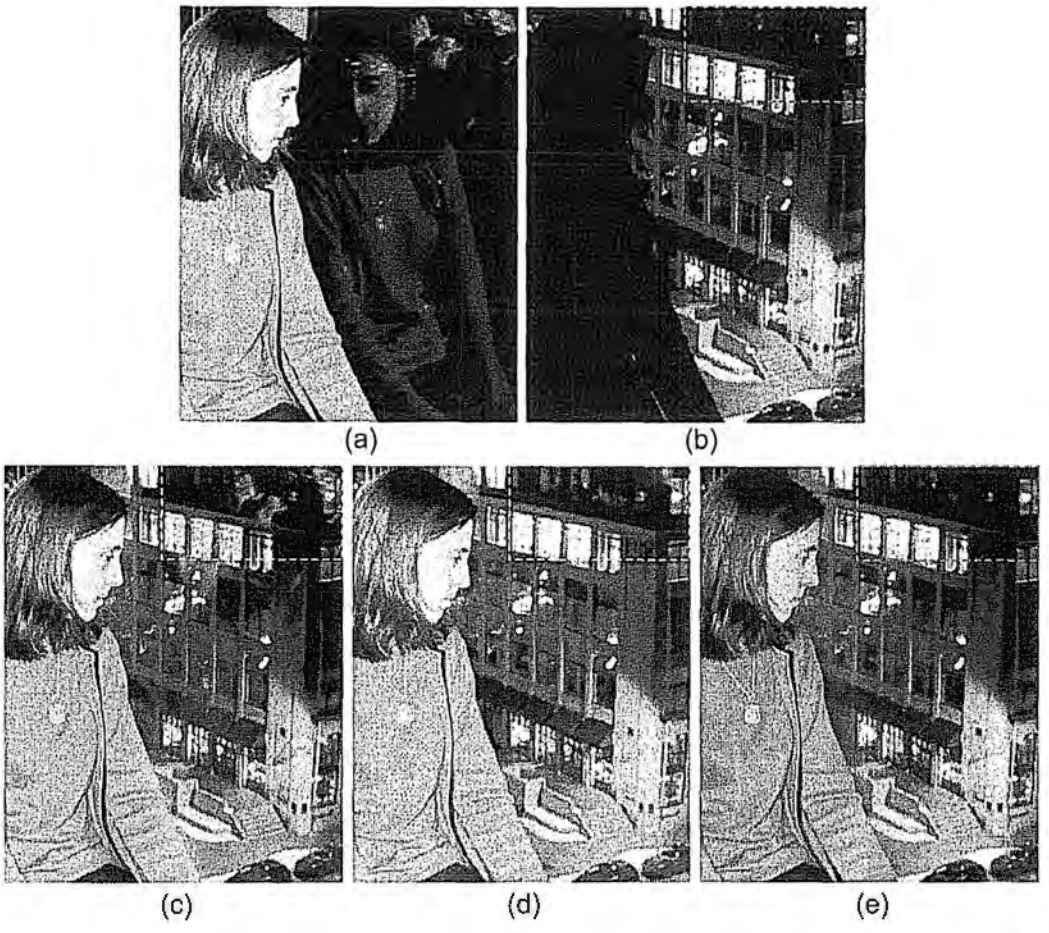

Figure 5.20: *Reflection removal. (a) Flash photo. (b) No-flash photo. (c) Mertens et al. [Mertens et al., 2009]. (d) Agrawal et al. [Agrawal et al., 2005]. (e) Our result (select the no-flash photo as the reference image). It is observed that our result (e) is slightly better than Agrawal et al.'s result (d) in the regions outlined by dashed rectangle. Note that [Mertens et al., 2009] and our method also corrected the over-exposedness of the flash image especially on the girls face. Data courtesy of Amit Agrawal.*

step to achieve acceptable results. Normally, they are running at minute level. Since consistency assessment is unnecessary, our algorithm works much faster in static scenes and takes less than 15 seconds to tackle the same amount of images.

5.3.5 Application in Flash and No-flash Photography

Photography in dark environment can be improved by combining flash photos with no-flash photos. Next, we will show how to use the proposed method to solve some annoying artifact problems we often encounter in flash and no-flash photography.

(*Spot removal*) Figure 5.19 shows two indoor images taken with flash and no-flash.

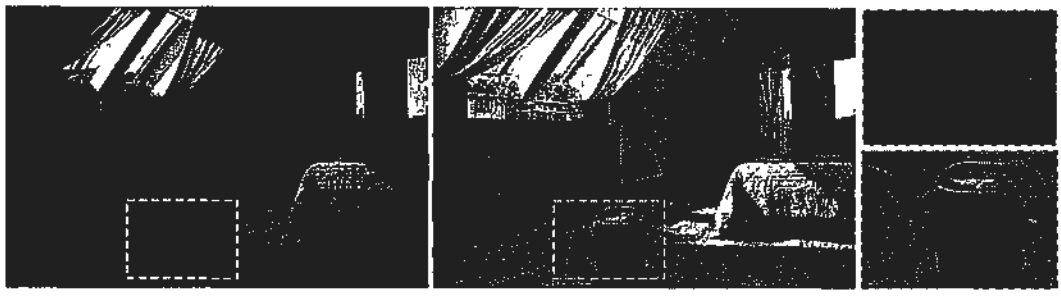

Figure 5.21: *A failure case in the example of Figure 5.5. Left: the reference image (image (a) of Figure 5.5). Middle: our composite result Right: close-up view of the risky regions where ghosting artifacts occurs*

The no-flash image is faithful to the ambient lighting, while the flash image reveals more details but suffers from hot spot artifacts. Our method with visibility assessment ($W = W_V$) can combine the advantages of them and generate a desirable image free of hot spot.

(*Reflection removal*) If there is transparent layer such as glass, flash also incurs reflection artifacts. As shown in Figure 5.20, a person is photographed from inside a glass enclosed room at night. Flash photo (a) can capture the person but exhibit reflection artifacts, while no-flash photo (b) can only take the distant building behind the glass. In this case, direct fusion [Mertens et al., 2009] incurs reflection artifacts as shown in Figure 5.20(c). Our proposed approach with RCA ($W = W_R$) seems to produce the best images with correctly removed reflection artifacts. Also, compared to Agrawal et al. [Agrawal et al., 2005], our method also corrected the overexposure of the flash image especially on the girl's face.

5.3.6 Limitations

The method we proposed also shares the common limitations in HDR technology. For example, it may not work well when the input exposures contain severe sensor noise or blurring artifacts caused by camera shake, since the gradient estimation might be inaccurate in these cases. One possible solution is to denoise or deblur the input images first and then proceed our HDR scheme. Besides, since ACA is developed based on the assumption that the stationary parts of the scene are predominant in the sequence, at least three exposures are required when perform deghosting with ACA in dynamic scenes. However, there is no such limitation for RCA. In most cases, two photographs

are good enough to get a pleasing result such as Figure 5.15. But RCA may fail to deghost if an unsuitable exposed image is selected as the reference view. For example in Figure 5.21, since the object in the risky region where motion occurred is underexposed in the reference image, it cannot provide effect gradient guidance for ghosting removal. However, this kind of failure is avoidable to some extent by taking the other exposures as the reference view as shown in Figure 5.5. As aforementioned, the principle is to take the image which is well exposed in the risky region as the reference view. Otherwise, the proposed method may not work well. In the future, we would like to introduce a more advanced model for ghosting removal to improve the current one-image-dependent reference view strategy.

5.4 Summary

Image gradients convey important information about the latent scene. In this chapter, we have shown that well utilization of image gradient makes it possible to handle the static and dynamic exposure composition in a simple but effective way. We have designed two kinds of consistency measures to deal with the two types of movement: foreground object movement and background object movement. Similar to [Goshtasby, 2005; Mertens et al., 2009; Shanmuganathan and Chaudhuri, 2009], the proposed method can free users from the tedious radiometric calibration and tone mapping steps. The effectiveness and efficiency of the proposed approach have been validated with various exposure sequences captured in different dynamic scenes.

5.5 Supplementary Results

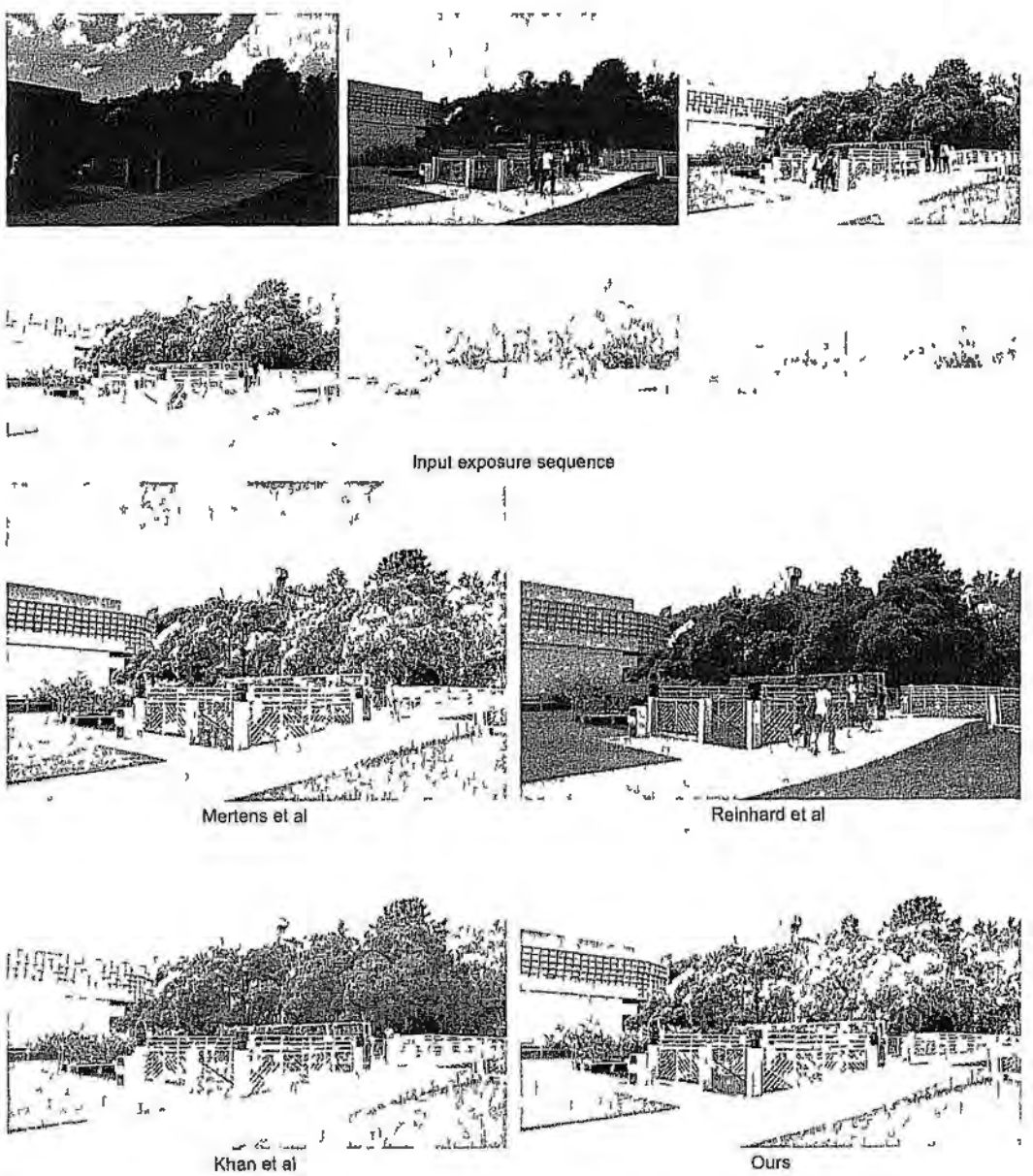

Figure 5.22 *Dynamic example with ACA The top two rows show the input exposure sequence The bottom results are obtained by Mertens et al 's result [Mertens et al 2009], Reinhard et al [Reinhard et al 2005], Khan et al [Khan et al 2006] and ours, respectively*

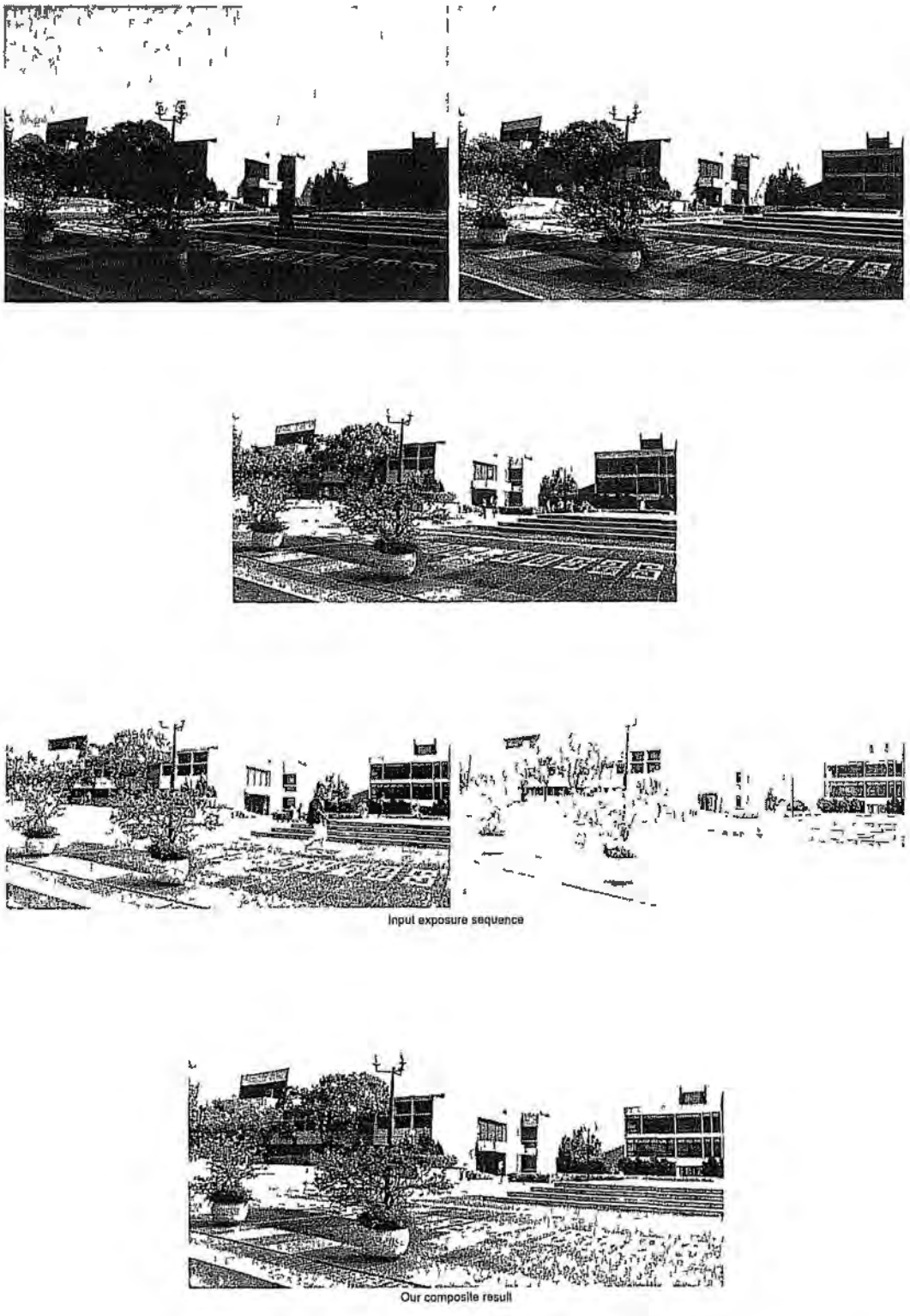

Figure 5.23 *Dynamic example with ACA. The top three rows show the input exposure sequence. The bottom row shows our artifact-free composite result.*

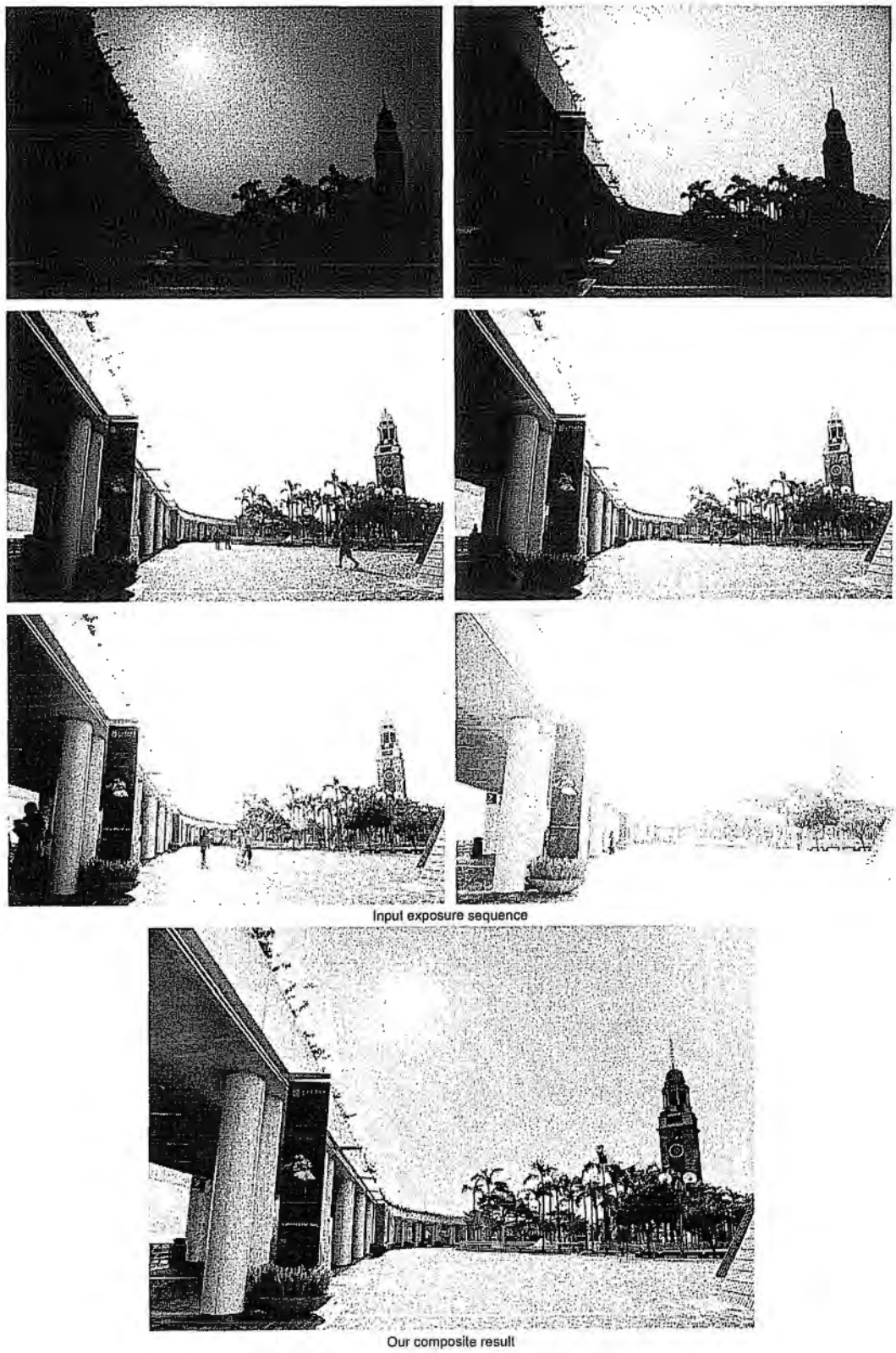

Figure 5.24: *Dynamic example with ACA. The top three rows show the input exposure sequence. The bottom row shows our artifact-free composite result.*

Figure 5.25 *Dynamic example with RCA Top two rows input sequence Third row our result (the third exposure serves as the reference)*

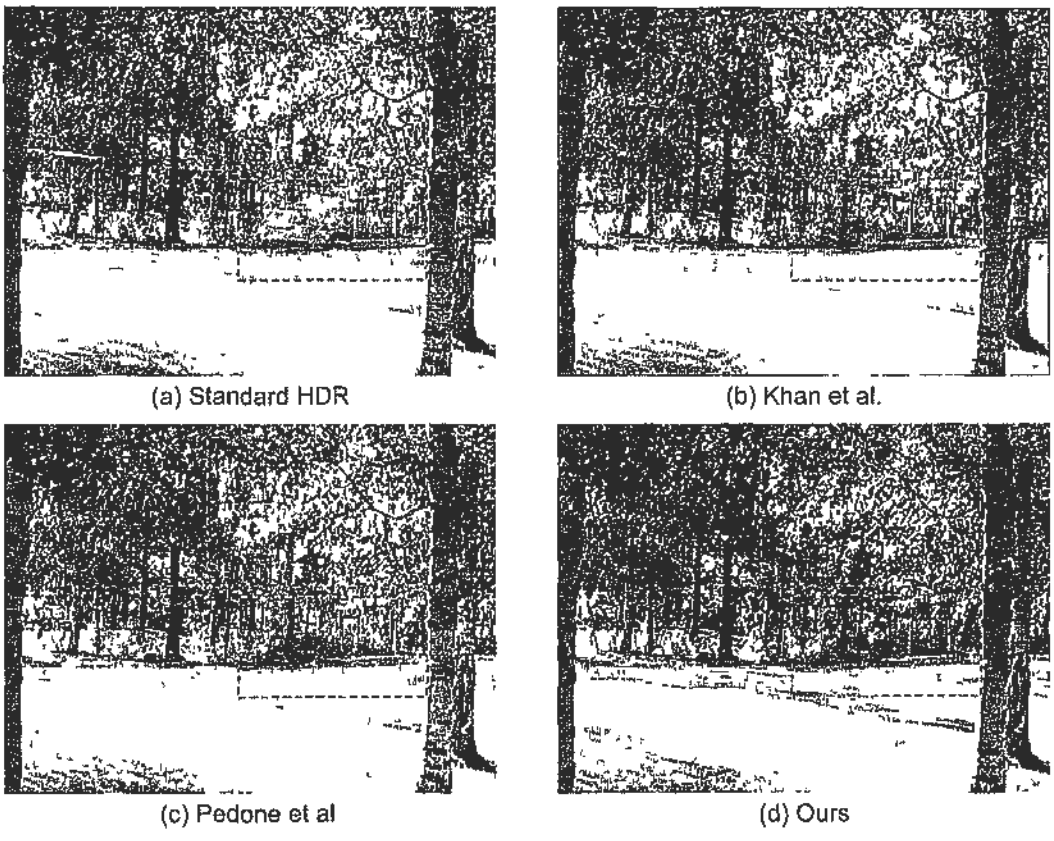

Figure 5.26 *Comparison of the example in Figure 5.25. (a) standard HDR (no deghosting), (b) Khan et al. [Khan et al. 2006], (c) Pedone et al. [Pedone and Heikkila 2008], (d) ours. Please notice the regions outlined by the dashed rectangles. Result (a) suffers severe ghosting artifacts caused by the moving car (see the blue dashed rectangle) and windblown leaves (see the red dashed rectangle). [Khan et al. 2006] and [Pedone and Heikkila 2008] can relieve the ghost problem incurred by the moving car, but cannot remove the others caused by the windblown trees, because as mentioned in the chapter, both of them cannot handle the frequent movement. Our method yielded the best result where all ghosts have been removed completely. Please enlarge to see more details. Data and results (a),(b),(c) courtesy of Matteo Pedone.*

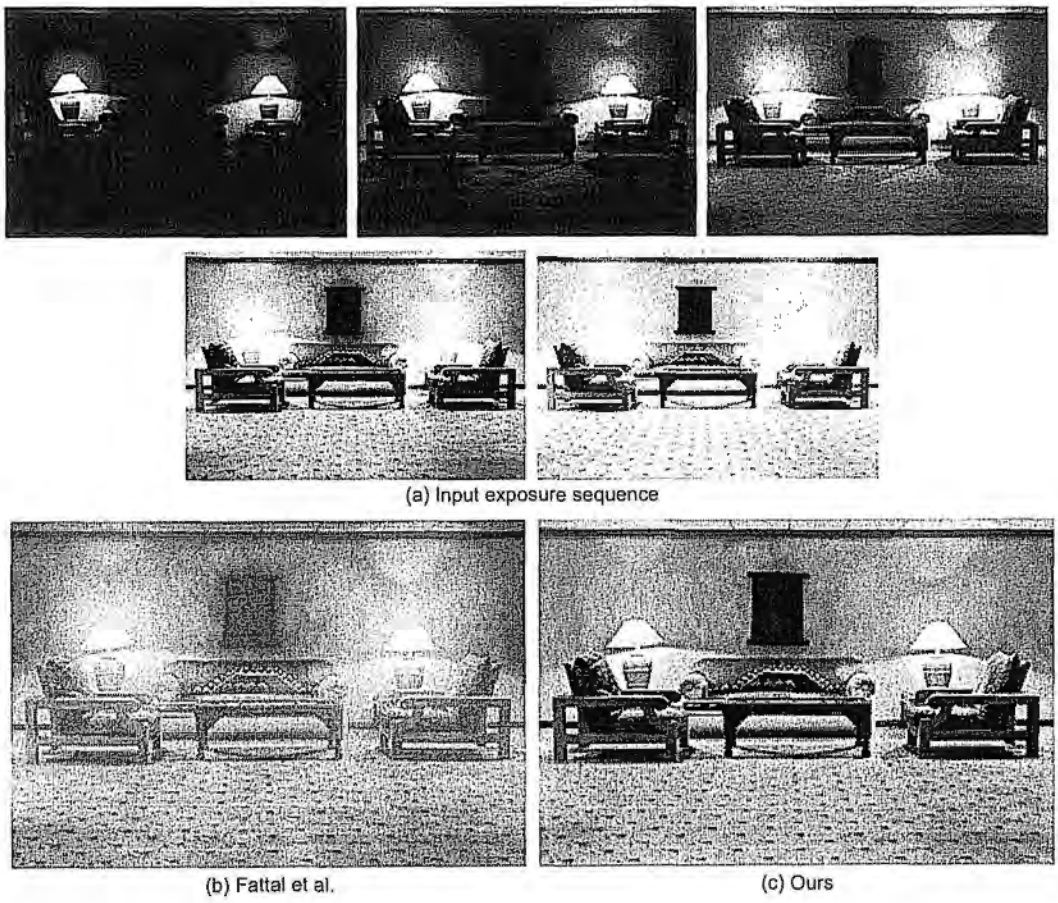

Figure 5.27: *Static example. (a) Input exposure sequence. (b) Tone mapping result published in the project web page of Fattal et al. [Fattal et al., 2002]. (c) Our result.*

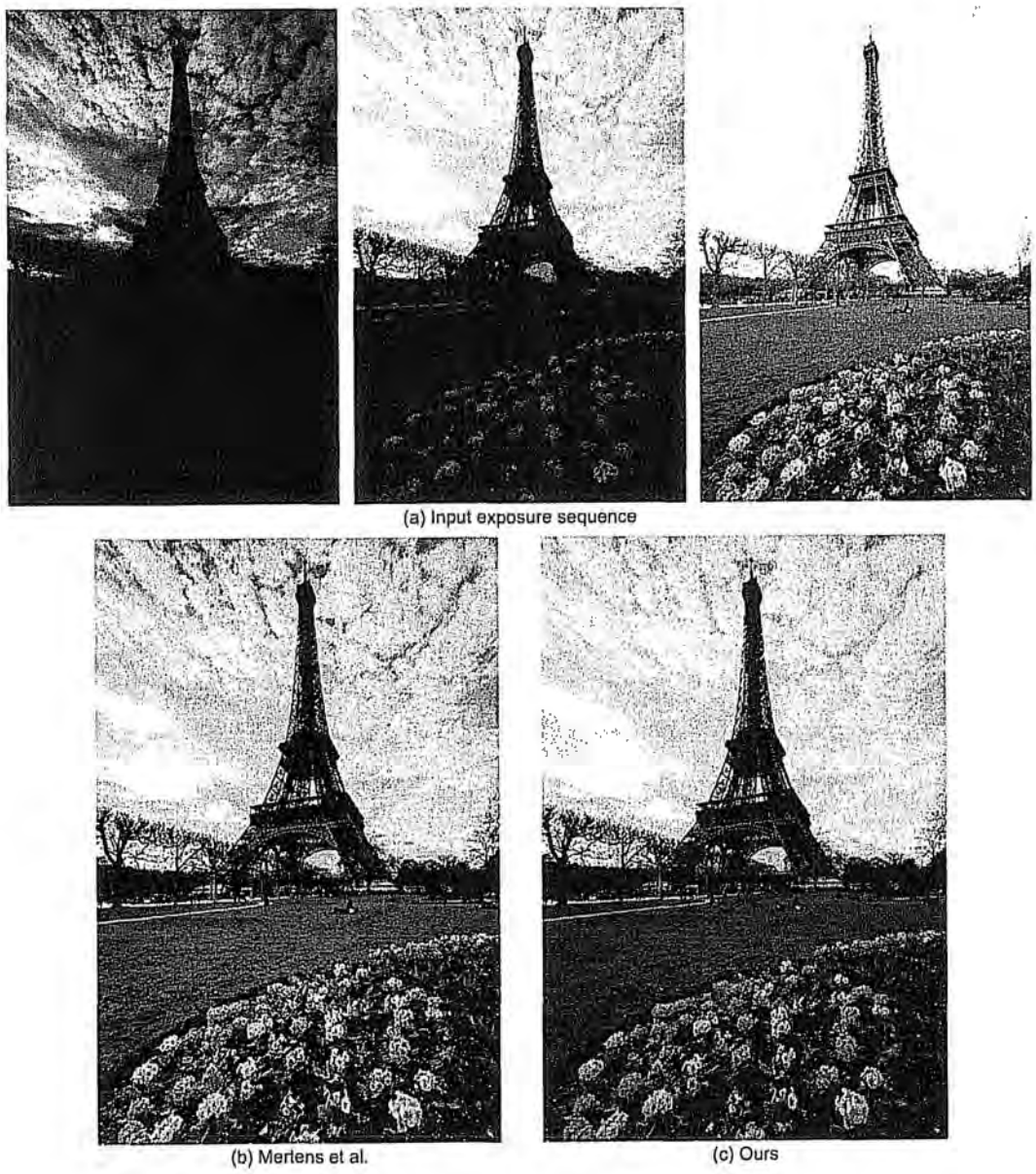

Figure 5.28: *Static example. (a) Input exposure sequence. (b) Result generated with the original codes of Mertens et al. [Mertens et al., 2009]. (c) Our result.*

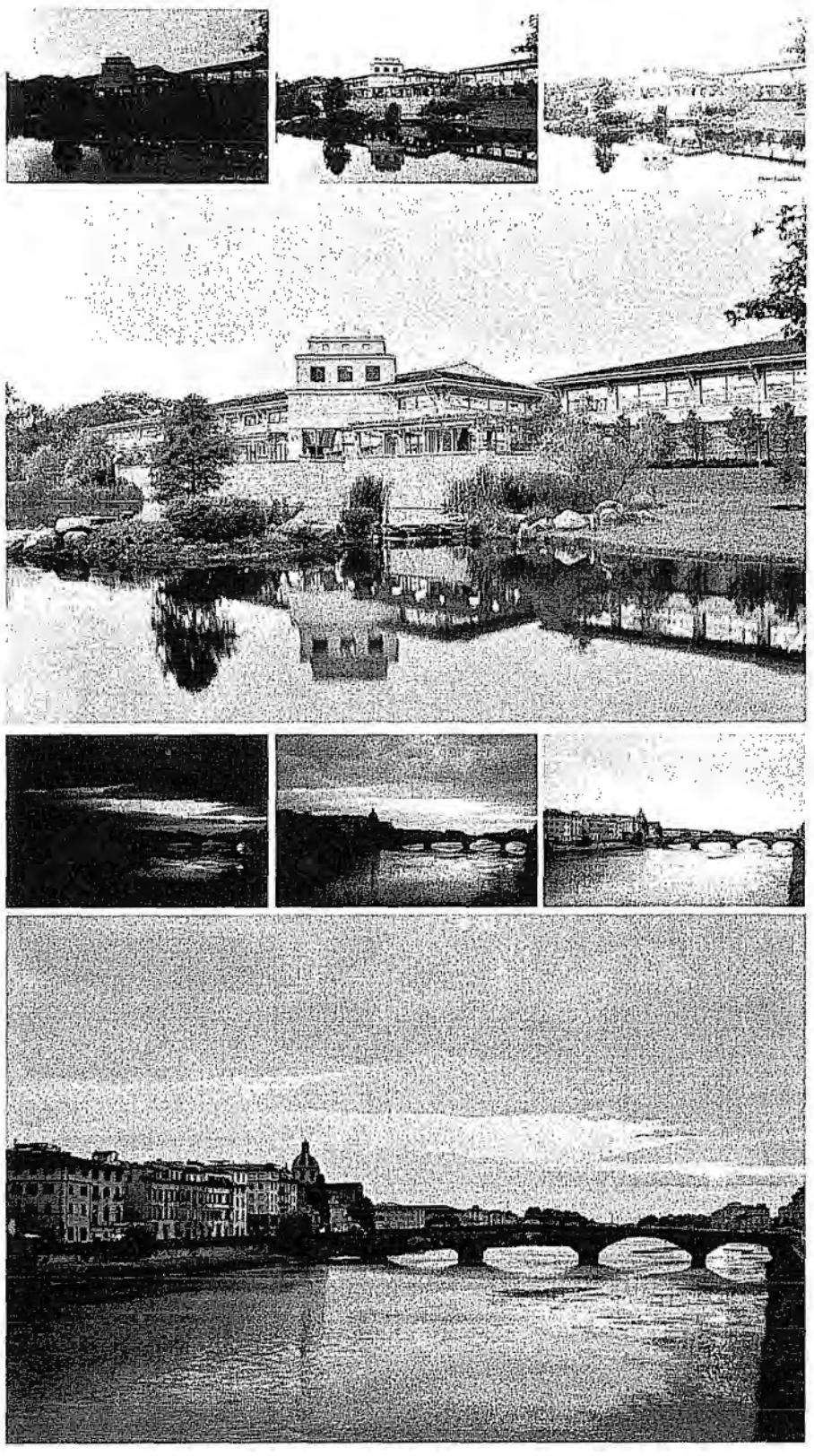

Figure 5.29: *Static example. Left: input exposure sequence. Right: composite image.*

Chapter 6

Conclusions and Future Work

This chapter closes the thesis with a summary of the main contributions and several directions for further work.

6.1 Contributions of the Thesis

To break the physical limits of cameras and turn the captured image to be what people are looking for, this thesis has presented a series of image enhancement algorithms which can improve the perceptual quality of the captured images in three aspects: resolution, focus effect and dynamic range. The main contributions can be summarized as follows:

6.1.1 Super-resolution

To enhance the resolution of a captured image, two kinds of super-resolution (SR) approaches are presented in Chapter 2 and Chapter 3. Chapter 2 aimed at super-resolving face images (i.e. face hallucination) with a learning-based framework [Zhang and Cham, 2008; Zhang and Cham, b]. Unlike previous learning-based work, face hallucination problem is addressed from a different perspective. In details, the problem is formulated as inferring the DCT coefficients in frequency domain instead of estimating pixel intensities in spatial domain. As shown in Section 2.3, DC coefficients can be estimated fairly accurately by simple interpolation-based methods. AC coefficients, which contain the information of local features of face image, cannot be estimated well using interpolation. An efficient learning-based inference model is proposed to infer the AC coefficients in Section 2.4. The proposed approach requires less memory and lower computation cost than conventional methods because firstly the data dimension is significantly reduced in the DCT domain. Secondly, clustering is implemented to remove the redundancy of the training set. Experiments were conducted to demonstrate

the effectiveness of the proposed method in producing high quality hallucinated face images.

Chapter 3 aimed at super-resolving generic images (i.e. face hallucination) with a reconstruction-based framework [Zhang and Cham, 2010b]. In detail, the SR process is straightforwardly divided into two steps: magnification and deblurring. Magnification is achieved using structure adaptive interpolation to avoid jaggy artifacts. Deblurring is a highly ill-posed blind deconvolution problem. Unlike previous work, we advocate solving it in an efficient and non-iterative way with the aid of little user intervention as stated in Section 3.3.2. Specifically, after introducing a parametric edge model, we show that the blurring kernel can be estimated accurately and quickly from the salient edges selected by user-drawn stroke. When the blurring kernel is fixed, the sharp image is recovered effectively with a maximum a posteriori (MAP) framework. Experiments on a variety of images demonstrate that the proposed algorithm is able to generate visually appealing super-resolved results with few artifacts.

6.1.2 Focus Editing

To change the focus of a image, a focus editing system [Zhang and Cham, 2009; Zhang and Cham, c] is presented in Chapter 4. In detail, the proposed system can accomplish the tasks of focus map estimation, image refocusing and defocusing. Given an image with a mixture of focused and defocused objects, we first detect the edges and then estimate the focus map based on the edge blurriness which is depicted explicitly with a well-parameterized model as stated in Section 4.4. In Section 4.5, the image refocusing problem is addressed in an elaborate blind deconvolution framework, where the image prior is modeled well by using both global and local constraints. Especially, we correct the defocused blurry edges to sharp ones with the aid of the parametric edge model and then render this cue as a novel local prior to ensure the sharpness of the refocused image. Experimental results demonstrate that the proposed system performs well in producing different styles of realistic images from a single input by focus editing.

6.1.3 Exposure Composition

To break the dynamic range limits of conventional cameras and simulate high dynamic range (HDR) photography, Chapter 5 presents a simple but effective method [Zhang

and Cham, 2010a; Zhang and Cham, a] that can accomplish the multi-exposure image composition in both static and dynamic scenes. Given multiple images with different exposures, the proposed approach is capable of producing a pleasant tonemapped-like HDR image by compositing them seamlessly with the guidance of gradient-based quality assessment. Especially, novel quality measures on visibility and consistency are developed in Section 5.2.2 based on the observation of gradient change among different exposures. Compared to previous work, our method is quite appealing in practice since it is computationally efficient and frees users from the tedious radiometric calibration and tone mapping process. More importantly, two kinds of consistency measures are designed by take advantaging of the gradient direction change in Section 5.2.2. One is named as accumulated consistency assessment (ACA), which can be used to remove all unwanted moving objects and produce a clean HDR image. The other is named as reference view guided consistency assessment (RCA), which is intended for compositing all exposures by taking one preselected image as a substrate. Hence, the proposed method is also able to produce a HDR image with some moving objects that the user desired. Various experimental results in static and dynamic scenes demonstrate the effectiveness of the proposed method.

6.2 Future Research Directions

From the current work, there are several interesting avenues for future research.

- The proposed learning-based method in Chapter 2 attempts to generate a plausible high-resolution (HR) face image by creating the required high frequency components from a face training set. Hence, the performance limited by the training set and depends on how well the low-resolution (LR) input matches the training samples. Currently, the training set for faces with a particular pose and expression is only applicable for the hallucination of faces under similar conditions. It would nice to allow the algorithm to handle faces with different poses or expressions as [Li and Lin, 2004] and [Jia and Gong, 2008]. A possible solution is to first detect the pose or expression of a face and then perform face hallucination using the corresponding training samples. But this requires a more comprehensive training set which collects face images with diverse poses and expressions. Besides, this algorithm is applicable to general SR problem and can be

generalized to tackle other types of images. Likewise, to super-resolve a certain type of images well, an appropriate training set which includes sufficient similar type of images needs to be prepared. For example, to super-resolve an image with flowers, it is important that the training set contains many flower images.

- The SR process in Chapter 3 is divided into two steps: magnification and deblurring. The challenging deblurring problem is addressed in an interactive way. User intervention is required to select a salient edge from the target image for suitable deblurring. Hence, it would be necessary to develop a system with friendly interface that can make the users manipulate the process easily. Also, we would like to avoid the user intervention and propose an automatic method to find the optimal blurring kernel. Besides, it is worthy to investigate how to estimate the blurring kernel accurately if no salient edges can be found in the target LR image.

- The image SR algorithms presented in Chapter 2 and Chapter 3 both have potential for tackling the video SR problem. But the following aspects need to be investigated. First, for the learning-based framework presented in Chapter 2, the key problem is to construct a good training set for video SR. The training samples can be obtained by exploiting the external information (e.g. HR images of the target scene as in [Kong et al., 2006]) beyond the video or the internal information inside the video (e.g. patch redundancy as in [Joshi et al., 2004; Protter et al., 2009; Glasner et al., 2009]). Second, for reconstruction-based framework presented in Chapter 3, a flexible way to find the optimal blurring kernel is desired. Third, an implicit or explicit temporal coherence constraint is needed to avoid observable flickering artifacts.

- As mentioned in Chapter 4, the proposed focus editing method cannot generate a desirable focus map in cases where the input image contains objects naturally blurry or many focus layers with large discontinuities. To relieve this issue, we hope to introduce user intervention as in [Yan et al., 2009; Bando and Nishita, 2007] to help the method handle the aforementioned tough cases. Besides, it is worthy to investigate the modeling of occlusion problems in the single image focus editing work. We would also like to extend the basic idea of this work to solve other low level vision problems such as space-variant deblurring.

- As mentioned in Chapter 5, the proposed exposure method may not work well when the input exposures contain severe sensor noise or blurring artifacts caused by camera shake. Therefore, we hope to investigate how to integrate the denoising or deblurring steps into the current scheme well, and extend the proposed method to more scenarios (e.g. camera shaking, high-ISO noise). As stated in Section 5.3.6, the RCA would fail to deghost if an unsuitable exposure is selected as the reference image. Therefore, it is desirable to have a more advanced model for ghost removal to improve the current one-image-dependent reference view strategy. We also found that proper retouching such as color correction and sharpness adjustment can make the result more impressive, so it would be nice to add some retouching techniques to the current framework. Also, it is worthy to further investigate the potential of this work in other related tasks such as flash photography, relighting and color transfer.

- The efficiency of all algorithms could be improved by optimized GPU implementation such as [Hou et al., 2008].

Bibliography

[Aggarwal and Ahuja, 2004] Aggarwal, M. and Ahuja, N. (2004). Split aperture imaging for high dynamic range. *International Journal of Computer Vision*, 58:7 - 17.

[Agranov et al., 2007] Agranov, G., Mauritzson, R., S.Barna, Jiang, J., Dokoutchaev, A., Fan, X., and Li, X. (2007). Super small, sub 2μm pixels for novel cmos image sensors. In *International Image Sensor Workshop*, pages 307–310.

[Agrawal et al., 2005] Agrawal, A., Raskar, R., Nayar, S. K., and Li, Y. (2005). Removing photography artifacts using gradient projection and flash-exposure sampling. *ACM Transactions on Graphics (TOG)*, 24(3):828–835.

[Bae and Durand, 2007] Bae, S. and Durand, F. (2007). Defocus magnification. In *Proc. Eurographics*.

[Baker and Kanade, 2000] Baker, S. and Kanade, T. (2000). Hallucinating faces. In *Proc. International Conference on Automatic Face and Gesture Recognition (FG)*, pages 83–88.

[Baker and Kanade, 2002] Baker, S. and Kanade, T. (2002). Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 24(9):1167–1183.

[Bando and Nishita, 2007] Bando, Y. and Nishita, T. (2007). Towards digital refocusing from a single photograph. In *Proc. Pacific Graphics*.

[Ben-Ezra et al., 2004] Ben-Ezra, M., Zomet, A., and Nayar, S. (2004). Jitter camera: High resolution video from a low resolution detector. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 135–142.

[Ben-Ezra et al., 2005] Ben-Ezra, M., Zomet, A., and Nayar, S. (2005). Video super-resolution using controlled subpixel detector shifts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 27(6):977–987.

[Bishop et al., 2003] Bishop, C. M., Blake, A., and Marthi, B. (2003). Super-resolution enhancement of video. In *Proc. Artificial Intelligence and Statistics*.

[Black and Sapiro, 1998] Black, M. J. and Sapiro, G. (1998). Robust anisotropic diffusion. *IEEE Transactions on Image Processing (T-IP)*, 7(3):421–432.

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

[Brajovic and Kanade, 1996] Brajovic, V. and Kanade, T. (1996). A sorting image sensor: An example of massively parallel intensity-to-time processing for low-latency computational sensors. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 1638–1643.

[Brown and Lowe, 2003] Brown, M. and Lowe, D. G. (2003). Recognising panoramas. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 1218–1225.

[Burt and Adelson, 1983] Burt, P. J. and Adelson, E. H. (1983). A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics (TOG)*, 2(4):217–236.

[Canny, 1986] Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 8:679–698.

[Capel, 2004] Capel, D. P. (2004). *Image Mosaicing and Super-Resolution (Cphc/Bcs Distinguished Dissertations.)*. SpringerVerlag.

[Capel and Zisserman, 2001] Capel, D. P. and Zisserman, A. (2001). Super-resolution from multiple views using learnt image models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 627–634.

[Chan and Wong, 1998] Chan, T. F. and Wong, C. K. (1998). Total variation blind deconvolution. *IEEE Transactions on Image Processing (T-IP)*, 7:370–375.

[Chang et al., 2004] Chang, H., Yeung, D. Y., and Xiong, Y. (2004). Super-resolution through neighbor embedding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 275–282.

[Chang et al., 2006] Chang, T. L., Liu, T. L., and Chuang, J. H. (2006). Direct energy minimization for super-resolution on nonlinear manifolds. In *Proc. European conference on computer vision (ECCV)*, pages 281–294.

[Cheeseman et al., 1994] Cheeseman, P., Kanefsky, B., Kraft, R., and Stutz, J. (1994). Super-resolved surface reconstruction from multiple images. Technical report, NASA.

[Chen et al., 2001] Chen, H., Xu, Y. Q., Shum, H. Y., Zhu, S. C., and Zheng, N. N. (2001). Example-based facial sketch generation with non-parametric sampling. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 433–438.

[Choi et al., 2004] Choi, E., Choi, J., and Kang, M. G. (2004). Super-resolution approach to overcome physical limitations of imaging sensors: An overview. *International Journal of Imaging Systems and Technology*, 14(2):36–46.

[Dai et al., 2009] Dai, S., Han, M., Xu, W., Wu, Y., Gong, Y., and Katsaggelos, A. (2009). Softcuts: a soft edge smoothness prior for color image super-resolution. *IEEE Transactions on Image Processing (T-IP)*, 18:969–981.

[Debevec and Malik, 1997] Debevec, P. E. and Malik, J. (1997). Recovering high dynamic range radiance maps from photographs. In *Proc. SIGGRAPH*, pages 369–378.

[Drago et al., 2003] Drago, F., Myszkowski, K., Annen, T., and Chiba, N. (2003). Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum*, 22:419–426.

[Durand and Dorsey, 2002] Durand, F. and Dorsey, J. (2002). Fast bilateral filtering for the display of high-dynamic-range images. *Computer Graphics Forum*, 21:257–266.

[Eden et al., 2006] Eden, A., Uyttendaele, M., and Szeliski, R. (2006). Seamless image stitching of scenes with large motions and exposure differences. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2498–2505.

[Eisemann and Durand, 2004] Eisemann, E. and Durand, F. (2004). Flash photography enhancement via intrinsic relighting. In *Proc. SIGGRAPH*.

[Elad and Feuer, 1999] Elad, M. and Feuer, A. (1999). Super-resolution reconstruction of image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 21(9):817–834.

[Elkhatib and Salama, 2008a] Elkhatib, T. A. and Salama, K. N. (2008a). High resolution imaging through integrated nanoholes image sensor. In *Proc. IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 245–248.

[Elkhatib and Salama, 2008b] Elkhatib, T. A. and Salama, K. N. (2008b). Super-resolution: Imaging beyond the pixel size limit. In *Proc. IEEE Custom Integrated Circuits Conference (CICC)*, pages 515–518.

[Eren et al., 1997] Eren, P. E., Sezan, M. I., and Tekalp, A. M. (1997). Robust, object based high resolution image reconstruction from low resolution video. *IEEE Transactions on Image Processing (T-IP)*, 6(10):1446–1451.

[Fan and Cham, 2000] Fan, G. L. and Cham, W. K. (2000). Model-based edge reconstruction for low bit-rate wavelet-compressed images. *IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT)*, 10:120–132.

[Fattal, 2007] Fattal, R. (2007). Image upsampling via imposed edges statistics. In *Proc. SIGGRAPH*.

[Fattal et al., 2002] Fattal, R., Lischinski, D., and Werman, M. (2002). Gradient domain high dynamic range compression. In *Proc. SIGGRAPH*, pages 249–256.

[Favaro and Soatto, 2005] Favaro, P. and Soatto, S. (2005). A geometric approach to shape from defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 27:406–417.

[Fergus et al., 2006] Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T., and Freeman, W. T. (2006). Removing camera shake from a single photograph. In *Proc. SIGGRAPH*.

[Ferzli and Karam, 2009] Ferzli, R. and Karam, L. J. (2009). A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). *IEEE Transactions on Image Processing (T-IP)*, 18:717–728.

[Fife et al., 2007] Fife, K., Gamal, A. E., and Wong, H.-S. (2007). A 0.5μm pixel frame-transfer ccd image sensor in 110nm cmos. In *Proc. IEEE International Electron Devices Meeting*, pages 1003–1006.

[Freeman et al., 2002] Freeman, W. T., Jones, T. R., and Pasztor, E. C. (2002). Example-based super-resolution. *IEEE Computer Graphics and Applications*, 22(2):56–65.

[Freeman et al., 2000] Freeman, W. T., Pasztor, E. C., and Carmichael, O. T. (2000). Learning low-level vision. *International Journal of Computer Vision (IJCV)*, 40(1):25–27.

[Frey and Dueck, 2007] Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science*, 315:972–976.

[Gallo et al., 2009] Gallo, O., Gelfand, N., Chen, W., Tico, M., and Pulli, K. (2009). Artifact-free high dynamic range imaging. In *Proc. IEEE International Conference on Computational Photography (ICCP)*.

[Georgiev et al., 2007] Georgiev, T., Intwala, C., and Babacan, D. (2007). Light-field capture by multiplexing in the frequency domain. Technical report, Adobe Systems Incorporated.

[Girod, 1993] Girod, B. (1993). What's wrong with mean-squared error? *Digital images and human vision, MIT Press*, pages 207–220.

[Glasner et al., 2009] Glasner, D., Bagon, S., and Irani, M. (2009). Super-resolution from a single image. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

[Goshtasby, 2005] Goshtasby, A. A. (2005). Fusion of multi-exposure images. *Image and Vision Computing*, 23:611–618.

[Grosch, 2006] Grosch, T. (2006). Fast and robust high dynamic range image generation with camera and object movement. In *Proc. Vision, Modeling and Visualization (VMV)*, pages 277–284.

[Grossberg and Nayar, 2003] Grossberg, M. D. and Nayar, S. K. (2003). Determining the camera response from images: What is knowable? *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 25:1455–1467.

[Hardie et al., 1997] Hardie, R. C., Barnard, K. J., and Amstrong, E. E. (1997). Joint map registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing (T-IP)*, 6(12):1621–1633.

[Hasinoff, 2008] Hasinoff, S. W. (2008). *Variable-Aperture Photography*. PhD thesis, University of Toronto, Dept. of Computer Science.

[Hasinoff and Kutulakos, 2007] Hasinoff, S. W. and Kutulakos, K. N. (2007). A layer-based restoration framework for variable-aperture photography. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.

[Hasinoff and Kutulakos, 2008] Hasinoff, S. W. and Kutulakos, K. N. (2008). Light-efficient photography. In *Proc. European conference on computer vision (ECCV)*.

[He and Kondi, 2005] He, H. and Kondi, L. P. (2005). A regularization framework for joint blur estimation and super-resolution of video sequences. In *Proc. IEEE International Conference on Image Processing (ICIP)*.

[Hou et al., 2008] Hou, Q., Zhou, K., and Guo, B. (2008). Bsgp: bulk-synchronous gpu programming. In *Proc. SIGGRAPH*.

[Ikeda, 1998] Ikeda, E. (1998). Image data processing apparatus for processing combined image signals in order to extend dynamic range. *U.S Patent 5801773*.

[Irani and Peleg, 1991] Irani, M. and Peleg, S. (1991). Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 53(3):231–239.

[Irani and Peleg, 1993] Irani, M. and Peleg, S. (1993). Motion analysis for image enhancement: Resolution, occlusion, and transparency,. *Journal of Visual Communications and Image Representation (VCIR)*, 4(4):324C335.

[Jacobs et al., 2008] Jacobs, K., Loscos, C., and Ward, G. (2008). Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications*, 28:84–93.

[Jia, 2007] Jia, J. Y. (2007). Single image motion deblurring using transparency. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Jia and Gong, 2008] Jia, K. and Gong, S. (2008). Generalized face super-resolution. *IEEE Transactions on Image Processing (T-IP)*, 17:873–886.

[Joshi et al., 2004] Joshi, M., Chaudhuri, S., and Rajkiran, P. (2004). Super-resolution imaging: Use of zoom as a cue. *Image and Vision Computing*, 22(14):1185–1196.

[Joshi et al., 2008] Joshi, N., Szeliski, R., and Kriegman, D. J. (2008). Psf estimation using sharp edge prediction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Joshi et al., 2009] Joshi, N., Zitnick, L., Szeliski, R., and Kriegman, D. J. (2009). Image deblurring and denoising using color priors. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Kang et al., 2003] Kang, S. B., Uyttendaele, M., Winder, S., and Szeliski, R. (2003). High dynamic range video. In *Proc. SIGGRAPH*, pages 319–325.

[Khan et al., 2006] Khan, E. A., Akyuz, A. O., and Reinhard, E. (2006). Ghost removal in high dynamic range images. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 2005–2008.

[Kimura, 1998] Kimura, T. (1998). Image pickup device. *Japanese Patent 10-069011*.

[Komatsu et al., 1993] Komatsu, T., Aizawa, K., Igarashi, T., and Saito, T. (1993). Signal-processing based method for acquiring very high resolution image with multiple cameras and its theoretical analysis. *Proc. Inst. Elec. Eng.,*, 140(1):19–25.

[Kong et al., 2006] Kong, D., Han, M., Xu, W., Tao, H., and Gong, Y. (2006). Video super-resolution with scene-specific priors. In *Proc. British Machine Vision Conference.*

[Kubota and Aizawa, 2005] Kubota, A. and Aizawa, K. (2005). Reconstructing arbitrarily focused images from two differently focused images using linear filters. *IEEE Transactions on Image Processing (T-IP)*, 14:1848–1859.

[Kubota et al., 2004] Kubota, A., Aizawa, K., and Chen, T. (2004). Reconstructing dense light field from a multi-focus images array. In *Proc. IEEE Conference on Multimedia and Expo (ICME)*, pages 2183–2186.

[Kutulakos and Hasinoff, 2009] Kutulakos, K. N. and Hasinoff, S. W. (2009). Focal stack photography: High-performance photography with a conventional camera. In *Proc. IAPR Conference on Machine Vision Applications (MVA)*, pages 332–337.

[Levin et al., 2007] Levin, A., Durand, R. F. F., and Freeman, B. (2007). Image and depth from a conventional camera with a coded aperture. In *Proc. SIGGRAPH.*

[Levin et al., 2004] Levin, A., Lischinski, D., and Weiss, Y. (2004). Colorization using optimization. In *Proc. SIGGRAPH.*

[Levin and Weiss, 2007] Levin, A. and Weiss, Y. (2007). User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 29(9):1647–1654.

[Levin et al., 2009] Levin, A., Weiss, Y., Durand, F., and Freeman, W. T. (2009). Understanding and evaluating blind deconvolution algorithms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[Li and Orchard, 2001] Li, X. and Orchard, M. (2001). New edge-directed interpolation. *IEEE Transactions on Image Processing (T-IP)*, 10(10):1521–1527.

[Li and Lin, 2004] Li, Y. and Lin, X. (2004). Face hallucination with pose variation. In *Proc. International Conference on Automatic Face and Gesture Recognition (FG)*, pages 723–728.

[Li et al., 2005] Li, Y., Sharan, L., and Adelson, E. H. (2005). Compressing and companding high dynamic range images with subband architectures. In *Proc. SIGGRAPH*, pages 836–844.

[Liang et al., 2008] Liang, C. K., Lin, T. H., Wong, B.-Y., Liu, C., and Chen, H. H. (2008). Programmable aperture photography: multiplexed light field acquisition. In *Proc. SIGGRAPH.*

[Lin and Chang, 2006] Lin, H. Y. and Chang, C. H. (2006). Depth from motion and defocus blur. *Optical Engineering*, 45(12).

[Lin et al., 2008] Lin, Z., He, J., Tang, X., and Tang, C.-K. (2008). Limits of learning-based superresolution algorithms. *International Journal of Computer Vision (IJCV)*, 80(3):406–420.

[Lin and Shum, 2004] Lin, Z. and Shum, H.-Y. (2004). Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 26(1):83 – 97.

[Liu et al., 2007] Liu, C., Shum, H. Y., and Freeman, W. T. (2007). Face hallucination: theory and practice. *International Journal of Computer Vision (IJCV)*, 75(1):115–134.

[Liu et al., 2005] Liu, W., Lin, D., and Tang, X. (2005). Hallucinating faces: Tensorpatch super-resolution and coupled residue compensation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 478–484.

[Ma et al., 2008] Ma, L., Zhang, Y., Lu, Y., Wu, F., and Zhao, D. (2008). Three-tiered network model for image hallucination. In *Proc. IEEE International Conference on Image Processing (ICIP).*

[Mertens et al., 2009] Mertens, T., Kautz, J., and Reeth, F. V. (2009). Exposure fusion: A simple and practical alternative to high dynamic range photography. *Computer Graphics Forum*, 28:161–171.

[Moreno-Noguer et al., 2007] Moreno-Noguer, F., Belhumeur, P. N., and Nayar, S. K. (2007). Active refocusing of images and videos. In *Proc. SIGGRAPH*.

[Morse and Schwartzwald, 2001] Morse, B. S. and Schwartzwald, D. (2001). Image magnification using level-set reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 333–340.

[Murakoshi, 1994] Murakoshi, M. (1994). Charge coupling image pickup device. *Japanese Patent 59-217358*.

[Muresan and Parks, 2002] Muresan, D. D. and Parks, T. W. (2002). Optimal face reconstruction using training. In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 373–376.

[Nagahara et al., 2008] Nagahara, H., Kuthirummal, S., Zhou, C., and Nayar, S. (2008). Flexible depth of field photography. In *Proc. European conference on computer vision (ECCV)*.

[Nayar and Branzoi, 2003] Nayar, S. K. and Branzoi, V. (2003). Adaptive dynamic range imaging: Optical control of pixel exposures over space and time. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1168–1175.

[Nayar and Mitsunaga, 2000] Nayar, S. K. and Mitsunaga, T. (2000). High dynamic range imaging: Spatially varying pixel exposures. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 472–479.

[Ng et al., 2005] Ng, R., Levoy, M., Bredif, M., Duval, G., Horowitz, M., and Hanrahan, P. (2005). Light field photography with a hand-held plenoptic camera. Technical report, Stanford University Computer Science.

[Ni and Nguyen, 2007] Ni, K. S. and Nguyen, T. Q. (2007). Image superresolution using support vector regression. *IEEE Transactions on Image Processing (T-IP)*, 16(6):1596–1610.

[Paris and Durand, 2006] Paris, S. and Durand, F. (2006). A fast approximation of the bilateral filter using a signal processing approach. In *Proc. European conference on computer vision (ECCV)*.

[Park et al., 2004] Park, S. C., Kang, M. G., Segall, C. A., and Katsaggelos, A. K. (2004). Spatially adaptive high-resolution image reconstruction of dct-based compressed images. *IEEE Transactions on Image Processing (T-IP)*, 13(4):573–585.

[Park et al., 2003] Park, S. C., Park, M. K., and Kang., M. (2003). Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21–36.

[Patti et al., 1997] Patti, A., Sezan, M. I., and Tekalp, A. M. (1997). Super resolution video reconstruction with arbitrary sampling lattices and non-zero aperture time. *IEEE Transactions on Image Processing (T-IP)*, 6:1064–1076.

[Patti and Altunbasak, 1999] Patti, A. J. and Altunbasak, Y. (1999). Super-resolution image estimation for transform coded video with application to mpeg. In *Proc. IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 179–183.

[Patti and Altunbasak, 2001] Patti, A. J. and Altunbasak, Y. (2001). Artifact reduction for set theoretic super resolution image reconstruction with edge adaptive constraints and higher-order interpolants. *IEEE Transactions on Image Processing (T-IP)*, 10(1):179–186.

[Pedone and Heikkilä, 2008] Pedone, M. and Heikkilä, J. (2008). Constrain propagation for ghost removal in high dynamic range images. In *Proc. Third International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 36–41.

[Pérez et al., 2003] Pérez, P., Gangnet, M., and Blake, A. (2003). Poisson image editing. *ACM Transactions on Graphics (TOG)*, 22(3):313–318.

[Petschnigg et al., 2004] Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., and Toyama, K. (2004). Digital photography with flash and no-flash image pairs. In *Proc. SIGGRAPH*.

[Pham et al., 2006] Pham, T. Q., van Vliet, L. J., and Schutte, K. (2006). Resolution enhancement of low quality videos using a high-resolution frame. In *Proc. Visual Communications and Image Processing (VCIP)*.

[Phillips et al., 2000] Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The feret evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 22:1090–1104.

[Phillips et al., 1998] Phillips, P. J., Wechsler, H., Huang, J., and Rauss, P. (1998). The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295–306.

[Protter and Elad, 2009] Protter, M. and Elad, M. (2009). Super-resolution with probabilistic motion estimation. *IEEE Transactions on Image Processing (T-IP)*, 18(8):1899–1904.

[Protter et al., 2009] Protter, M., Elad, M., Takeda, H., and Milanfar, P. (2009). Generalizing the non-local-means to super-resolution reconstruction. *IEEE Transactions on Image Processing (T-IP)*, 18(1):36–51.

[Rajagopalan et al., 2004] Rajagopalan, A. N., Chaudhuri, S., and Mudenagudi, U. (2004). Depth estimation and image restoration using defocused stereo pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 26:1521–1525.

[Reinhard et al., 2002] Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J. (2002). Photographic tone reproduction for digital images. In *Proc. SIGGRAPH*, pages 267–276.

[Reinhard et al., 2005] Reinhard, E., Ward, G., Pattanaik, S., and Debevec, P. (Dec. 2005). *High Dynamic Range Imaging: Acquisition, Display and Image-Based Lighting*. Morgan Kaufmann Publishers.

[Roth and Black, 2009] Roth, S. and Black, M. J. (2009). Fields of experts. *International Journal of Computer Vision (IJCV)*, 82(2):205–229.

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

[Rudin et al., 1992] Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259C268.

[Saito, 1996] Saito, K. (1996). Electronic image pickup device. *Japanese Patent 08-340486*.

[Schechner and Kiryati, 2000] Schechner, Y. Y. and Kiryati, N. (2000). Depth from defocus vs. stereo: how different really are they? *International Journal of Computer Vision (IJCV)*, 39:141–162.

[Schultz and Stevenson, 1996] Schultz, R. and Stevenson, R. (1996). Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing (T-IP)*, 5(6):996–1011.

[Shan et al., 2010] Shan, Q., Jia, J., and Brown, M. S. (2010). Globally optimized linear windowed tone mapping. *IEEE Transactions on Visualization and Computer Graphics (T-VCG)*, 16.

[Shan et al., 2008a] Shan, Q., Jia, J. Y., and Agarwala, A. (2008a). High-quality motion deblurring from a single image. In *Proc. SIGGRAPH*.

[Shan et al., 2008b] Shan, Q., Li, Z., Jia, J., and C.Tang (2008b). Fast image/video upsampling. In *Proc. SIGGRAPH ASIA*.

[Shanmuganathan and Chaudhuri, 2009] Shanmuganathan, R. and Chaudhuri, S. (2009). Bilateral filter based compositing for variable exposure photographs. In *Proc. Eurographics*.

[Sidibe et al., 2009] Sidibe, D. D., Puech, W., and Strauss, O. (2009). Ghost detection and removal in high dynamic range images. In *Proc. European Signal Processing Conference (EUSIPCO)*.

[Street, 1998] Street, R. A. (1998). High dynamic range segmented pixel sensor array. *U.S Patent 5789737*.

[Subbarao et al., 1995] Subbarao, M., Wei, T.-C., and Surya, G. (1995). Focused image recovery from two defocused images recorded with different camera settings. *IEEE Transactions on Image Processing (T-IP)*, 4(12):1613–1628.

[Sun et al., 2008] Sun, J., Sun, J., Xu, Z., and Shum., H.-Y. (2008). Image super-resolution using gradient profile prior. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Sun et al., 2003] Sun, J., Tao, H., and Shum, H. Y. (2003). Image hallucination with primal sketch priors. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 729–736.

[Takeda et al., 2009] Takeda, H., Milanfar, P., Protter, M., and Elad, M. (2009). Super-resolution without explicit subpixel motion estimation. *IEEE Transactions on Image Processing (T-IP)*, 18(9):1958–1975.

[Tappen et al., 2003] Tappen, M. F., Russell, B. C., and Freeman, W. T. (2003). Exploiting the sparse derivative prior for super-resolution and image demosaicing. In *Proc. IEEE Workshop on Statistical and Computational Theories of Vision*.

[Tran et al., 2003] Tran, T. D., Liang, J., and Tu, C. (2003). Lapped transform via time-domain pre- and post-filtering. *IEEE Transactions on Signal Processing (T-SP)*, 51(6):1557–1571.

[Tu and Tran, 2002] Tu, C. and Tran, T. D. (2002). Context based entropy coding of block transform coefficients for image coding. *IEEE Transactions on Image Processing (T-IP)*, 11:1277–1283.

[Tumblin et al., 2005] Tumblin, J., Agrawal, A., and Raskar, R. (2005). Why i want a gradient camera. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

[van Beek, 1995] van Beek, P. (1995). *Edge-based image representation and coding*. PhD thesis, Delft University of Technology.

[van den Berg and Friedlander, 2007] van den Berg, E. and Friedlander, M. P. (2007). Spgl1: A solver for large-scale sparse reconstruction. In *Available: http://www.cs.ubc.ca/labs/scl/spgl1*.

[van den Berg and Friedlander, 2008] van den Berg, E. and Friedlander, M. P. (2008). Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912.

[Veeraraghavan et al., 2007] Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A., and Tumblin, J. (2007). Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *Proc. SIGGRAPH*.

[Wang et al., 2005] Wang, Q., Tang, X., and Shum, H. Y. (2005). Patch based blind image super resolution. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 709–716.

[Wang et al., 2002a] Wang, Y., Ostermann, J., and Zhang, Y. (2002a). *Video Processing and Communications*. Prentice Hall.

[Wang et al., 2007] Wang, Y., Yang, J., Yin, W., and Zhang, Y. (2007). A new alternating minimization algorithm for total variation image reconstruction. Technical report, Rice University CAAM.

[Wang et al., 2002b] Wang, Z., Bovik, A. C., and Lu, L. (2002b). Why is image quality assessment so difficult? In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 3313–3316.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (T-IP)*, 13(4):600–612.

[Ward, 2003] Ward, G. (2003). Fast, robust image registration for compositing high dynamic range photographs from handheld exposures. *Journal of Graphics Tools*, 8:17–30.

[Wen, 1989] Wen, D. D. (1989). High dynamic range charge-coupled device. *U.S Patent 4873561*.

[Xiong et al., 2009] Xiong, Z., Sun, X., and Wu, F. (2009). Image hallucination with feature enhancement. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Yan et al., 2009] Yan, C.-Y., Tien, M.-C., and Wu, J.-L. (2009). Interactive background blurring. In *Proc. ACM International Conference on Multimedia (MM)*, pages 817–820.

[Yang et al., 2008a] Yang, H., Gao, J., and Wu, Z. (2008a). Blur identification and image super-resolution reconstruction using an approach similar to variable projection. *IEEE Signal Processing Letters (SPL)*, 15:289–292.

[Yang and Schonfeld, 2010] Yang, J. and Schonfeld, D. (2010). Virtual focus and depth estimation from defocused video sequences. *IEEE Transactions on Image Processing (T-IP)*, 19:668–679.

[Yang et al., 2008b] Yang, J., Schonfeld, D., and Mohamed, M. (2008b). Focused video estimation from defocused video sequences. In *SPIE Proc. of Electronic Imaging: Science and Technology, Conference on Visual Communications and Image Processing (VCIP)*.

[Zhang and Cham, a] Zhang, W. and Cham, W.-K. Gradient-directed multi-exposure composition. *submitted to IEEE Transactions on Image Processing (T-IP). (In review)*.

[Zhang and Cham, b] Zhang, W. and Cham, W.-K. Hallucinating face in the dct domain. *IEEE Transactions on Image Processing (T-IP) (Accepted with minor revision)*.

[Zhang and Cham, c] Zhang, W. and Cham, W.-K. Single image refocusing and defocusing. *submitted to IEEE Transactions on Image Processing (T-IP). (In review)*.

[Zhang and Cham, 2008] Zhang, W. and Cham, W.-K. (2008). Learning-based face hallucination in dct domain. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zhang and Cham, 2009] Zhang, W. and Cham, W.-K. (2009). Single image focus editing. In *Proc. IEEE International Conference on Computer Vision workshop on Color and Reflectance in Imaging and Computer Vision (ICCV-CRICV)*.

[Zhang and Cham, 2010a] Zhang, W. and Cham, W.-K. (2010a). Gradient-directed composition of multi-exposure images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zhang and Cham, 2010b] Zhang, W. and Cham, W.-K. (2010b). High quality artifact-free super-resolution. In *Proc. IEEE International Conference on Image Processing (ICIP)*.

[Zhang and Wu, 2008] Zhang, X. and Wu, X. (2008). Image interpolation by adaptive 2d autoregressive modeling and soft-decision estimation. *IEEE Transactions on Image Processing (T-IP)*, 17:887–896.

[Zomet et al., 2001] Zomet, A., Rav-Acha, A., and Peleg, S. (2001). Robust super-resolution. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 645–650.