SINGLE IMAGE DEBLURRING WITH OR WITHOUT FACE PRIOR AND ITS APPLICATIONS

BY LIN ZHONG

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Dimitris Metaxas

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2015

ABSTRACT OF THE DISSERTATION

Single Image deblurring with or without face prior and its applications

by Lin Zhong

Dissertation Director: Dimitris Metaxas

The motion blur is one of the most difficult challenges in photography, which is generated from the relative motion between the sensor and the scene during exposure time. These blur artifacts degrade the visual experience, and the performance of various applications, such as, object detection, facial analysis. Therefore, it is significant to remove the blur and restore sharp and clean images. Our work focuses on the general single image deblurring, and face image deblurring with face prior.

State-of-the-art single image deblurring techniques are sensitive to image noise. Even a small amount of noise, which is inevitable in low-light conditions, can degrade the quality of blur kernel estimation dramatically. We propose a new method for handling noise in blind image deconvolution based on new theoretical and practical insights. Based on the observations on directional filter, our method applies a series of directional filters at different orientations to the input image, and estimates an accurate Radon transform of the blur kernel from each filtered image. Finally, we reconstruct the blur kernel using inverse Radon transform. Experimental results on synthetic and real data show that our algorithm achieves higher quality results than previous approaches on blurry and noisy images.

The human face is one of the most essential focuses in numerous applications. Although

significant progress has been made in the image deblurring area, few of them can obtain promising results on blurry face images. Many state-of-the-art single image deblurring approaches estimate the blur kernel based on analyzing the edge profiles of the input image. However, the detection of strong edges is very difficult on human faces, since the human faces do not contain as much texture as natural images. We propose to utilize the global face structure information to help with the strong or salient edge detection. Our method outperforms the existing methods in extensive evaluations on synthetic and real face images.

Facial expression is a significant application on sharp and restored face images. To improve the general facial expression recognition performance, we present a new idea to analyze facial expression by exploring the common and specific information among different expressions.

Acknowledgements

I would like to express my supreme gratitude to my advisor, Professor Dimitris N. Metaxas, for his encouragement, guidance and great support in the past five years. It is Prof. Metaxas, who directed me into the fields of computer vision and machine learning, and enabled me to develop an understanding of these areas. His insightful guidance and suggestions always directed me towards the interesting and challenging topics in these fields, yet still gave me great freedom to pursue independent work. Without his guidance and persistent help, my achievements during the Ph.D. period and this dissertation would not have been possible.

I would like to thank other committee members: Prof. Ahmed Elgammal, Prof. Kostas Bekris, and Prof. Dimitris Samaras (Stony Brook University) for their advice, help and valuable suggestions regarding this thesis. It is an honor for me to have each of them serves in my committee.

I would also like to thank Dr. Jue Wang (Adobe Research), Dr. Sunghyun Cho (Samsung Electronics). They were my mentors when I conducted internship at Adobe Creative Research lab Seattle in 2012 summer. Both of them spent a lot of time and effort on helping me solving different problems and difficulties. They also offered me a lot of assistance in the paper revision and the patent application.

I also thank many professors and researchers who have helped me in many aspects, especially to Dr. Qingshan Liu (Nanjing University of Information Science and Technology), Dr. Peng Yang (Intelligence Automation Inc), Prof. Junzhou Huang (The University of Texas at Arlington), Dr. Minwoo Park (Object Video), Dr. Sen Wang (Qualcomm), and Dr. Chao Chen.

Finally, I would like to thanks all my friends and colleagues in CBIM. I benefited a lot from their friendship and help, which made my years at Rutgers pleasurable, fruitful and memorable.

Dedication

This dissertation is dedicated to my parents with whom I share my success and frustration. Without their endless encouragement, love and support, I can hardly finish my Ph.D. degree and achieve my goals.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the development of hand-held cameras and smartphones, people take more and more photos than before. However, the conditions for the photography may not be ideal in our daily lives, such as, indoors, low light environment. In these cases, amateur photographers often get blurry and noisy photos. Usually, retaking the image with the same content is impossible. Besides daily life photos, blur also happens to satellite imaging, video recording and so on. It is really important to develop an efficient, robust and fast deblurring method to restore a sharp image with an blurry input. Although many methods have been proposed to solve this problem, they can not restore promising latent images in many real cases. This thesis tries to contribute to the single image deblurring area, and the proposed methods show their superiority over the state-of-the-art methods in many aspects.

1.1 Background

The single image deblurring or blind deconvolution problem has attracted much attention since early 1960s. However, it is very difficult to develop an efficient and robust method for various blurry images since little or no general prior can be applied to diverse images. Thus, until now, it is still an open question for the image processing and computer vision community.

Generally speaking, the image blur degradation is generated from the imperfection during the capturing and imaging process. The degradation is quite complicated in real cases and could result from different sources, such as out-of-focus blur, low-pass filter, motion blur, moving objects or combination of them. Some sample blurry images are shown in Fig. 1.1. The underline reasons incurring different types of blur are different, so various methods are proposed to solve different types of blur respectively.

Out-of-focus blur happens when the desired object is not in the proper distance when

(a) out-of-focus blur (b) low-pass filter (c) motion blur

(d) moving object (e) combination

Figure 1.1: Some examples with different types of blur.

capturing. When the 3-D scene with objects of different depth is projected to a 2-D image plane, some objects would be out-of-focus, and produce blur artifacts. The proper distance is defined by the lens focal length and the aperture of the camera. Only the objects with the proper distance can be captured sharply. Since the degree of defocusing is largely related to the depth of the objects from the camera and the wavelengths, the blur can be modeled by geometry properties. Grossmann [39], Darrell and Wohn [28] tried to extract the depth information of the scene by measuring the degree of the blur. More recently, Shen et al. [76] derived a blur map using local contrast prior and the guided filter, and then utilized the $L1-2$ norm prior to obtain a better restored image.

Low-pass filter blur is the result of filtering an image with a low pass filter, such as a 2-D Gaussian function. This type of blur normally occurs in a number of situations, such as the atmospheric turbulence blurs the astronomical images, low resolution images. Since these blur in practice can be approximated by a Gaussian blur, Hummel et al. [43] proposed an approach to deblur the Gaussian blur by restricting the band-limited functions. More generally, the Gaussian blur or low-resolution problem are considered as super-resolution problems. Suppose the restored image is natural and can be composed by a combination of small patches learned from

natural images or selected examples, Yang et al. [103] provided a promising method to produce super-resolution images.

Moving object blur normally occurs in dynamic scenes with one or multiple moving objects relative to the static scene. The moving direction and speed in the 3D scene are difficult to estimate from a 2D blurry image, and thus dynamic scene deblurring is deeply challenging. In [54], an adaptive segmentation method is utilized to separate the moving object from the static scene, and the blur kernels are modeled independently with the corresponding data term and regularization term.

Motion blur, which generates from the relative motion between the recording sensor and the static scene during the exposure time, is one of the major sources of the blur. Motion blur normally happens to hand-held camera in low-light condition when the exposure time is long. The blurry image is normally modeled as the convolution result of the latent sharp image and the motion blur, which can be considered as the trajectory of the recording sensor. The deblurring methods are generally categorized into two types : blind deconvolution and non-blind deconvolution methods. Blind deconvolution methods need to recover both the blur kernel and the latent image from the single input blurry image [19, 56, 58, 75]. More literature reivew can be found in the Relevant work Sec 2.1.

On the other hand, for non-blind deconvoluiton methods, the blur kernels are given, and only the latent images are needed to be recovered. So they could be considered as one component of the blind deconvolution method. Since the convolution is a linear operator, classical linear image restoration is firstly used to solve this problem [50, 67]. Later, more priors are employed to generate more promising latent images, such as, the local salient edge prior [75], models of natural image patches [120], generic vs specific priors [77]. Cho et al. [20] also proposed a method to handle the outliers i.e., saturated pixels and non-Gaussian noise in convolution.

In the deblurring areas, there are many research topics regarding to different kinds of blurs as mentioned above. Since the motion blur is a commonly seen degradation, and happens to most handlheld cameras, our thesis will focus on single image deblurring, which recovers the latent image and blur kernel from a single input blurry image.

1.2 Problem Statement

Low-light photography is very common in our daily life. In such situations, the camera would slow its shutter speed and increase the exposure time. The motion during the exposure time is unavoidable to hand-held cameras. Even some cameras are equipped with vibration reduction function, it is still very common for amateur photographers to get blurry images, especially in low-light conditions. The blur results from the movement of the camera or the capturing sensor, then the blurry image can be approximated as the convolution result of the latent sharp image and the blur kernel. Moreover, the blurry image also contains a lot of noise because of the long exposure time and high ISO setting. With slower shutter speed, the amount of salt-and-pepper noise will increase because of the photo-diode leakage currents. High ISO setting will amplify the take-in signal including noise, and make the imaging sensor more sensitive to noise. Thus, we normally get more blurry and noisy images in low-light conditions than ideal light environments.

The blur and noise of the images degrade the visual experience greatly, and the photo retake is impossible in most of the cases. Thus, restoring the latent sharp images from blurry and noisy images is of great significance. Many approaches have been proposed to solve the deblurring problem, but the noise effects are largely ignored. Unfortunately, the noise would be amplified in solving the convolution linear system, and thus most of the existing deblurring methods would fail when the noise occurs in real cases. One of the major problems we would like to solve in this thesis is how to handle the noise in single image deblurring.

Human faces are the most important and attractive areas in the photos. The deblurring and refinement of the face is even more important than other areas in an image. Most of the existing deblurring methods focus on the natural images without considering the properties of human faces. Since human face does not follow the priors learned from natural images, and also does not contains much salient edges, most existing methods fail in recovering high quality face images from blurry images. On the other hand, the face structure information has been largely utilized in many other applications, such as face detection, facial landmark localization. Another major goal for our thesis is exploring good ways to employ global face structure information in face image restoration and refinement.

The restoration of sharp face images from blurry images would be beneficial to many face related applications, such as face identification, facial expression analysis. These applications would get better performance on sharp face images than blurry face images. The problem of how to boost the facial expression recognition performance is also discussed in our thesis.

1.3 Main Contributions

The main contributions of this thesis are summarized as follows:

1) Since most state-of-the-art single image deblurring methods are sensitive to noise, and noise is also unavoidable in low-light photography, we proposed a new method to handle noise in blind image deconvolution based on new theoretical and practical insights. We found that applying a directional low-pass filter to the input blurry and noisy image can greatly reduce the noise level, while preserving the blur information in the orthogonal direction to the filter. Based on this observation, we proposed a noise-aware kernel estimation method by applying a series of directional filters in different directions. After the kernel is estimated, we introduced a final noise-aware nonblind deconvolution method to restore a sharp and clean image from the input blurry image.

2) The performances of most single image deblurring methods degrade greatly on face images, since the texture on face is limited, and salient edge detection is very hard on blurry face image. The face landmark localization can implicitly help the salient edge detection since its model contains the information of the face structure. We proposed a face image deblurring method based on the face landmark detection. Extensive experiments show the effectiveness of the proposed face deblurring method. The face deblurring algorithm can be used as the preprocessing before various facial analysis, and thus boost the recognition performance.

3) In the facial expression recognition area, we provided a solid validation for an important psychology discovery, that only partial area of the face (corresponding to underlying facial muscles) are discriminative for expression recognition. A two-stage multi-task sparse learning framework is proposed to formulate the commonalities among expressions, and

to find out the locations of common and specific patches for expressions. Multi-scale image division strategy is utilized to generate patches of different size for facial expression analysis. More convincing conclusion about facial parts (muscles) could be achieved, since they are of different sizes. The common and specific patches can be combined to improve the performances of state-of-the-arts. Patches across different scales can also been fused to further boost the performance.

1.4 Organization

The remainder of this thesis is organized as follows.

Chapter 2 reviews the relevant work in single image deblurring, and further image deblurring with face prior. Besides reviewing the existing deblurring methods, we also analyze their advantages and limitations. The literature of the facial expression recognition is also reviewed, since it is an important application on the restored face image.

Chapter 3 introduces the proposed single image deblurring method based on the directional filter in detail. It includes the introduction and proof of the directional filter, which can greatly reduce the noise, while keeping the blur information intact in the orthogonal direction. The details of the noise-aware kernel estimation and final nonblind deconvolution are also illustrated in this chapter.

The method proposed in the previous chapter can not handle face image well, since salient edges are difficult to detect on face images. Chapter 4 introduces how to employ face landmark detection (containing the face structure prior) to help the face image deblurring. The robustness of landmark detection on blurry image is first illustrated. The details of how to generate the initial blur kernel based on the landmark detection results and the later iterative algorithm are included.

In chapter 5, the facial expression recognition on face image is introduced. More specifically, it is a multi-scale active patch learning algorithm based on multi-task sparse learning. This chapter introduces how to construct multiply tasks of expression recognition, and how to employ multi-task sparse learning to explore the common and specific patches, which can further improve the recognition performance.

Finally, chapter 6 summarizes the work and contributions of this thesis, along with more discussions about the limitations and the future work.

Chapter 2

Relevant Work

2.1 Single Image Deblurring

Single image deblurring, or blind deconvolution, has been studied for decades. In this problem a blurry image b is modeled as:

$$b = l * k + n, \tag{2.1}$$

where l is the latent sharp image, k is the blur kernel, and n is noise. Estimating k and l from a single observation b is a severely ill-posed problem, requiring additional assumptions or priors on k and l to make the the estimation possible.

In the last '90s, Yitzhaky et al. [107] proposed to use parametric blur kernels to model the k. The directional blur kernels are defined by angle and length, which only have one dimension. Rav-Acha [66] also modeled the motion blur in a similar parametric way. However, the shapes of the real motion blurs are quite complex and more complicated than simple parametric models. The parametric blur kernel would be a too restrictive assumption for real cases, and thus these methods fail very often and can not get high-quality results.

More recently, Some previous methods took the $MAP_{k,l}$ approach to jointly estimate k and l [34, 75]. To better constrain the problem, these approaches force the estimated l to satisfy the natural image prior, i.e., the gradient magnitudes of l follows a heavy-tailed distribution. Although these methods can work well in some cases, they suffer from the well-known MAP failure [57] that often leads to non-satisfactory results in practice. The high computational cost of these methods further hinders their practical usage.

On the other hand, the MAP_k framework has recently emerged as a practical deblurring solution for handling real world data. Representative approaches are developed by Cho and Lee [19], Xu *et al.* [100], Cho *et al.* [21], Joshi *et al.* [47], Lin *et al.* [117], and Jia

[45]. These approaches first explore edges and estimate k only, which is a better-constrained problem than $MAP_{k,l}$, and then apply a non-blind deconvolution method to recover l given the estimated k. A common property of these methods is that they all extract and rely on some "good" image edges, rather than the entire image, to estimate k. Thus, how to select proper edges to use becomes a critical issue for these edge-based kernel estimation methods. Previous approaches have shown the importance of the selected edges, which should be strong and well-separated from nearby image structures, and then the blur information can be extracted reliably from them. For instance, Cho and Lee [19] and Lin *et al.* [117] identify edges with highest gradient magnitudes; Xu *et al.* [100] utilizes the usefulness map to eliminate small edges; Cho *et al.* [21] and Joshi *et al.* [47] explicitly detect step or sharp edges on blurry image using a set of heuristic rules. Hu *et al.* [41] tries to explore good regions for deblurring.

2.2 Face Image Deblurring

The faces are the most informative and important areas on face images. However, we often get blurry face images in low-light photography. The restoration of the blurry faces is even more important. Unfortunately, the previous methods do not perform well on face images mainly for two reasons: first, the face images are quite different from natural images, and the prior learned on natural images can not be applied directly. Second, the human faces do not contain much salient edges. Most of the edges are soft edges and hard to detect. There are only quite few deblurring approaches are proposed to handle face blurry image specifically. Pan et al. [63] tries to find exemplar sharp image which is similar to the blurry image, and transfer the edge gradients in the exemplar images to deblur the blurry face image. The main problem for this method is the exemplar database can not be completed enough to match arbitrary faces with various poses, expressions, and shapes.

2.2.1 Facial Landmark Localization

Facial landmark localization has been studied for many years in computer vision area, which would be beneficial to many applications, including face recognition, facial expression analysis, video editing. The methods to locate the facial landmarks can be generally categorized into two

categories: parametric template based methods and regression based methods.

Early successes in facial landmark localization are achieved by the well-known framework of the Active Shape Model (ASM) [8] and the Active Appearance Model (AAM) [25]. These methods fit a generative model of the global facial appearance to a given image by optimizing over the template's parameter space. So these methods are effective for many cases, and robust to local corruptions. However, these methods need expensive iterative steps to get the optimal solution, and also very easy to break down on extreme poses, and expressions, due to the limitations of the model flexibility.

Recently, new regression based methods [12, 26, 99] have been proposed. These methods consider landmark localization as a regression task directly and and a holistic regressor is used to compute the landmark coordinates from raw input pixels. These methods overcome some disadvantages of parametric based methods because of their greater flexibility and effective sub-pixel localization capability. These methods are also more efficient since no iterative fitting step is required. Powerful deep convolutional neural networks (DCNN) have been successfully utilized in the regression framework [105, 119]. These methods achieve two-fold advantages: 1) geometric constraints among facial points are implicitly utilized; 2) huge amount of training data can be leveraged, and thus the state-of-the-art performance.

2.3 Facial Expression Analysis

Most facial expression analysis methods generally follow the aforementioned two categories: *AU-based* and *message and sign judgement* methods. Although our method belongs to the latter category, it is still necessary to give a completed review on related works on expression analysis.

AU-based facial expression analysis inspired by the well-known study on facial activity, *Facial Action Coding System (FACS)* [31]. In this system, the subtle changes in facial appearance are encoded into 32 action units (AUs) with individual linguistic description. Since each basic expressions can be decomposed into several related AUs, the expression recognition problem can be transferred to AUs detection problem. Bartlett *et al.* [5] recognized six single upper face AUs, but no simultaneous AUs are considered in combination. Tian *et al.* [81] detected

16 AUs from face image sequences using lip tracking, template matching and neural networks. More works have been done on spontaneous facial expression data by automatic recognition of AUs [6, 7, 22, 23, 46, 91]. Some differences between spontaneous and deliberate facial behavior are also studied by [24]. Recently, AU detection and AU-based expression recognition methods make a lot of significant progresses. Tong *et al.* [85] explore the dynamic and semantic relationships of facial AUs to improve their recognition performance. A dynamic Bayesian network is built by Tong *et al.* [84] for better facial activity understanding. Senechal *et al.* [72] combines different types of features using SimpleMKL learning algorithm to extract geometric and appearance information simultaneously. Sandbach *et al.* [69, 70] exploit 3D motion-based features between frames of 3D facial geometry sequences for dynamic AU detection and further expression recognition. AU-based methods decompose the facial expressions into different individual muscle activities, and then infer the expression categories based on the AU detection results. These methods can have great representation power, but AU detection itself is quite difficult and it is still an open problem to the community.

Message and sign judgement facial expression analysis methods generally consist of the two main steps: facial representation and expression recognition.

Facial representation derives a set of features from original facial images to effectively represent all faces. Different features have been applied to either the whole-face or specific face regions to extract the facial appearance changes, such as Gabor [6, 40, 60], haar-like features [96], local binary patterns (LBP) [62, 73]. Zafeirious *et al.* [110] explored the graph structures with landmarks to represent the variations among different expressions. In Shan *et al.* [74], facial images are equally divided into small regions, and then LBP features are extracted from these empirically weighted sub-regions to represent the facial appearance. The LBP features are shown to be effective in expression recognition, so our method will also utilize the LBP features with the same sub-region division strategy. Different from their work, we will focus on learning the effective sub-regions statistically.

Expression recognition aims to correctly categorize different facial representations. Support Vector Machine (SVM) [6, 74, 104] is the most popular and effective learning method in facial expression recognition. Shan's work [74] is the most similar work with ours, so it will be considered as the baseline. For fair comparison, our method will also employ SVM as the the

classification algorithm.

Besides these works, there are also some works utilizing the geometric features [14, 64, 65, 116], such as the location of facial feature points (corners of the eyes, mouth, etc.). Some methods perform facial expression analysis based on 3D face models [14,71,106]. More works on fusion of audio and visual information can be found in [111].

2.3.1 Multi-task Sparse Learning

Sparsity methods have attracted much attention in computer vision, multimedia and medical image communities, and have been employed in many applications, such as face recognition [97], background substraction [42], image annotation and retrieval [114], and shape prior based segmentation [115]. Many algorithm are proposed to solve these problems of sparsity priors, such as greedy methods (basis pursuit (BP) [16], matching pursuit [61], orthogonal matching pursuit (OMP) [86]), or $L1$ norm relaxation and convex optimization [11, 35, 52].

Multi-task sparse learning is an inductive transfer machine learning approach. It aims to learn a problem together with some related problems for better performance [13, 29]. Multi-task sparse learning is then designed in [3] for feature selection, through encouraging multiple predictors from different tasks to share similar parameter sparsity patterns. Multi-task sparse learning also obtained a rewarding performance on handwritten character recognition in [36]. Yuan *et al.* [109] developed a visual classification algorithm by learning the shared parts among different representation tasks. Recently, Chen *et al.* [15] provided a faster solution to multi-task sparse learning problems.

Suppose there are T related tasks, and $(x_i^t, y_i^t), i = 1, 2, ..., N_t$ is the training set of task t, where each sample is represented by K-dimensional features, $x_i^t \in R^K$, and $y_i^t \in \{-1, 1\}$ indexes x_i^t is negative or positive. w^t is a K-dimensional vector of representation coefficients for task t. All the w^ts are the rows of the matrix $W = [w_k^t]_{t,k}$, while every column of the matrix W is a T-dimensional vector that means the representation coefficients from the k-th feature across different tasks, $w_k = [w_k^1, w_k^2, ..., w_k^T]'$. Multi-task sparse learning aims to learn the shared sparse information among all the tasks. The formulation with L_1/L_2 mixed-norm

regularization is as follows:

$$\arg\min_{W} \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} J^t(w^t, x_i^t, y_i^t) + \lambda \sum_{k=1}^{K} \|w_k\|_2 \tag{2.2}$$

where $J^t(w^t, x_i^t, y_i^t)$ is the cost function of the tth task, λ is a constant to balance the sparsity, and $\sum$ is the mathematic format for L_1 norm. The regularization term encourages most columns of matrix W to be zero, and the remaining non-zero columns indicate the corresponding features are shared features across all the tasks.

Chapter 3

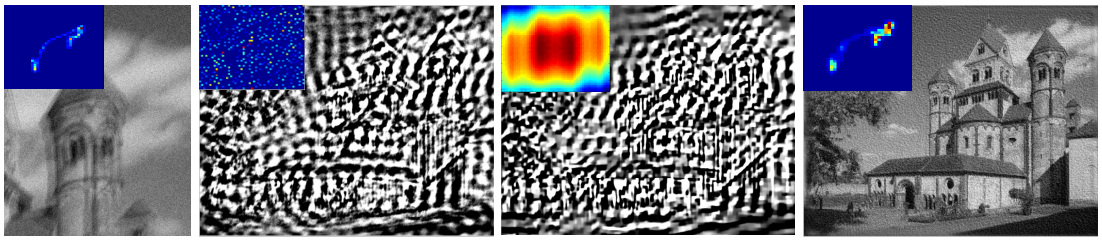
Single Image Deblurring using Directional Filters

3.1 Problem Background

Taking handheld photos in low-light conditions is challenging. Since less light is available, longer exposure times are needed – and without a tripod, camera shake is likely to happen and produce blurry pictures. Increasing the camera light sensitivity, i.e., using a higher ISO setting, can reduce the exposure time, which helps. But it comes at the cost of higher noise levels. Further, this is often not enough, and exposure time remains too long for handheld photography, and many photos end up being blurry *and* noisy. Although many techniques have been proposed recently to deal with camera shake, most of them assume low noise levels. In this work, we do not make this assumption and aim to restore a sharp image from a blurry and noisy input.

Many single image blind deconvolution methods have been recently proposed [19, 21, 34, 37, 47, 53, 57, 75, 100]. Although they generally work well when the input image is noise-free, their performance degrades rapidly when the noise level increases. Specifically, the blur kernel estimation step in previous deblurring approaches is often too fragile to reliably estimate the blur kernel when the image is contaminated with noise, as shown in Fig. 3.1. Even assuming that an accurate blur kernel can be estimated, the amplified image noise and ringing artifacts generated from the non-blind deconvolution also significantly degrade the results [20, 48, 108, 112].

To handle noisy inputs in single image deblurring, Tai and Lin [78] first apply an existing denoising package [2] as preprocessing, and then estimate the blur kernel and the latent image from the denoised result. This process iterates a few times to produce the final result. However, applying existing denoising methods is likely to damage, at least partially, the detailed blur



(a) Input (b) Cho and Lee [19] (c) Levin et al. [58] (d) Our method

Figure 3.1: Previous deblurring methods are sensitive to image noise. (a) Synthetic input image with 5% noise and the ground truth kernel (overlayed). It is cropped to better show blur and noise. (b) Estimated kernel and latent image by Cho and Lee [19]. (c) Results by Levin et al. [58]. (d) Results of our method.

information that one can extract from the input image, thereby leading to a biased kernel estimation. In 3.2, we illustrate that standard denoising methods, from bilateral filtering to more advanced approaches such as Non-Local Means [10] and BM3D [27], have negative impacts on the accuracy of kernel estimation.

In this chapter, we propose a new approach for estimating an accurate blur kernel from a noisy blurry image. Our approach still involves denoising and deblurring steps. However, we carefully design the denoising filters and deblurring procedures in such a way that the estimated kernel is not affected by the denoising filters. That is, we shall see that, unlike existing approaches, we can theoretically guarantee that our approach does not introduce any bias in the estimated kernel.

Our approach is derived from the key observation that if a directional low-pass linear filter is applied to the input image, it can reduce the noise level greatly, while the frequency content, including essential blur information, along the orthogonal direction is not affected. We use this property to estimate 1D projections of the desired blur kernel to the orthogonal directions of these filters. These projections, also known as the *Radon transform*, will not be affected by applying directional low-pass filters to the input image, except for the noise reduction. Based on this observation, we apply a series of directional low-pass filters at different orientations, and estimate a slice of kernel projection from each image. This yields an accurate estimate of the Radon transform. Finally, we reconstruct the blur kernel using the *inverse Radon transform*. Once a good kernel is obtained, we incorporate denoising filtering into the final deconvolution process to suppress noise and obtain a high-quality latent image. Results on synthetic and

real noisy data show that our method is more robust and achieves better results than previous approaches.

3.2 Side Effects of Denoising as Preprocessing

Before introducing our approach, we first analyze the negative impact of employing denoising as preprocessing on kernel estimation. In single image deblurring, a blurry and noisy input image b is usually modeled as:

$$b = \ell * k + n, \tag{3.1}$$

where ℓ, k and n represent the latent sharp image, blur kernel, and additive noise, respectively, $*$ is the convolution operator. Solving ℓ and k from input b is a severely ill-posed problem, and the additional noise n makes this problem even more challenging.

Assuming that ℓ is known, a common approach to solve for k is:

$$k = \underset{k}{\operatorname{argmin}} \left\{ \|b - k * \ell\|^2 + \rho(k) \right\}, \tag{3.2}$$

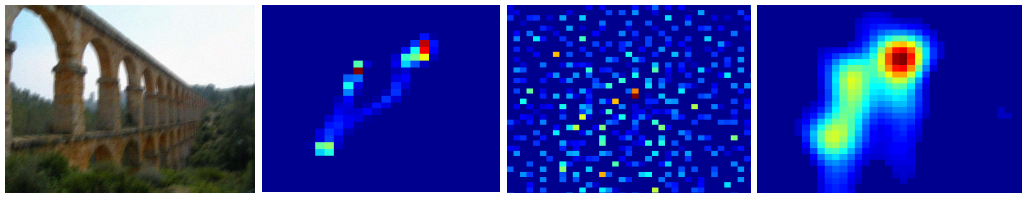
where $\rho(k)$ is the additional regularization term that imposes smoothness and/or sparsity prior on k. Without considering the regularization term, this becomes a least-squares problem and the optimal k can be found by solving the following linear system:

$$\mathbf{L}^T\mathbf{L}\mathbf{k} = \mathbf{L}^T\mathbf{b} = \mathbf{L}^T(\mathbf{b}' + \mathbf{n}), \tag{3.3}$$

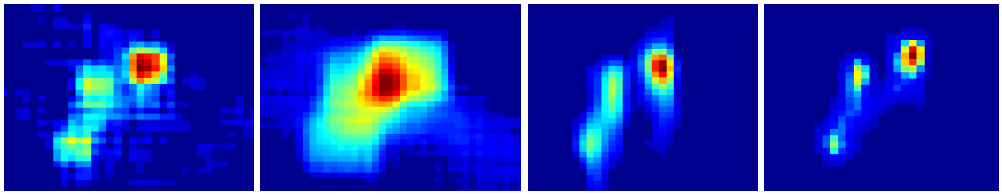
where $\mathbf{k}$ and $\mathbf{b}$ are the corresponding vector forms of k and b, respectively, and $\mathbf{L}$ is the matrix form of ℓ. We also introduce the noise-free blurry image $\mathbf{b}' = \mathbf{b} - \mathbf{n}$. We estimate the relative error of $\mathbf{k}$ with respect to the noise in $\mathbf{b}$ using the condition number of the linear system, that is:

$$\begin{aligned} \frac{e(\mathbf{k})}{e(\mathbf{b})} &= \frac{\|(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{n}\| / \|(\mathbf{L}^T\mathbf{L})^{-1}\mathbf{L}^T\mathbf{b}'\|}{\|\mathbf{L}^T\mathbf{n}\| / \|\mathbf{L}^T\mathbf{b}'\|} \\ &\leq \|(\mathbf{L}^T\mathbf{L})\| \cdot \|(\mathbf{L}^T\mathbf{L})^{-1}\| = \kappa(\mathbf{L}^T\mathbf{L}), \end{aligned} \tag{3.4}$$

where $e(\mathbf{k})$ and $e(\mathbf{b})$ are relative errors in $\mathbf{k}$ and $\mathbf{b}$, respectively. Thus, the noise $\mathbf{n}$ in the input image will be amplified at most by the condition number $\kappa(\mathbf{L}^T\mathbf{L})$ for kernel estimation,



(a) Input (b) True kernel (c) No denoising (d) Gaussian filter



(e) Bilateral filter (f) Non-local means (g) BM3D (h) Our method

Figure 3.2: The side effects of employing different denoising methods as preprocessing step in single image deblurring. (a) the synthetic input image with 5% noise. (b) the ground truth kernel. (c) the blur kernel estimated without applying any denoising method to the input image (a). (e)-(g) the estimated blur kernels after applying different denoising filters. (h) the kernel estimated by our method.

where $\mathbf{L}^T\mathbf{L}$ is often called the deconvolution matrix and has a block-circulant-with-circulant-block (BCCB) structure [51]. Eq. 3.4 shows that the upper bound on the error in the estimated kernel is proportional to the amplitude of the noise in input image. Building on this result, one can attempt to apply sophisticated denoising filter to the blurry image to reduce the noise amplitude, hoping that this will improve the kernel estimate. However, denoising filters also alter the profile of edges, e.g., [9]. This information is critical to accurate kernel estimation, and as we shall see, the benefits of the noise reduction are often outweighed by the artifacts caused by the profile alteration.

To illustrate it, we first look at a simple noise reduction method, Gaussian smoothing. Convolving with a Gaussian G_g decreases the noise level. However, the kernel estimation then becomes:

$$\begin{aligned} k_g &= \underset{k_g}{\operatorname{argmin}} \|b * G_g - \ell * k_g\|^2 \\ &= \underset{k_g}{\operatorname{argmin}} \|(\ell * k + n) * G_g - \ell * k_g\|^2 \\ &\approx \underset{k_g}{\operatorname{argmin}} \|\ell * (k * G_g - k_g)\|^2 = k * G_g, \end{aligned} \tag{3.5}$$

where k is the blur kernel for the original input image and k_g is the optimal solution after Gaussian denoising. Eq. 3.5 shows that the estimated kernel k_g is a blurred version of the

actual kernel k. Further, since G_g is a low-pass filter, the high frequencies of k are lost and recovering them from k_g would be very difficult, if possible at all. This result comes from the initial noise reduction and is independent of the kernel estimation method.

Although more sophisticated denoising methods are better at preserving high frequencies, denoising remains an open problem for which no perfect solution exists. Since no information about the blur kernel can be observed in uniform regions of the blurry image, edges are the main source of information that drives deblurring algorithms either implicitly or explicitly, e.g., [19, 21, 47, 100]. Even small degradations introduced by state-of-the-art denoising techniques can have a strong impact on deblurring results as shown in Fig. 3.2. In this experiment, we apply bilateral filtering [83], non-local means [10] and BM3D [27] to a test image with 5% noise, i.e., noise of standard deviation 0.05 when the intensity range is $[0, 1]$, and then use Cho and Lee's method [19] to estimate the blur kernel. The estimated kernels are not accurate due to the side effects of denoising.

The recent approach of Tai and Lin [78] first applies an existing commercial denoising package (NeatImage [2]) to the input image, then iteratively applies a motion-aware non-local mean filtering and deblurring to refine the results. Although special treatment has been added into the process, both the commercial denoising package and the non-local means filter have the same negative impacts on kernel estimation as we will show in 3.4.

3.3 Methodology

In the previous section, we have shown that there is a tension between noise reduction and edge preservation. The former helps to estimate a more accurate kernel, but the latter hinders it. Our experiments showed that even state-of-the-art denoising filters do have negative impacts on kernel estimation. In this section, we resolve this problem by using directional blur and the Radon transform to estimate the kernel. Our approach reduces the noise without degrading blur information, thereby producing better kernels.

3.3.1 Applying directional filters

We now show that directional low-pass filters can be applied to an image without affecting its Radon transform, while decreasing its noise level. We consider the directional low-pass filter f_θ:

$$I(\mathbf{p}) * f_\theta = \frac{1}{c}\int_{-\infty}^{\infty} w(t) I(\mathbf{p} + t\mathbf{u}_\theta) dt, \tag{3.6}$$

where I is an image, $\mathbf{p}$ is a pixel location, t is the spatial distance from one pixel to $\mathbf{p}$, c is the normalization factor defined as $c = \int_{-\infty}^{\infty} w(t)dt$, and $\mathbf{u}_\theta = (\cos\theta, \sin\theta)^T$ is a unit vector of direction θ. The profile of the filter is determined by $w(t)$, for which we use a Gaussian function: $w(t) = \exp(-t^2/2\sigma_f^2)$, where σ_f controls the strength of the filter.

Filtering the image affects the estimated kernel. With the same argument as for Eq. 3.5, the kernel that we estimate from the filtered image $b_\theta = b * f_\theta$ is:

$$k_\theta = k * f_\theta. \tag{3.7}$$

Similarly to filtering with a 2D Gaussian G_g, applying f_θ averages pixels and reduces the noise level. Since f_θ filters only along the direction θ, it has nearly no influence on the blur information in the orthogonal direction. We exploit this property to estimate the projection of the *original* kernel k along the direction θ. The projection can be formulated as Radon transform [21,82], which is the collection of integrals of a signal (i.e., k) along projection lines. The particular value on Radon transform corresponding to one projection line $\rho = x\sin(\theta) + y\cos(\theta)$ is:

$$R_{\theta'}(\rho) = \int\int k(x,y)\delta(\rho - x\sin(\theta) - y\cos(\theta))dxdy, \tag{3.8}$$

where $k(x,y)$ indicates the value at the coordinate (x,y) on kernel k. θ and ρ are the angle and offset of the projection line, respectively. Thus, the projection of kernel k_θ along the projection direction θ is:

$$R_{\theta'}(k_\theta) = R_{\theta'}(k * f_\theta) = R_{\theta'}(k) * R_{\theta'}(f_\theta) = R_{\theta'}(k), \tag{3.9}$$

where $R_{\theta'}(\cdot)$ is the Radon transform operator to the direction θ', and $\theta' = \theta + \pi/2$. It is a linear operator, and one can verify that $R_{\theta'}(f_\theta)$ is a 1D delta function, given the definition of f_θ (Eq. 3.6). Eq. 3.9 shows f_θ has no impact on the Radon transform of the blur kernel to the

orthogonal direction of the filter. This is the foundation of the proposed approach. An example is shown in Fig. 3.3.

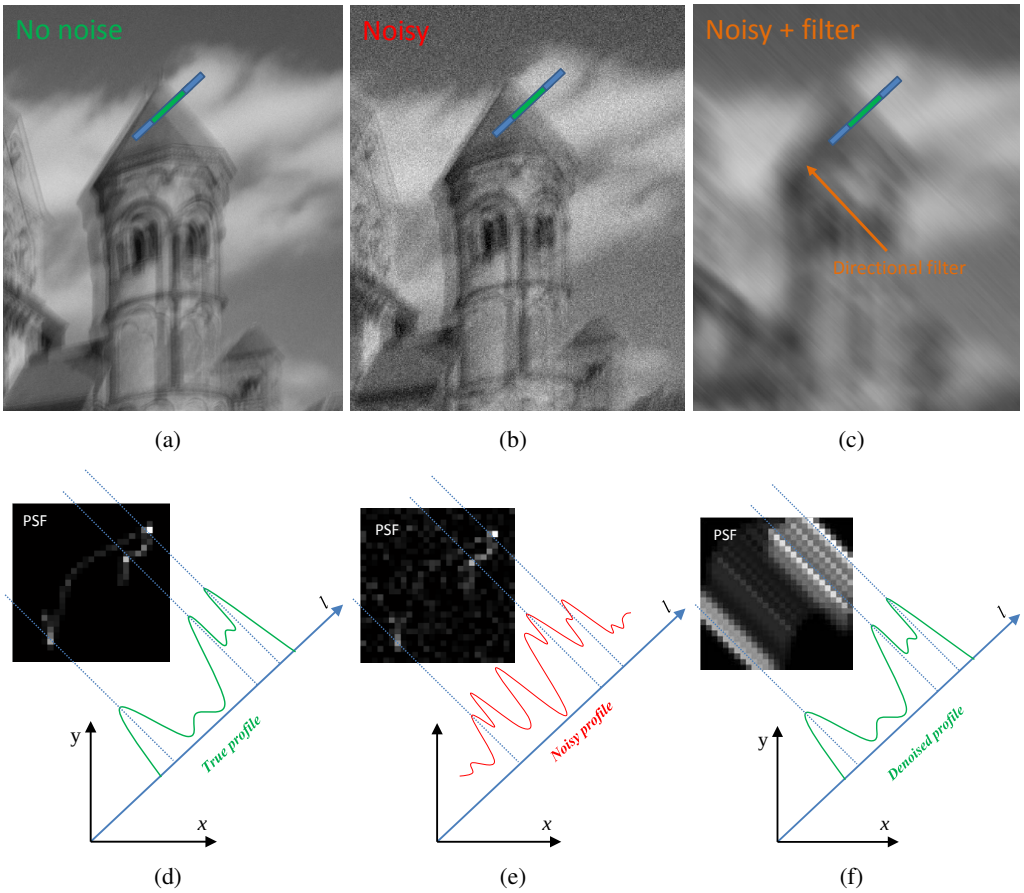


Figure 3.3: Directional denoising mechanism in single image deblurring. (a)(b)(c) are the synthetic image before adding noise, after adding noise, and after applying a directional filter($\theta = 3\pi/4$), respectively. (d)(e)(f) are the corresponding estimated blur kernels and their Radon transforms in the same direction. Note that the estimated kernel in (f) is largely damaged by the directional filter, but its Radon transform is the same as the one in (d).

3.3.2 The algorithm

We now explain how we recover the sharp image, with the kernel estimation first, and then the deconvolution step.

Noise-aware kernel estimation

Based on the above analysis, we apply a directional blur f_θ, estimate the combined blur kernel k_θ, and then project it along the same direction of the filter to get the corresponding Radon

Algorithm 1 *Multiscale noise-aware blind deconvolution*

Input: The pyramid $\{b_0, b_1, ..., b_n\}$ by down-sampling the input blurry and noisy image b, where $b_0 = b$.

Output: blur kernel k_0 and latent image ℓ_0.

1: Apply an existing nonblind approach ([19] in our implementation) to estimate k_i and ℓ_i for b_i,$i = n, ..., 1$.
2: Upsample ℓ_1 to generate initial ℓ_0.
3: **repeat**
4: Apply N_f directional filters to the input image b_0, each filter has a direction of $i \cdot \pi / N_f$, $i = 1, ..., N_f$, where N_f is the number of directional filters.
5: For each filtered image b_θ, use ℓ_0 as the latent image to estimate k_θ.
6: For each optimal kernel k_θ, compute its Radon transform $R_{\theta'}(k_\theta)$ as in Eq. 3.9, along the direction $\theta' = \theta + \pi/2$.
7: Reconstruct k_0 from the series of $R_{\theta'}(k_\theta)$ using inverse Radon transform.
8: Update ℓ_0 based on the new k_0 using a noise-aware nonblind deconvolution approach.
9: **until** k_0 converges.
10: With the final estimated kernel k_0, use the final deconvolution method described in 3.3.2 to generate the final output ℓ_0.

transform. We repeat this process to get a set of projections. Finally, we compute the 2D kernel using the inverse Radon transform [82]. The advantage of this strategy is that it greatly reduces noise when applying f_θ, while keeping the computed Radon transform intact. However, so far, we have assumed that the latent image ℓ is known when estimating the blur kernels. This is not the case in practice, and even with state-of-the-art kernel estimation techniques, recovering k_θ from b_θ, which is a blurry image convolved with an additional directional blur, has proven to be challenging. The additional filter tends to make nearby edges "collide" with each other, which in turn introduces errors in the estimated kernel.

For a more reliable kernel estimation, we adopt the multiscale blind deconvolution framework commonly used in previous approaches [19, 100]. We create an image pyramid of the input image b as $\{b_0, b_1, ..., b_n\}$, where b_0 is the original resolution, and estimate the blur kernel in a bottom-up fashion from b_n to b_0. Since noise is largely removed by image downsizing, we apply an existing approach by Cho and Lee [19] to estimate the blur kernels k_i and latent images ℓ_i from layer b_n to b_1. Only for the full resolution layer b_0, we apply the directional filter f_θ and then estimate the kernel using the robust deconvolution technique described later in this section. The process is described in Algorithm 1. Steps 4 to 7 are also illustrated in

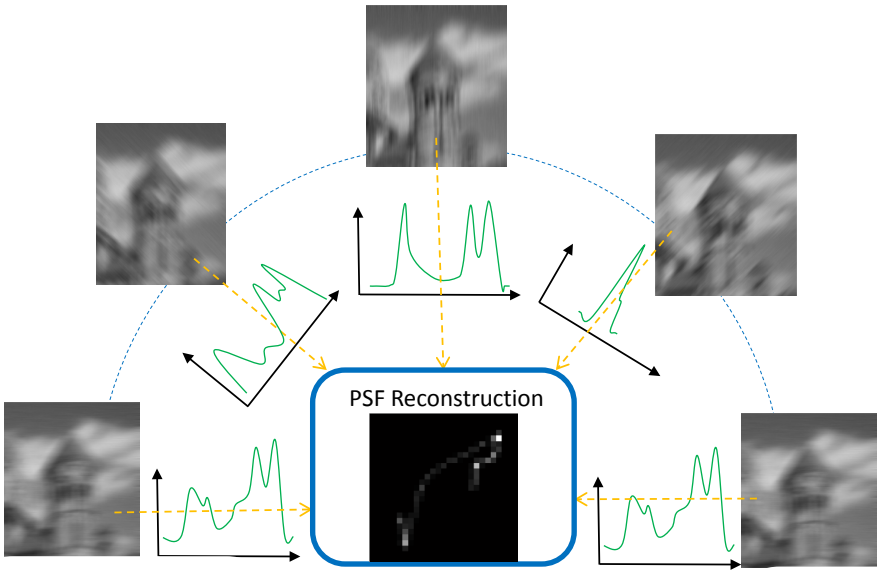


Figure 3.4: Illustration of applying directional filters for blur kernel estimation from a noisy input image. We apply directional filters in different orientations to the input image. From each filtered image a corresponding kernel is computed first, then projected along the same direction to generate the correct radon transform of the true kernel. The final blur kernel k_0 is reconstructed using inverse Radon transform [21].

Fig. 3.4. Specifically, in Step 5, although each filtered image b_θ is severely blurred with the additional filtering, the latent image ℓ_0, initialized from the multiscale process, is relatively sharp and clean, which allows us to estimate k_θ as:

$$k_\theta = \underset{k_\theta}{\operatorname{argmin}} \left\{ \|\nabla b_\theta - k_\theta * \nabla \ell_0\|^2 + \rho(k_\theta) \right\}, \tag{3.10}$$

where ∇ is the gradient operator. This process is robust to noise because ∇b_θ is a low-pass filtered image. In Step 8, nonblind deconvolution is employed to update ℓ_0 based on the new k_0. However, existing methods do not work well in this case since we need to estimate a clean ℓ_0 from a noisy image b_0, and the results of previous methods are prone to inaccuracy. To generate a noise-free ℓ_0, we minimize the following energy function that aims for limiting the impact of noise on the result:

$$\|\nabla \ell_0 * k_0 - \nabla b_0\|^2 + w_1 \|\nabla \ell_0 - u(\nabla \ell_1)\|^2 + w_2 \|\nabla \ell_0\|^2, \tag{3.11}$$

where $u(\cdot)$ is the upsampling function, and w_1 and w_2 are pre-defined weights. The second term encourages the gradient of ℓ_0 to be similar to the upsampled gradient field of ℓ_1, which is from the previous level in the pyramid. Since ℓ_1 contains much less noise due to image

downsizing, incorporating this term can effectively reduce the noise level in ℓ_0. This non-blind deconvolution step is an intermediate step in blur kernel estimation that produces sufficiently accurate images at a limited computational cost. In the next section, we describe a more sophisticated non-blind deconvolution algorithm for generating high-quality final latent image given the estimated kernel.

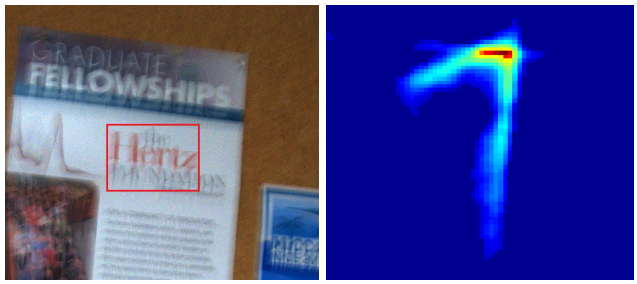
It is worth mentioning that for simplicity, in the above discussion we assume b_1 is almost noise-free after downsizing the image by half. However, this will not be true, if severe noise presents in b_0. To deal with severe noise, we will only use previous methods to estimate blur kernels from b_n to b_2 in Step 1 of the algorithm, and then apply noise-aware kernel estimation from Step 2 to 9 to the last two layers b_1 and b_0. We applied this modified version of the algorithm to examples with 10% noise (Gaussian noise with standard deviation of 0.1) in 3.4.

Discussion

Cho et al. [21] also use the Radon transform to recover the blur kernel. However, their approach to compute the kernel projection is different from ours. They rely on heuristics to identify straight edges in the images, and extract the projections from these edges. Because this process relies on a few arbitrary thresholds to locate and analyze the edges, it is sensitive to noise. We also show that it performs poorly on noisy inputs in the experimental section. In comparison, our approach does not rely on such arbitrary thresholds and performs well on noisy images.

Final noise-aware nonblind deconvolution

Once an accurate k_0 is estimated, we use it to estimate a good latent image ℓ_0 from the noisy input b_0. This is not a trivial task when b_0 contains severe noise [108]. However, since k_0 is fixed at this stage, it is safe to apply existing denoising methods in the process. This is in sharp contrast to Tai and Lin's method [78] where denoising and kernel estimation interfere with each other.



(a) Input image (b) Estimated kernel

(c) Zoran&Weiss [120] (d) Cho et al. [20] (e) Our method

Figure 3.5: Comparison results of our final noise-aware nonblind deconvolution with other recent nonblind deconvolution methods. The results are obtained using the same input image and the estimated kernel. (c),(d),(e) show the zoom-in results.

In our approach, we minimize the following energy function to estimate the final ℓ_0:

$$\|\ell_0 * k_0 - b_0\|^2 + w_3\|\ell_0 - \text{NLM}(\ell_0)\|^2, \tag{3.12}$$

where $\text{NLM}(\cdot)$ is the non-local means denoising operation [10], and w_3 is a balancing weight. Minimizing this energy function will ensure that the deblurred result is noise-free, and can best fit with k_0 and b_0 as well.

Directly minimizing this energy is hard because $\text{NLM}(\ell_0)$ is highly nonlinear. We found that iterating the following two steps yields a good result in practice:

$$\ell_0' = \text{NLM}(\ell_0), \tag{3.13a}$$

$$\ell_0 = \underset{\ell_0}{\text{argmin}} \left\{ \|\ell_0 * k_0 - b_0\|^2 + w_3\|\ell_0 - \ell_0'\|^2 \right\}. \tag{3.13b}$$

For initialization, we set ℓ_0' to be zero (a black image). Solving Eq. 3.13b yields a noisy ℓ_0 that also contains useful high-frequency image structures. In the alternating minimization process, the noise in ℓ_0 is gradually reduced, while the high-frequency image details are preserved. To show the effectiveness of our method, we compare it with other two recent non-blind deconvolution methods, i.e., Zoran and Weiss [120] and Cho et al. [20] in Fig. 3.5.

3.4 Experimental Results

We implemented our method in Matlab on an Intel Core i5 CPU with 8GB of RAM. We apply directional filters along 36 regularly sampled orientations, that is, one sample every 5°. The computation time is a few minutes for a one-megapixel image. For all the experiments, we set the extent σ_f of the directional filter to 30 pixels. We also set $w_1 = 0.05$ and $w_2 = 1$ (Eq. 3.11), and $w_3 = 0.05$ (Eq. 3.12).

3.4.1 Synthetic data

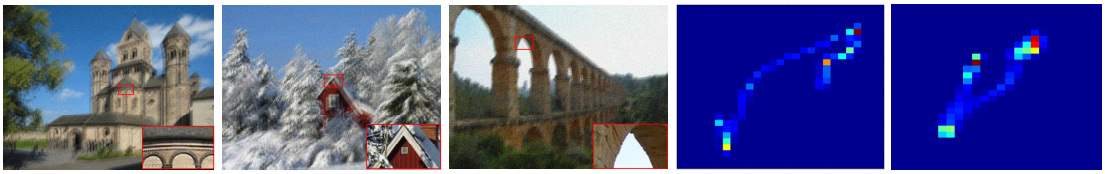
We first conducted experiments on images that we convolved with a known blur kernel and to which we added noise in a controlled fashion. This allows us to report quantitative measures in addition to visual results.

Comparisons with Tai and Lin's [78] method

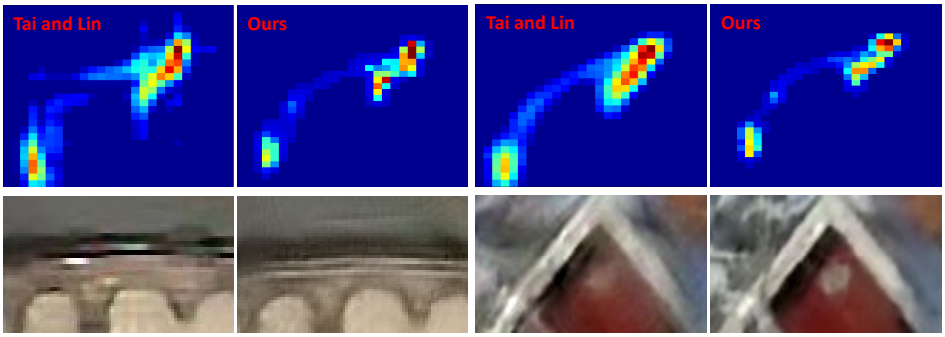
Tai and Lin's method [78] is the most related work to ours since it also seeks to handle noisy images. This section focuses on comparing this method with our approach. We first ran comparisons on synthetic images (Fig. 3.6), where the latent sharp images were blurred using two blur kernels provided by Levin et al. [57]. We then added Gaussian noise with zero mean and standard deviations of 0.05 and 0.1 for a [0,1] intensity range. Tai and Lin kindly provided the results for their method. The comparison shows that visually our estimated blur kernels are closer to the ground truth, and our estimated latent images contain more details and less ringing artifacts. We also evaluate the results quantitatively by computing the Peak Signal-to-Noise Ratio (PSNR) and Structural SIMilarity (SSIM) (Table 3.1).

Comparisons with other methods

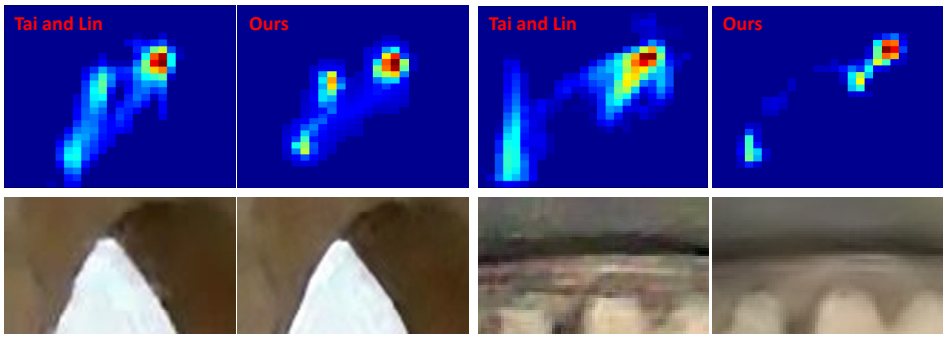
We also conducted experiments to explore how noise affects the performance of other state-of-the-art single-image blind deconvolution methods. Using the "Aque" image and the blur kernel shown in 3.6(e), we generated 10 input images with noise from 1% to 10%. We then applied different blind deconvolution methods to these test images, and measure the PSNR curve of each method (3.7). The accuracies of previous methods degrade rapidly when the



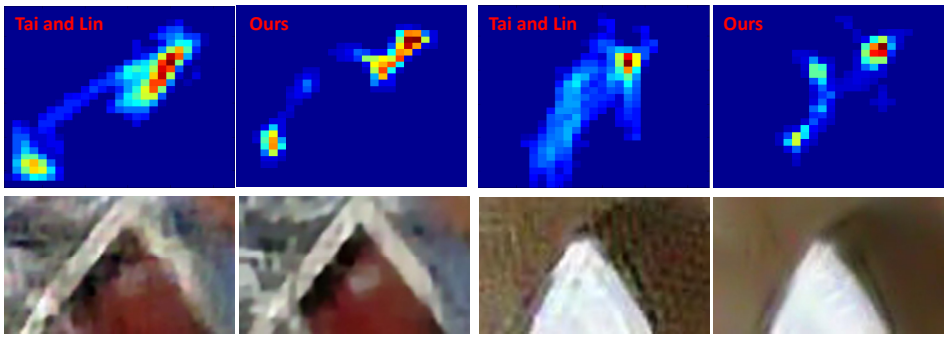
(a) Abbey(input, 5% noise) (b) Chalet(input, 5% noise) (c) Aque(input,10% noise) (d) Kernel 1 (e) Kernel 2



(f) Abbey (result, 5% noise) (g) Chalet (result, 5% noise)



(h) Aque (result, 5% noise) (i) Abbey (result, 10% noise)



(j) Chalet (result, 10% noise) (k) Aque (result, 10% noise)

Figure 3.6: Comparing Tai and Lin's method [78] and our method on synthetic data. Three input blurry image examples with different levels of noise are shown in (a),(b),(c). (d) and (e) are the ground truth blur kernels from Levin et. al. [57]. (d) is used for the examples "Abbey" and "Chalet", and (e) is used for the example "Aque". (f-k) show the estimated kernels and the latent images of Tai and Lin's method and our method with 5% noise and 10% noise. Due to the space limit only the areas highlighted by the bounding boxes in (a-c) are shown. Full size images for comparison are in the supplementary material.

		PSNR		SSIM	
	Noise	5%	10%	5%	10%
Abbey	Tai	22.43	21.05	.8122	.7242
	Ours	**22.73**	**21.61**	**.8150**	**.7270**
Chalet	Tai	19.79	18.95	.8244	.7162
	Ours	**22.80**	**19.35**	**.8273**	**.7200**
Aque	Tai	26.58	24.53	.8206	.7415
	Ours	**28.46**	**25.58**	**.8512**	**.7469**

Table 3.1: The comparison experiments of our method and Tai and Lin [78] on synthetic blurry images with different amount of noises. The performances are evaluated by PSNR and SSIM, comparing the generated latent images with the ground truth.

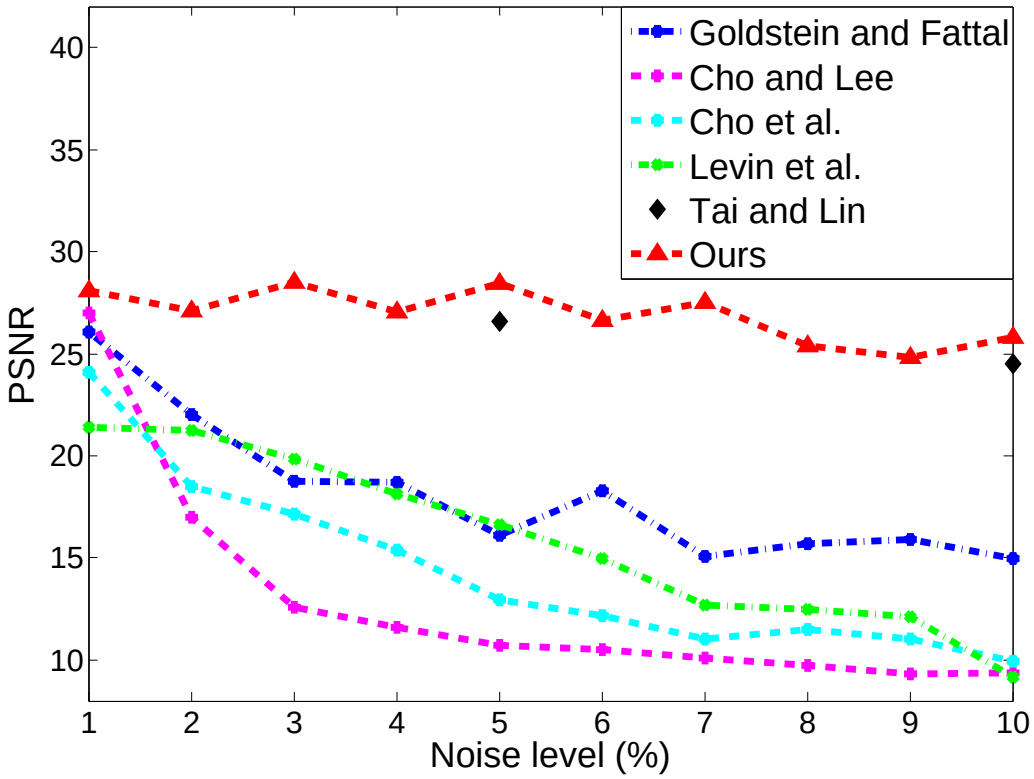
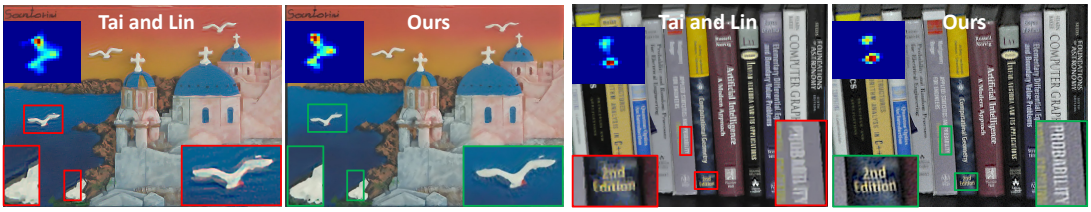


Figure 3.7: The PSNR curves of various blind deconvolution algorithms, including Goldstein and Fattal [37], Cho and Lee [19], Cho et al. [21], Levin et al. [58] and our method, on the 10 synthetic test images with noise level from 1% to 10%, generated by the "Aque" image and the kernel shown in 3.6(e). The two data points of Tai and Lin's method [78] are shown as black diamonds, which are provided by the authors. While the PSNR values are closer to ours, the visual difference is still significant; our approach produces cleaner images (3.8). All images are included in the supplementary material.

noise level increases. On the contrary, our method is more robust, i.e., it works more reliably in the presence of noise, and achieves satisfactory results even when the input noise level is high. This figure also includes two data points of the Tai and Lin's method [78] provided by the authors themselves.

3.4.2 Results on real examples

We first compared our method and Tai and Lin's method on real-world images shown in their original paper [78], and the results are shown in 3.8. The results of other state-of-the-art methods can be found in [78]. Our estimated kernels are sharper than Tai and Lin's. The close-ups show that our method recovers more high-frequency details. For the boundaries of objects, our results have less noticeable ringing artifacts. Overall, our approach produces visually more satisfying results.



(a) Santorini (b) Books

Figure 3.8: Comparisons of Tai and Lin's method and our method on real-world images from [78]. Our results contain more high-frequency details and less ringing artifacts. Zoom-in regions are shown in bounding boxes.

We further show our results on real-world photographs that were captured under common low-light conditions with a Nikon D90 DLSR camera and a $18-105$mm lens. We compare our results with those of other state-of-the-art methods, including Goldstein and Fattal [37], Cho and Lee [19], Cho et al. [21], Levin et al. [58]. The results (Fig. 3.9) show that our recovered latent images exhibit less artifacts, such as noise and ringing, and contain more high-frequency details at the same time. These observations are consistent across all test images. We provide additional examples in supplemental material.

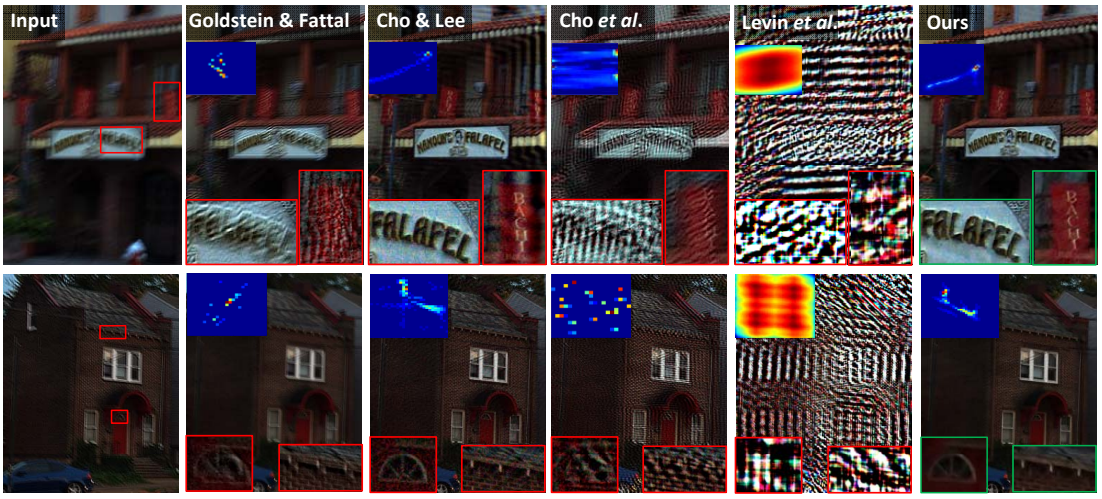


Figure 3.9: Comparisons on real-world examples, where we compare our results with the results of Goldstein and Fattal [37], Cho and Lee [19], Cho et al. [21], Levin et al. [58]. More results are in the supplementary material.

3.5 Conclusion

We have shown that most state-of-the-art image deblurring techniques are sensitive to image noise. In this chapter, we propose a new single image blind deconvolution method that is more robust to noise than previous approaches. Our method uses directional filters to reduce the noise while keeping the blur information in their orthogonal direction intact. By applying a series of such directional filters, we showed how to recover correct 1D projections of the kernel in all directions, which we use to estimate an accurate blur kernel using the inverse Radon transform. We also introduced a noise-tolerant non-blind deconvolution technique that generates high-quality final results. The effectiveness of the proposed approach is demonstrated on several comparisons on synthetic and real data.

Chapter 4

Image Deblurring with Face Prior

4.1 Problem Background

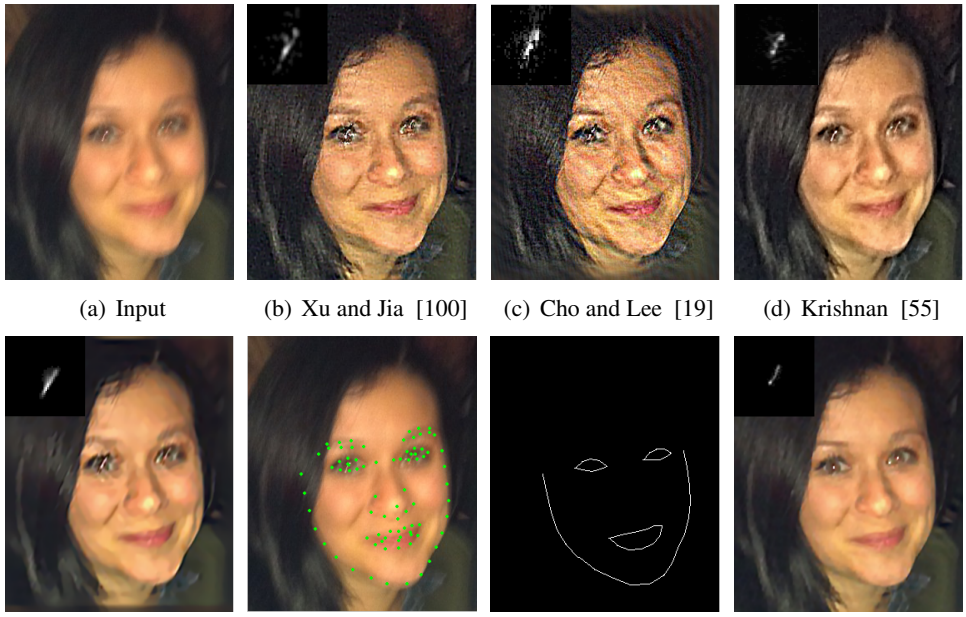
Single image deblurring has been studied for decades, and has attracted much attention with significant progress in recent years [19, 20, 37, 48, 117]. The purpose of the image deblurring is to restore the sharp image and recover the latent blur kernel from single blurry image. The motion blur is normally modeled as a spatially-invariant model:

$$b = \ell * k + n, \tag{4.1}$$

where ℓ, k and n represent the latent sharp image, blur kernel, and additive noise respectively, $*$ is the convolution operator. The image deblurring problem is an ill-posed problem, because for a given blurry image, there would be many pairs of latent images and blur kernels which would meet this equation.

Normally, additional information or constraints are required to constrain the solution. The statistic prior knowledge of natural image is commonly used to constrain the solving process, such as heavy-tailed gradient distribution [34, 56, 58, 75], and L_1, L_2 priors [57]. Since these priors are obtained from natural images, they are effective for generic cases. However, for specific cases, such as face images and text images, these priors would not be able to achieve promising results because the pre-learned priors from natural images are not suitable for these specific images. Specific priors or constraints, which capture the object properties, are needed to perform well on these specific cases, such as text object properties [18], and face structure properties [63].

Implicit or explicit extraction of salient edges is another category of constraint which achieved great success in many blind deconvolution methods [19, 21, 100, 117]. These methods typically employ a salient edge prediction step, which is mainly based on heuristic image



(a) Input (b) Xu and Jia [100] (c) Cho and Lee [19] (d) Krishnan [55]

(e) Zhong et al. [117] (f) Landmark (g) Mask (h) **Our result**

Figure 4.1: The deblurring results of the state-of-the-art methods on a real face image. (a) is the input blurry face image. (b) - (e) are the results of all different methods. (f) is the facial landmark localization results on this blurry face. (g) is the generated mask from (h). The final result of our method is shown in (h).

processing method to explore high contrast local edge structures. These methods may fail on images with less textures, since few salient edges could be recovered to provide enough blur information. For example, face images have similar skin color components and low contrast edges. Thus, existing edge-prediction methods can not robustly detect the edges and obtain promising results. Moreover, only local structure information is used to locate the salient edges without considering the latent global object structure. Therefore, for these methods, the ambiguity of selecting salient edges will be difficult to eliminate, and will degrade the performance greatly.

We test some state-of-the-art single image deblurring methods on a given blurry face. The restored latent images and the estimated kernels are shown in Fig. 4.1. We can see the existing methods can not achieve promising results because of the above mentioned difficulties. Fig. 4.1(f) shows the landmark localization result on this blurry face image, and the mask (in Fig. 4.1(g)) generated from the landmarks captures the main structure of the face. Our estimated kernel and restored image are shown in Fig. 4.1(h), which demonstrate the superiority over the

compared methods.

In this work, we propose a facial landmark localization based face image deblurring method to address the above-mentioned problems. Facial landmark localization module is trained on face images with manual annotations, which embed the facial structure information. Thus, the landmarks detected on the face would be used as a global constraint for salient edge selection. We generate a mask by connecting the landmarks and use the mask to generate an initial blur kernel.

4.2 Methodology

Existing state-of-the-art edge-based deblurring methods employ heuristic edge selection explicitly or implicitly. Since face images do not contain much texture, it is very difficult to identify salient edges. Although previous methods normally use the coarse-to-fine strategy to eliminate the ambiguities in edge selection, the local structure information can not provide enough evidence for accurate edge prediction. In our method, the landmark localization is employed to constrain the edge selection with the global face structure, which would improve the accuracy of kernel estimation.

4.2.1 Landmark Localization on Blurry Images

Recently, landmark localization methods made huge progress in accuracy and robustness, especially with the development of regression frame [12, 26, 99], and the employment of deep convolutional neural networks (DCNN) [105, 119]. The landmark localization utilizes coarse-to-fine strategy to refine the facial landmarks,

Most of the facial landmark localization methods are designed for sharp images. Nevertheless, they are robust to blur at certain extent since the face image used for training varies in resolution. Moreover, the coarse-to-fine strategy is employed and the initial positions of the landmarks are first estimated on low-resolution down-sampled images, and then the locations are refined on higher resolution images. Take the Zhou et al. [119] as an example. Their method is based on the deep convolutional neural network framework. The first level networks predict the bounding boxes for the inner points and contour points separately. Then, the second

level predicts an initial estimation of the positions. Because the inner points would have more gradient evidence and can be refined further on finer scale, the third and fourth levels are designed further to improve the predictions of mouth and eyes by considering them as individual components.

In our cases, the motion blur is different from the low-resolution blur. But when the images are down-sampled, the difference will decrease greatly. Thus, we would like to explore how robust is the facial landmark localization method to motion blur. Given a sharp image with resolution of 320×240 pixels, we blur it with different sizes of kernels, and perform the facial landmark localization on the synthesized blurry images. In Fig. 4.2, we can see the landmark localization is robust to the increasing size of blur kernel with relatively small kernel.

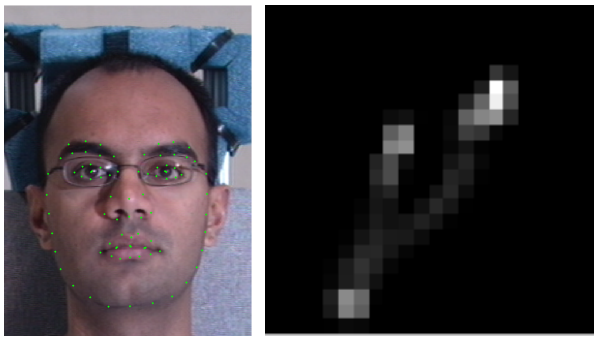
The landmarks can capture the main structure and the salient edges on the face from kernel size of 2×2 to 22×22. The proposed deblurring method is also able to obtain promising sharp latent images. When the kernel size increases to 32×32 pixels, the contour points are erroneous, but the inner points with eyes and mouths are still robust (in Fig. 4.2(e)). With partial correct information with the inner face points, the proposed method can still recover promising latent images (in Fig. 4.2(i)). When the kernel size increases to super large (47×47 pixels), the landmark localization method totally fails (in Fig. 4.2(f)), and thus our method can not recover the latent image neither (in Fig. 4.2(j)).

4.2.2 Face deblurring with Landmarks

Most blind deconvolution methods contain two main steps: kernel estimation and non-blind deconvolution. Those methods iterate these two steps until converge to the optimal solution. Kernel estimation tries to estimate the motion blur kernel from the blurred image. For edge-based methods, salient edges need to be predicted beforehand from the blurred image, and treated as the gradients from the latent sharp image to estimate the blur kernel:

$$k = \underset{k}{\operatorname{argmin}} \left\{ \|\nabla b - k * \nabla \ell\|_2^2 + \lambda * \rho(k) \right\}, \tag{4.2}$$

where ∇b is the gradient calcuated form the blurred image, $\nabla \ell$ is the gradient of the predicted salient edges, k is the blur kernel, $\rho(k)$ is the prior on k, which could be L_2, L_0 norm or other constraints and λ is the weight for the prior.



(a) Sharp Image (b) Blur kernel

(c) 12×12 (d) 22×22 (e) 32×32 (f) 47×47

(g) Restored image (size 12×12) (h) Restored image (size 22×22) (i) Restored image (size 32×32) (j) Restored image (size 47×47)

Figure 4.2: The performance of the landmark localization method [119] on blurry images with increasing kernel sizes. The restored latent images from these blurring images and their detected landmarks are also shown here. (a) shows the sharp face image and the landmark detection result. The resolution of the sharp image is 320×240 pixels. The blurry images in (c) - (f) are generated by convoluting the sharp image with different sizes of the kernel shown in (b). (c) - (f) show the facial landmark localization results on these blurry images. With the landmark localization results, the corresponding deblurring results are shown in (g) - (j). Landmark localization method is robust to motion blur when the kernel size is relatively small (smaller than 22×22 on this image), and becomes worse with bigger kernels. The proposed method can handle partial incorrectness (e.g., size 32) with landmark localization, but will also fail when the landmarks are totally misplaced (e.g., size 47). Please zoom in for better view.

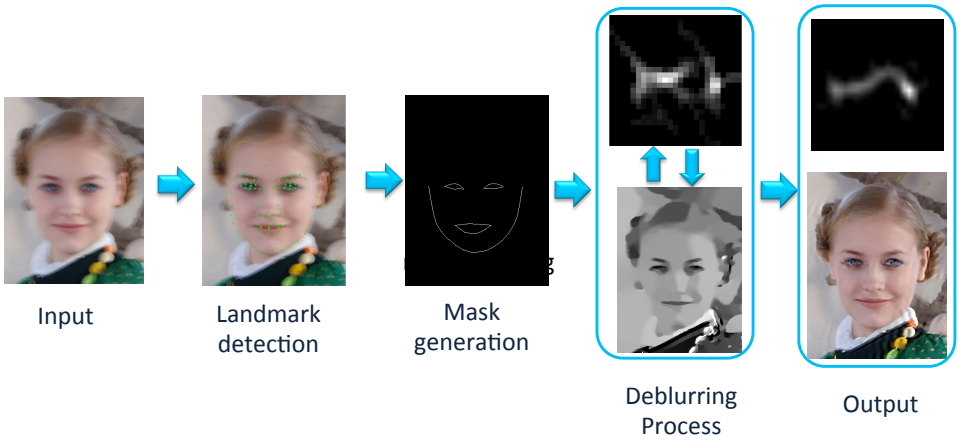


Figure 4.3: The framework of our deblurring method. Given a blurred face image, landmark localization method is applied to extract the salient points. A mask is then generated from the landmarks of the face contour, mouth, and eyes. The mask would be used to estimate an initial blur kernel. After a few iterations of kernel estimation and non-blind deconvolution in the deblurring process, the method would converge to the optimal solution.

The salient edges (i.e., large gradient) of face could come from the face contour, mouth lips, eyes, eyebrows, hair, nose, glasses if exist. Pan et al. [63] show that the deblurring process would still get promising results with only edges from eyes, mouth and face contour. The landmark localization would robustly capture the salient edges from the face contour, mouth, and eyes. As shown in Fig. 4.3, given an blurred face image, we first locate the facial landmarks, and connect the landmarks on face contour, mouth and eyes to generate the mask. To generate a smoothing mask, dilation and erosion are employed. Then, an initial blur kernel can be estimated by:

$$k = \underset{k}{\operatorname{argmin}} \left\{ \|\nabla b - k * M * \nabla b\|_2^2 + \lambda \|k\|_2^2 \right\}, \tag{4.3}$$

Where M is the estimated binary mask, the other symbols are defined the same as in 4.2. λ is set to 1 in our experiments. In this kernel estimation, the L_2 norm regularization on the kernel is applied, which would make the optimization to be a quadratic problem and thus be solved fast by conjugate gradient descent methods. With the constraint of the mask, the initial blur kernel could be estimated. Compared to other existing methods, the initial estimated kernel is more accurate since it utilizes the global face structure prior which is embedded in the mask.

With the initial start point, iterative kernel estimation and non-blind deconvolution would

be performed. To handle the large blur and eliminate the edge ambiguities, the L_0 norm is employed to restore the latent image l. The L_0 norm is showed to be very effective in removing the ringing artifacts [102]. The non-blind deconvolution would be :

$$k = \underset{I}{\operatorname{argmin}} \left\{ \|\nabla b - k * \nabla \ell\|_2^2 + \gamma \|\nabla \ell\|_0 \right\}, \tag{4.4}$$

where γ is set to be 0.002 in the experiment. Since Equ. 4.4 is half-quadratic, we need to decompose the problem with auxiliary variables to get an approximate solution as Xu et al. [101]. The Equ. 4.4 can be rewritten using the auxiliary variables $W = (w_x, w_y)^T$ as :

$$k = \underset{I,W}{\operatorname{argmin}} \left\{ \|\nabla b - k * \nabla \ell\|_2^2 + \beta \|W - \nabla \ell\|_2^2 + \gamma \|W\|_0 \right\}, \tag{4.5}$$

We note that Equ. 4.5 can be splitted into two sub-problems and efficiently solved by alternatively minimizing I and W independently. In each iteration, The optimal solution of I would be obtained by

$$k = \underset{I}{\operatorname{argmin}} \left\{ \|\nabla b - k * \nabla \ell\|_2^2 + \beta \|W - \nabla \ell\|_2^2 \right\}, \tag{4.6}$$

This sub-problem is a quadratic problem and has a closed-form solution which can be computed quickly.

The optimal solution of W can be estimated from

$$k = \underset{I}{\operatorname{argmin}} \left\{ \beta \|W - \nabla \ell\|_2^2 + \gamma \|W\|_0 \right\}, \tag{4.7}$$

which could be obtained by :

$$W = \begin{cases} \nabla \ell, & if\ |\nabla \ell|^2 \geq \frac{\gamma}{\beta} \\ 0, & otherwise \end{cases}$$

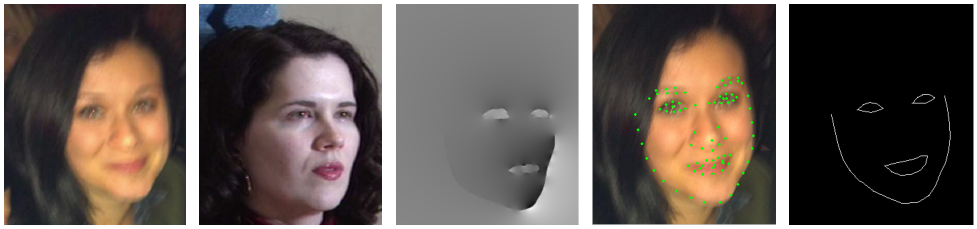
Even this is an approximate solution for Equ. 4.4, it has been proved to be effective in image smoothing [101], and image deblurring [102]. Our method also shows its capabilities in eliminating ambiguities and removing ringing artifacts.

After the iterative optimization process, the blur kernel would be estimated. Since L_0 norm in ℓ estimation focuses on the most strong gradient, the estimated latent image would be too

smooth. We employ the final non-blind deconvolution methods proposed in Levin et al. [56] to recover the final latent image.

4.2.3 Comparison with Pan et al. [41]

Among the existing single image deblurring methods, Pan et al. [41] is the most related work to ours, which explores the face structure information to help with the deblurring process, and also employs L_0 norm constraint in the intermediate non-blind deconvolution step. However, there are a few major differences between their methods and ours. The biggest difference is the way in extracting the global face structure. Their method tries to find a exemplar image from their image database, which should be similar with the input blurred image. The salient edges or large gradients from the exemplar image are then directly used in the deblurring process. This strategy faces two major problems. Firstly, the training database is limited, so it may not be able to find good exemplar image for an arbitrary input image. Take the input in Fig. 4.4 as an example, [41] would first find an exemplar image for the input image from their training set (i.e., Multi-PIE dataset [38]). Since the dataset only contains face images with upright pose with fixed angles, it can not cover all the arbitrary angles.



(a) input (b) Exemplar image (c) Predicted gradients (d) **Landmark localization (ours)** (e) **Generated mask (ours)**

Figure 4.4: Face structure information extraction comparison between Pan et al. [41] and our method. The selected exemplar image is quite different from the input image, including pose, facial expression, face shapes. Thus the predicted gradients are also erroneous, and would unavoidable introduce bias in deblurring. Our method can get more accurate facial landmarks and more promising mask, since we do not directly rely on any limited exemplar image set.

Moreover, it is also quite difficult to match the blurred images and sharp images with unknown blur kernel and different people. Therefore, the selected exemplar image could be quite different from the input image in many ways, such as, the head pose, expressions, and face

shape. Some examples are shown in 4.4(a) and 4.4(b). On the contrary, the mask calculated by our method fits the original face better, and also our method has better generalization capability with various pose and angles.

4.3 Experiments

The robustness of the landmark localization is first investigated on blurred image with blur kernels of different sizes. Then our method is compared with other state-of-the-art methods on both synthetic and real face images. All our experiments run on a Macbook laptop with 2.4 GHz Intel Core i5 processor and 8 GB memory. To deblur a face image with 320×240 pixels, our method takes around 20 seconds.

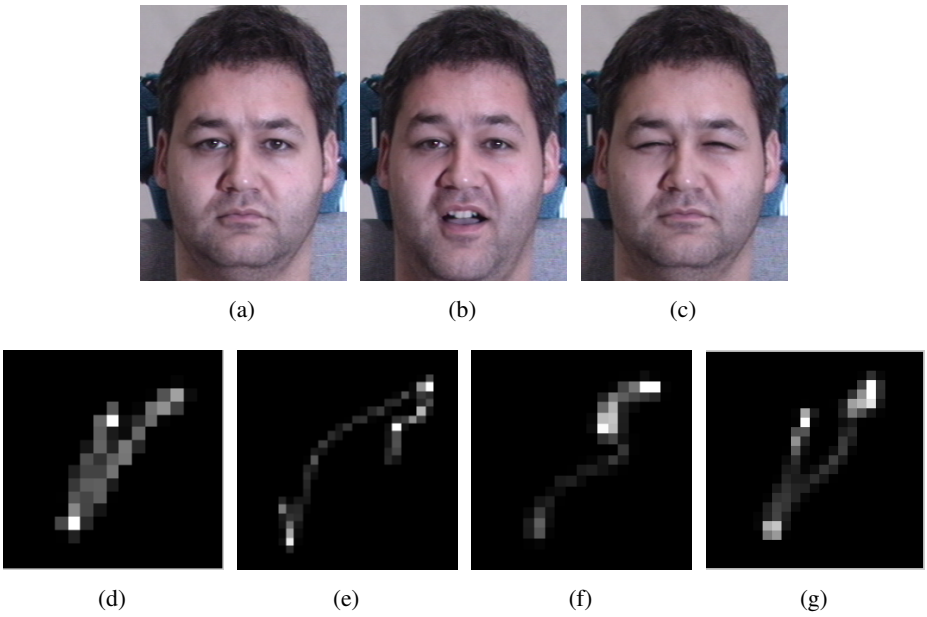


Figure 4.5: The sample sharp images and blur kernels in our experiments. Each subject would have three different expressions for quantitative analysis, such as (a) - (c). (d) - (g) are the 4 selected blur kernels used for synthesizing blurred face images.

4.3.1 Landmark Localization Robustness

The landmark localization method provides the global facial structure information for our deblurring method. Thus its accuracy affects the performance of our proposed deblurring method

directly. The robustness of landmark localization method on blurred face image is rarely analyzed in the literature. In this chapter, we would investigate the landmark localization method proposed in [119] quantitatively. We randomly select 27 face images with 9 subjects from Multi-PIE database [38]. Each subject would have 3 images with different expressions. The four ground truth kernels we used to synthesize the blurred image are from Levin et al. [57]. The sample images and blur kernels can be found in Fig. 4.5. For each kernel, we resize it to different sizes from 7×7 to 47×47 pixels with every 5 pixels. Each image will be convoluted with the kernels with different sizes to generate blurred images, as in Fig. 4.2.

To measure the difference between two detected facial landmarks on image p and q, the average distance between the corresponding points is normalized by distance of two eye outside corners.

$$dis = \frac{1}{T} \sum_{t=1}^{T} \frac{|p_t - q_t|}{|p_\ell - q_r|} \tag{4.8}$$

where t is the index of the facial landmarks. There are $T = 83$ points in our experiments. ℓ, q are the indexes for the outside corners of left eye and right eye.

The landmark localization method [119] is performed on these blurred images. Some of the landmark results are shown in Fig. 4.2. The error distance of the detected landmarks is calculated by the Equ. 4.8. The landmark locations detected on the corresponding sharp image are considered as the ground truth. We then calculate the mean error distance for the images with the same size of kernel, which is shown in the Fig. 4.6. As expected, the performance of the landmark localization method degrades with increasing size of blur kernels. The average error distance increases at the beginning since the detection of the inner face landmarks are robust to the blur to some extent. However, when the kernel size increases up to 32×32 pixels, the blur would be too big for the landmark localization method, and thus the error distance increases quicker.

4.3.2 Comparison with Existing Deblurring Methods

To further evaluate our method, we compare our method with some state-of-the-art methods on both synthetic and real images.

Synthetic images: To quantitatively evaluate the proposed method, we run our method and

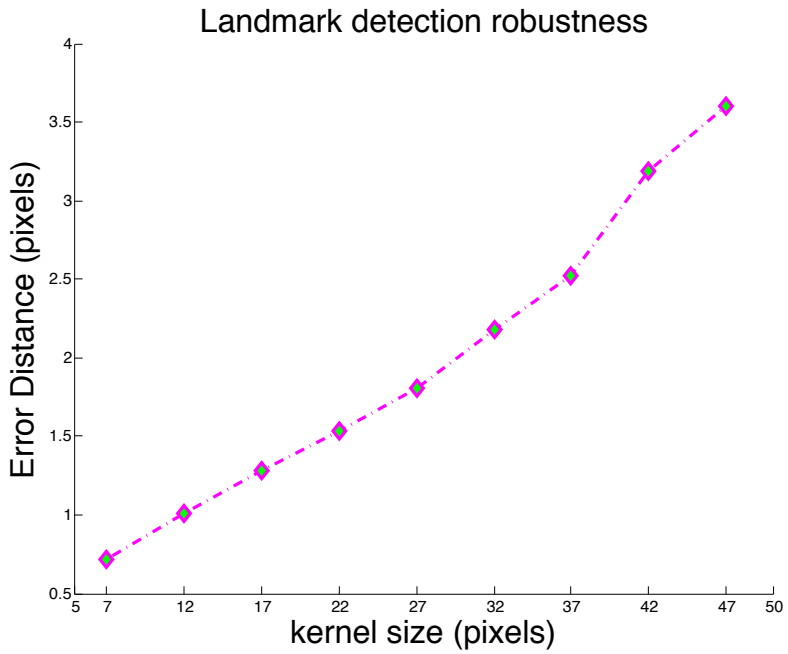


Figure 4.6: The robustness of landmark localization on blurry images. With the increasing blur kernel, the landmark detection will be more erroneous and the average error increases. When the kernel size increases up to 27×27, the landmark localization may fail, and thus the error increases even quicker. The resolution of the test images is 320×240 pixels.

some of the state-of-the-art methods on the synthesized blurry images, which are produced by different size of kernels. The Fig. 4.7 shows the results of different methods on the blurry images generated by a 17×17 kernel. To avoid the artifacts produced by the non-blind deblurring step, we use the restored image with the ground truth kernel as the ground truth for comparison (Fig. 4.7(c)). We compared our method with a few existing methods, i. e., Xu and Jia [100], Cho and Lee [19], Kristinan et al. [55], and Zhong et al. [117]. The average Peak signal-to-noise ratios (PSNR) of the restored results for different kernel sizes are shown in Fig. 4.8. With the kernel size increases, the performance of all the methods degrades. Our method outperforms the other methods when the kernel size is smaller than 42×42. This is because the landmark localization method can perform well when the kernel size is not super large, and thus provide face structure information to help with the deblurring. Meanwhile, the other methods, which only focus on local structures or take advantages of priors from natural images, can not get promising results as our method. When the kernel size increases up to 42×42, our method would perform similar with other methods, since the face prior is not accurate because of the

erroneous landmark locations.

(a) Sharp Image (b) Blurry image (c) Deblur with real kernel (d) Xu and Jia [100]

(e) Cho and Lee [19] (f) Krishnan [55] (g) Zhong et al. [117] (h) **Our result**

Figure 4.7: An synthetic example for the comparison of different existing methods. (b) is the synthesized blurry image of (a) with kernel size of 17×17 pixels. (c) is the non-blind deconvolution method with the ground truth kernel, and it will be used as the ground truth for restored image comparison. (d)-(h) show the recovered image by different state-of-the-art methods, and the corresponding PSNR are also shown.

Real Images: We also compared our method with other state-of-the-art deblurring methods on some real blurred images. The real images are different from synthesized blurred images, since they may contain complicated types of noises, saturated pixels and non-uniform kernels. The comparison results are shown in Fig. 4.1, Fig. 4.9, and Fig. 4.12. For image deblurring, the promising restored images should be sharp without blur, and also clean without noise and artifacts. In Fig. 4.9, our method performs slightly better than Zhong et al. [117], and gets sharper edges for the face contours. The other methods can not get promising results without a lot of noise and ringing artifacts. Our method also performs very robustly and consistently outperforms the compared state-of-the-arts methods on several real images as in Fig. 4.1 and Fig. 4.12.

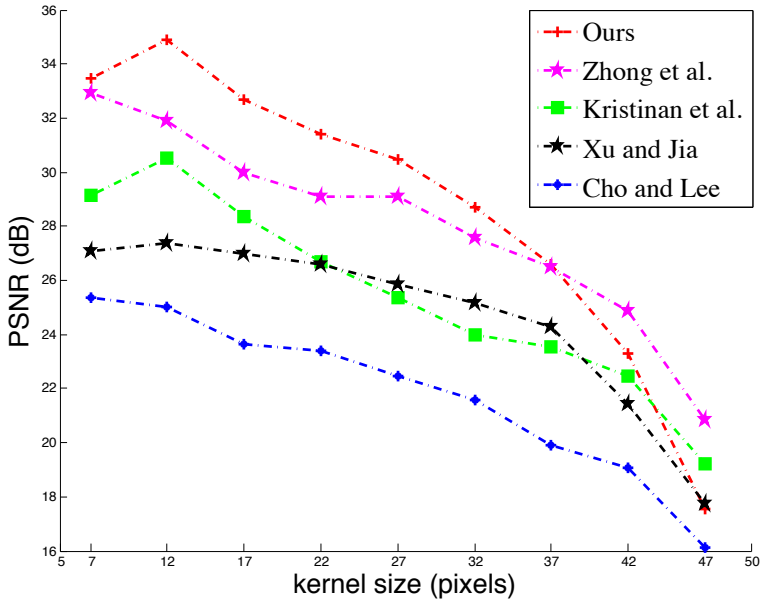
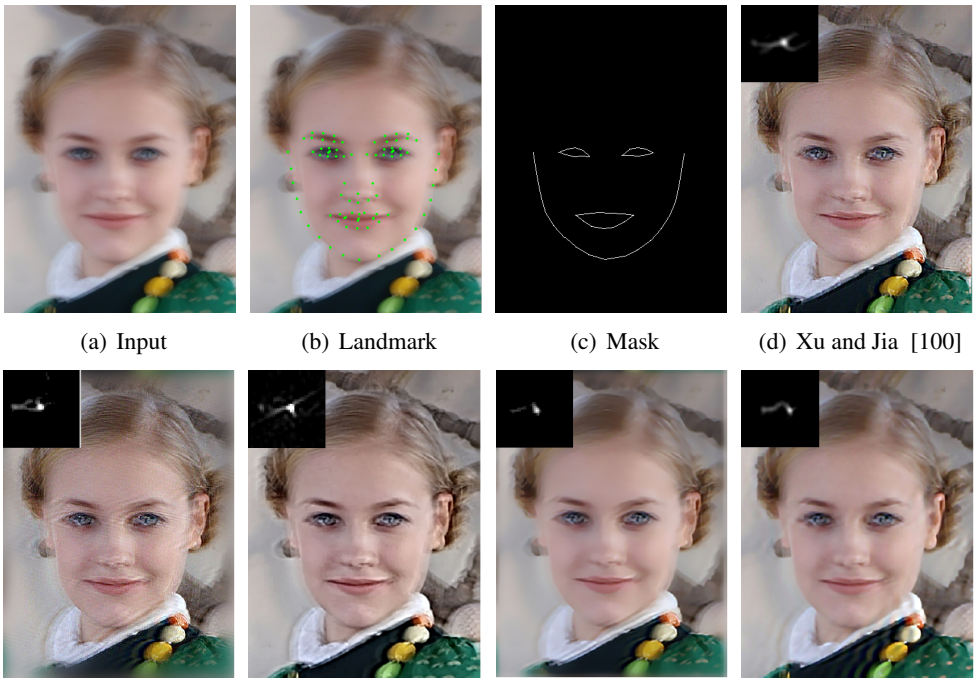


Figure 4.8: The quantitative analysis of different blind deconvolution methods. Our method shows its superiority over the compared method when the kernel size is reasonable. This is mainly because our method utilizes the additional face structure prior provided by landmark localization. When the landmark localization method fails with super large kernels, our method would perform similar with the other methods.



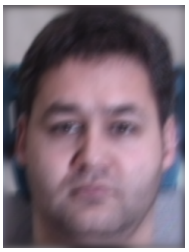
(a) Input (b) Landmark (c) Mask (d) Xu and Jia [100]

(e) Cho and Lee [19] (f) Krishnan [55] (g) Zhong et al. [117] (h) **Our result**

Figure 4.9: The Comparison results of the proposed method with other state-of-the-art methods. Our method contains cleaner and sharper result with much less noise and artifacts.

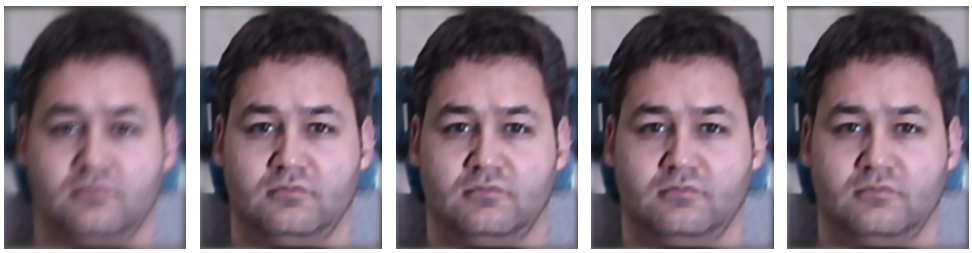
4.3.3 Iterative Landmark Localization and Deblurring

Our method strongly relies on the accuracy of landmark localization result, so iteratively run the landmark localization on the restored face image would potentially improve the accuracy of landmarks. We test our method on a face example blurred with a kernel of 17×17 pixels, and the results are shown in Figure. 4.10. It seems simply adding more iterations cannot improve the results much.



(a) Input

(b) Landmark (iter1) (c) Landmark (iter2) (d) Landmark (iter3) (e) Landmark (iter4) (f) Landmark (iter5)



(g) Result (iter1) (h) Result (iter2) (i) Result (iter3) (j) Result (iter4) (k) Result (iter5)

Figure 4.10: The deblurring results with iterative landmark localization and deblurring.

Our iterative deblurring process uses strong L_0 norm to suppress small changes, so it is also robust to the errors in landmark localization to some extent. We would like to explore if the iteration can improve the result in real cases potentially. To eliminate the errors in landmark localization on blurry images, we use the sharp image to extract the landmark localization

instead. The results are shown in Figure. 4.11. We could see our method can produce similar result when the landmark localization contains reasonable errors.

(a) Landmark (gt) (b) Landmark (blurry) (c) Result (gt) (d) Result (blurry)

Figure 4.11: Landmark error effects on face deblurring.

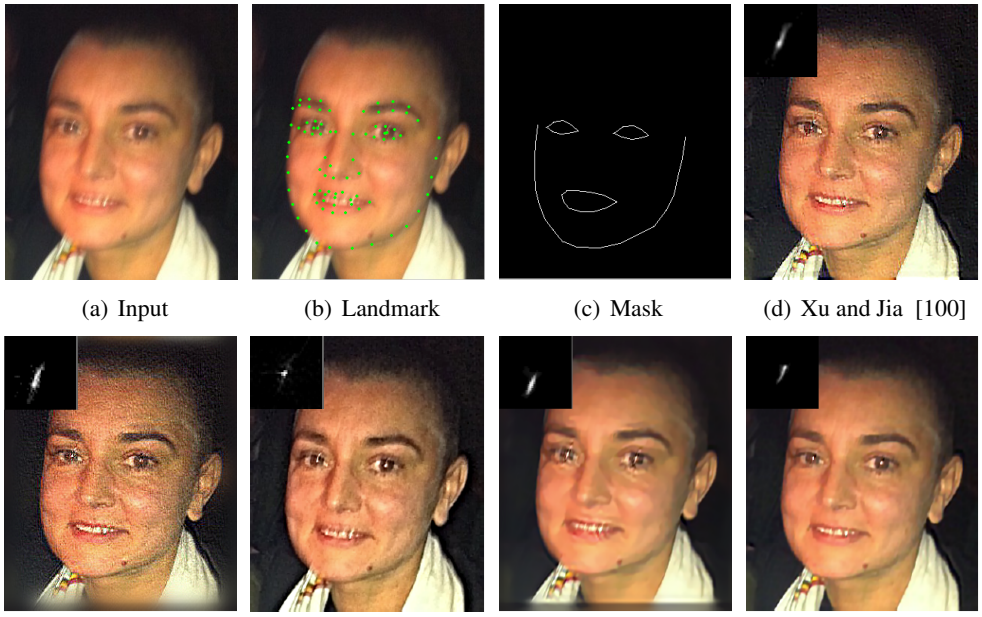
On the other hand, computational cost is one of the essential factor for image deblurring algorithm. Iteratively applying landmark localization and deblurring will greatly increase the running time for a single image. Therefore, our method only contains single localization and deblurring loop in most cases.

4.3.4 Failure Case

Our method strongly relies on the landmark localization results to identify the salient edges. Even the landmark localization method is already robust on various faces, it may still have some errors on images with large kernel size. When the landmark localization only contains partial correct information, our deblurring method can also recover the incorrect part by iterative salient edge detection and non-blind deconvolution (e.g., Fig. 4.2(i)). However, when the kernel size is too big, salient edges are more difficult to extract in the later edge extraction. So if the landmark localization is wrong, our method will also fail and can not get promising results (e.g., Fig. 4.2(j)).

4.4 Conclusion

In this chapter, we proposed a new face blind deconvolution method based on facial landmark localization. Since previous methods either focus on the local edge structure or utilize the



(a) Input (b) Landmark (c) Mask (d) Xu and Jia [100]

(e) Cho and Lee [19] (f) Krishnan [55] (g) Zhong et al. [117] (h) **Our result**

Figure 4.12: The Comparison results of the proposed method with other state-of-the-art methods. Our method contains cleaner and sharper results with much less noise and artifacts.

natural image prior, they can not get promising results on facial images. The facial landmark localization is embedded with global face structure information, thus it would be employed to eliminate the ambiguities in identifying salient edges. Our method utilizes the detected landmarks to predict the salient edges of faces from blurry images. The robustness of landmark localization on blurry image is analyzed in the experiment. Extensive experiments on both synthetic images and real images show the effectiveness of the proposed method in restoring sharp and clean images from blurred face images.

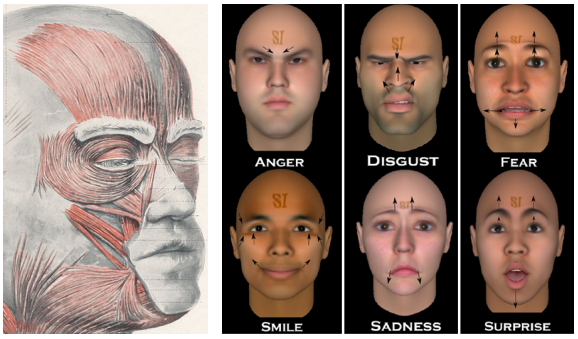
The face shape prior is utilized to help with the salient edge detection on face in this work, and then boost the face image deblurring. The underline reason for this success is that the faces have relative stable shape pattern. Thus, this approach can be extended to deblur other structured objects, such as cars, human body, particular animals. The extraction and utilization of other structured object information to boost the deblurring performance will be our future work.

Chapter 5

Learning Multi-scale Active Facial Patches for Expression Analysis

5.1 Problem Background

Facial expressions play significant roles in our daily communication, due to their abilities to reflect human emotions, and social interaction. In the past three decades, automatic facial expressions recognition has become an increasingly fascinating topic in the computer vision and pattern recognition communities for their extensive applications, such as human-computer interface, multimedia, and security [68,92,94]. However, as the basis of expression recognition, the exploration of the functional facial features is still an open problem.



(a) (b)

Figure 5.1: (a) Illustration of facial muscles distribution [30]. (b) Major AUs for six expressions. The arrows represent for AUs.

Studies in psychology show that facial features of expressions are located around *mouth, nose*, and *eyes*, and their locations are essential for explaining and categorizing facial expressions. Through electrical muscle stimulation, Duchenne [1, 30] found that most expressions are invoked by a small number of facial muscles around the mouth, nose and eyes (See Figure 5.1(a)). This indicates that most of the descriptive regions for each expression are located

around certain face parts. Moreover, expressions can be generally categorized into six popular *"basic" or "universal" expressions* [44]: anger, disgust, fear, happiness, sadness and surprise. (Of course, there are a lot of complex expressions which are not basic expressions). These expressions seems to be universal across different ethnicities and cultures [33]. As shown in Figure 5.1(b), each of these basic expressions can be further decomposed into a set of several related action units (AUs) [31], e.g., happiness can be roughly decomposed to cheek raiser and lip corner puller. However, few existing methods statistically utilize these prior knowledge about facial muscle and AUs to aid facial expression analysis in computer vision community.

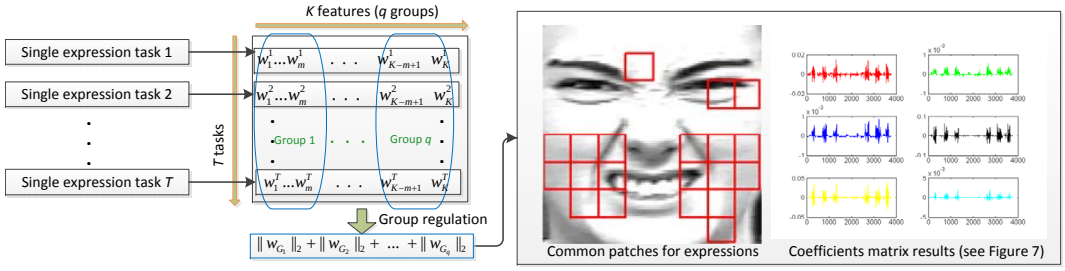


Figure 5.2: Discovering the common patches across six expressions using multi-task sparse learning (MTSL). Each single expression task is the binary classification task for one expression (See Figure 5.4). Expression tasks are combined in a MTSL model to select out the common patches under the group sparsity constraint.

Previous expression recognition methods can be generally categorized into two groups: *AU-based* methods and *message and sign judgement* methods. **AU-based** methods [32, 80, 81, 85, 87] recognize expressions by detecting AUs, which have more descriptive power, but these methods suffer from the difficulties of AU detection. **Message and sign judgement** methods [60, 74, 96] reveal the differences among expressions by facial appearance variations, which has been proved to be more reliable on single still images. However, these methods treat different facial parts equally or assign different weights to them empirically, thus lacking statistical support for the weight settings. This motivates us to fully make use of the prior knowledge from facial muscles and AU studies to extract the most discriminative regions, which can further assist expression analysis.

Inspired by the locations of AUs, we divide human face into non-overlapping patches on different scale levels, and then conceptually group these patches into three categories: *common facial patches*, *specific facial patches*, and the *rest*. *Common facial patches* are active ones for

all expressions. *Specific facial patches* are only active for one particular expression. Therefore, the most important facial patches are the common ones shared by all expressions; specific patches are only a few and only useful to discriminate a particular expression; the rest of the patches are of less help to expression recognition. The effective facial patches corresponding to different facial muscles may have different sizes, and the optimal size of patches is hard to be determined and it may vary among different areas of the face, we employ three different scale patches in our method to cover effective facial parts with different sizes. That is, each face image is divided into 8×8, 6×6 and 4×4 patches. The patches are smaller when the face image is divided into more patches. An example of face division (8×8) is illustrated in Figure 5.3(a).

A two-stage multi-task sparse learning framework is proposed to explore common and specific patches statistically on each scale level respectively. In the first stage, the binary classification problem for each expression is treated as an individual task (see Figure 5.4), then a multi-task sparse learning (MTSL) model is built based on these related tasks to extract the common facial patches. In the second stage, the face verification task (see Figure 5.5) is designed to be coupled with the previous classification task for one expression. In this way, another MTSL model can be constructed to find out the specific patches for this particular expression. Similarly, the specific patches for all the expressions can be figured out separately. After the common and specific patches on different scales are learned by the two-stage multi-task sparse learning model, they are combined to boost the facial expression recognition accuracy.

For all the scales, the common and specific patches, found by extensive experiments on the Cohn-Kanade database [49] and the MMI database [90], not only confirm the psychology discoveries of the facial muscles and AUs, but also provide more accurate appearance locations. Moreover, these common and specific patches at the same scale can be used to boost the performance of expression recognition. The learned patches are shown to be effective across different databases, e.g. Cohn-Kanade database [49], GEMEP-FERA [88]. Only using relatively small number of patches ($\sim 1/3$ of the face), our method still outperforms other methods in expression recognition. Finally, the learned effective patches at different scales can be also combined to further improve the expression recognition performance.

Our contributions are:

1) We provide a solid validation for an important psychology discovery, that only partial area of the face (corresponding to underlying facial muscles) are discriminative for expression recognition.

2) A two-stage multi-task sparse learning framework is proposed to formulate the commonalities among expressions, and to find out the locations of common and specific patches for expressions.

3) Multi-scale image division strategy is utilized to generate patches of different size for facial expression analysis. More convincing conclusion about facial parts (muscles) could be achieved, since they are of different sizes.

4) Extensive experiments with 3 different scales on three public databases demonstrate that these active patches are effective in recognizing expressions. The common and specific patches can be combined to improve the performances of state-of-the-arts. Patches across different scales can also been fused to further boost the performance.

The rest of the chapter is organized as follows. Section 2 reviews Related work of facial expression analysis and multi-task sparse learning. Section 3 presents our framework to learn common and specific patches based on multi-task sparse learning. These effective patches on all different scales are learned in this section. The experimental results on three public databases are shown in section 4. We conclude the chapter in section 5.

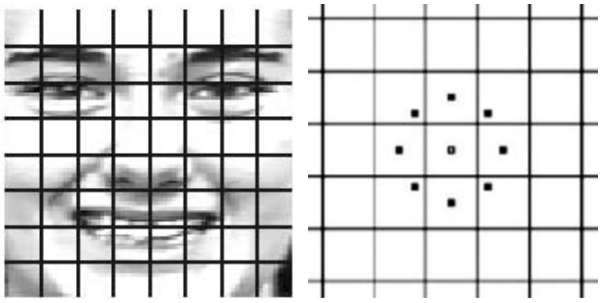
5.2 Methodology

In this section we first introduce the multi-scale facial appearance representation strategy, and then the learning procedures of common and specific patches at each scale level are illustrated. Finally, we design the classifier with these learned effective patches.

5.2.1 Multi-scale Appearance Representation

Facial expressions are usually manifested by local facial appearance variations. However, it is not easy to automatically localize these local active areas on a facial image. A facial image is

divided into p local patches, and then local binary pattern (LBP) features are used to represent the local appearance of the patch. These features have been proven to be a powerful descriptor in expression recognition [74] and face verification [95]. Since the facial parts corresponding to facial muscles do not have equal size, multi-scale division strategy should be applied to facial images to generate facial patches with different sizes. These patches could offer a more complete coverage for the effective facial parts than the generated patches if only one single scale is applied. The facial patches should have reasonable size, which can not be too big to cover too much facial parts or too small to have no physic meaning. In our method, we adopt three different scale sizes, We set $p = 8 \times 8, 6 \times 6, 4 \times 4$ in the experiments with the image size of 96×96. We denote these three scales as $S8, S6, S4$, respectively. An example of division for $S8$ is shown in Figure 5.3(a). For each patch, the uniform LBP features are extracted with the LBP operator $LBP_{8,1}$, as shown in Figure 5.3(b), and mapped to a m-dimensional histogram ($m = 59$ in our method).



(a) (b)

Figure 5.3: (a)A cropped facial image is divided into 64 patches. (b) LBP feature example. ($LBP_{P,R}$ refers to a neighborhood size of P equally spaced pixels on a circle of radius R that form a circularly symmetric neighbor set. $P = 8, R = 1$ for this example.)

Based on these local patches, the common patches across all expressions on each scale level are learned for expression recognition. Then, some specific patches for each expression are explored to enhance the performance. Finally, these learned patches on different scales are fused to further boost the recognition performance.

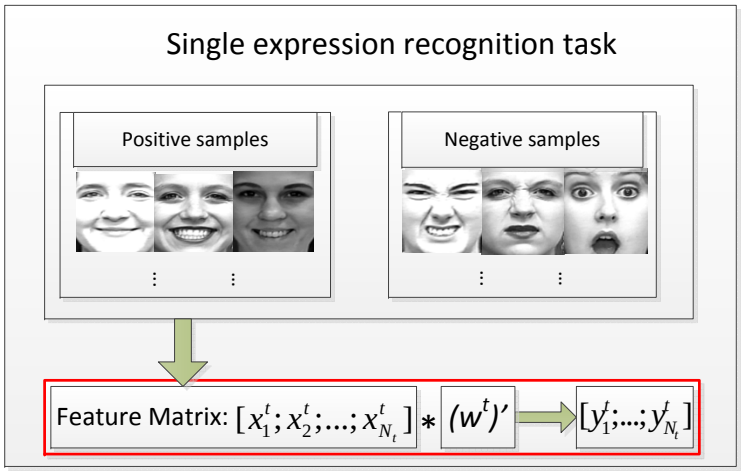


Figure 5.4: Illustration of one single expression task. Each task is a binary expression classification problem. Take Expression task of happiness for example here.

5.2.2 Learning Common Patches Across Expressions

Discovering the common patches across all the expressions is actually equivalent to learning the shared discriminative patches for all the expressions. Since Multi-task sparse learning (MTSL) can learn common representations among multiple related tasks [3], our problem can be transfered into a MTSL problem. T related tasks are defined as T discriminative patch learners for T facial expressions respectively (we set $T = 6$ for six basic expressions). Supposing each image has p patches, it can be represented by $(p \times m)$-dimensional LBP-based histogram features. Let $K = p \times m$. However, equation (1) cannot directly model our problem. Different from the MTSL model described in Equation(2.2), we focus on the selection of common patches instead of individual features. Since a group of consecutive features stand for one patch, and the number of common patches are not large, group sparsity prior can be assumed [113, 114]. Our problem is modeled as the following MTSL problem, in which the regularization term of Equation(2.2) is modified to a patch level sparse constraint:

$$\arg\min_{W} \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} J^t(w^t, x_i^t, y_i^t) + \lambda \sum_{j=1}^{p} \|w_{G_j}\|_2 \tag{5.1}$$

Here, w_{G_j} is a sub-matrix of matrix W, where G_j denotes the j-th patch, as shown in Figure 5.2. Figure 5.4 illustrates how to set up each task. In each task, images of one particular expression are considered as positive samples, while others are negative samples. This regularization term encourages the representation coefficients of the features in most patches to be

zero, and then the remaining non-zero patches indicate the shared important representation for all the expressions. The cost function of J^t is defined as a logistic loss function:

$$J^t(w^t, x_i^t, y_i^t) = ln(1 + exp(-y_i^t x_i^t \cdot w^t)). \tag{5.2}$$

To solve this patch-based multi-task sparse learning, the proposed algorithm is based on the accelerated gradient method proposed in [98]. The algorithm comprises two main steps: the *generalized gradient mapping step* and the *aggregation step*. The two steps alternately update two matrices in each iteration, i.e., a weight matrix sequence W_s and an aggregation matrix sequence V_s, s is the iteration index number. The updating of W_{s+1} is the *generalized gradient mapping step*, which uses the current aggregation matrix V_s to update matrix W_s. During this updating, we heuristically enforce the group sparsity prior which makes the representation coefficients of the features in one patch to be zeros under the condition in step 5-9 of Algorithm 1. The updating of V is the *aggregation step*, in which we construct a linear combination of W_{s+1} and W_s to update V_{s+1} (step 11 in Algorithm 1). The detailed problem solving procedure are summarized in Algorithm 1.

5.2.3 Learning Specific Patches For Individual Expression

Although learned common patches can discriminate all facial expressions, the performance could not be the best, because each expression also has its special properties besides the common properties. Here, we aim to explore some specific facial patches for each expression with the help of face verification, and then they are used to further boost the performance of common facial patches.

The motivation to employ the face verification task is that those special facial patches are important face regions, which are not only useful for recognize this expression, but also very significant for identifying the subjects. Take an expression e for example. Recognition e task will prefer to select out those patches which are useful only to recognizing the expression. Since we have the assumption that those patches should be the important face regions, and thus they are also very discriminative to face verification task, a multi-task sparse learning model can be used to couple these two tasks and select out those shared important patches between these two tasks more robustly. The learned patches should embed some specific signatures of

Algorithm 2 *Algorithm for learning common patches*

Input : Training data $\{(x_i^t, y_i^t), i = 1, .., N_t\}$, define $X^t = [x_1^t; ...; x_{N_t}^t], Y^t = [y_1^t; ...; y_{N_t}^t], V = [v^1; ...; v^T]$. t indicates the task index, and $t = 1, ..., T$. j is the group index, and $j = 1, ..., p$.
Initialize : W_0 takes equal weights, $V_0 = W_0$ and $a_0 = 1$. Tuning parameter λ and step size η.
for $s = 0...S$ **do**
  $w_{s+1}^t = v_s^t -$
  $\eta[\frac{1}{1+exp(-(Y^t)^{'} X^t v_s^t)} exp(-(Y^t)^{'} X^t v_s^t)(-(X^t)^{'} Y^t)]$
  if $\|w_{G_j,s+1}\|_2 \geq \lambda\eta$ **then**
    Set $w_{G_j,s+1} = (1 - \frac{\lambda\eta}{\|w_{G_j,s+1}\|_2}) w_{G_j,s+1}$
  else
    Set $w_{G_j,s+1} = 0$
  end if
  $a_{s+1} = \frac{2}{s+3}, \delta_{s+1} = W_{s+1} - W_s$
  $V_{s+1} = W_{s+1} + \frac{1-a_s}{a_s} a_{s+1}\delta_{s+1}$
  if $\|\delta_{s+1}\|_2 \leq \epsilon$ **then**
    break
  end if
end for
Normalization : $w^t = \frac{w^t}{\|w^t\|_2}$
$w_{G_j} = \sum w_k^t$, where w_{G_j} is the weight for patch j, and $w_k^t \in G_j$
Output : order w_{G_j} decreasingly, and output the top patches as the common patches for all expressions.

the face identity.

This multi-task sparse learning framework for specific patches is the same as the framework of learning the common patches except the different task design. The individual expression analysis task is organized in the same way as in Figure 5.4. Figure 5.5 illustrates how to organize the task of face verification. For face verification, we need to compare two images and label them as the same person or not, so we organize the training data of this task by the feature difference between two images. Assuming $(x_i^2, y_i^2)_{i=1}^{N_2}$ is the training set, x_i^2 is the feature difference between two images in i-th image pair. $y_i^2 \in \{-1, 1\}$ indicates whether the two images in i-th pair come from one subject or not. N_2 is the number of image pairs. The superscript 2 means this task is the second task in the multi-task sparse learning model. The procedure for solving this problem is the same with Algorithm 1. Because there are six expressions, six multi-task sparse learning models are needed to be built to learn their specific patches respectively.

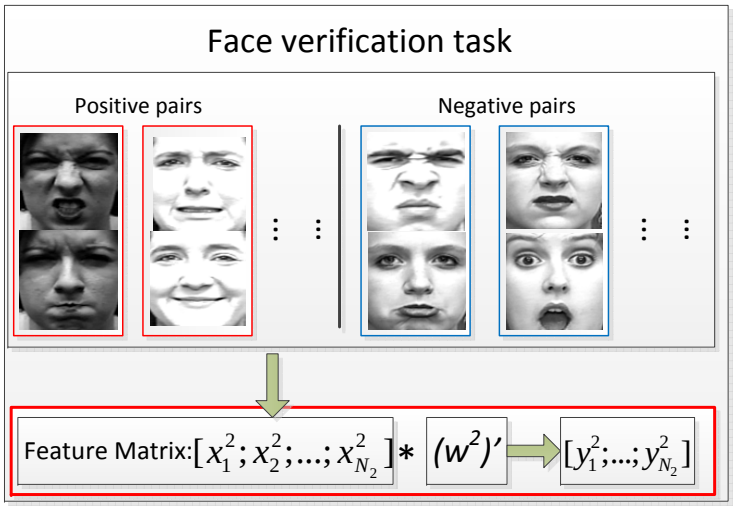


Figure 5.5: The design of Face Verification task. Image pairs from the same subject are considered as positive samples. Otherwise, as negative samples.

The specific patches have overlap with the learned common patches. Since the common patches will be used for all expressions, the overlapped patches are removed from the specific patches. The rest patches are considered as the final specific patches.

5.2.4 Classifier Design

With the extracted common and specific patch features based on the training data, classifiers are then built based on these features for testing data. Multi-task sparse learning model can directly give out classification results [109]. However, to fairly compare with previous work [59, 74], SVM is adopted to learn the expression classifiers and the one-against-all strategy is employed to decompose the six class problem into multiple binary classification problems. Each binary classification will output a confidence value of the test sample belonging to this class. The class label with the highest confidence will be the final classification result of this sample. The performances of common patches and the combination of common and specific patches are evaluated respectively. For expression e, denotes the common patches as P_c, and the specific patches as $\{P_s^e\}_{e=1}^6$. When both common and specific patches are investigated, the features from P_c and P_s^e are concatenated to represent facial images, and train the SVM classifiers; While only use the features of P_c when common patches are tested.

5.3 Experiments

We evaluate the learned common and specific patches for facial expression recognition. All methods are compared on three datasets, the Cohn-Kanade database [49], the MMI database [90] and the GEMEP-FERA [88], which are widely used for facial expression recognition algorithms. Our methods are denoted as CPL and CSPL respectively (see Table 5.1). To efficiently evaluate the performance of our proposed methods, they are compared with [74], which is the most recent comprehensive study on expression recognition with remarkable results. In [74], two methods are evaluated, denoted as ADL and AFL respectively. ADL uses Adaboost to select important patches and then performs SVM on the extracted LBP features of these patches. AFL uses all the patches to train the classifier without feature selection. MCPL, MCSPL, MADL and MAFL are the methods using multi-scale patches for CPL, CSPL, ADL and AFL, respectively. For fair comparison, all the methods are based on the same patch(sub-region) division strategy, same feature representation, and the same classification method (SVM). The only difference among the methods is the patches they use. All method abbreviations are listed in table 5.1. 10 folds cross-validation is employed for all methods.

Table 5.1: Method abbreviations.

ADL	only use patches selected by **AD**aboost are used.
AFL	**A**ll patches of the whole **F**ace are used.
MADL	only use **M**ulti-scale patches selected by **AD**aboost are used.
MAFL	**M**ulti-scale **A**ll patches of the whole **F**ace are used.
CPL	only use **C**ommon **P**atches. (our method)
CSPL	use **C**ommon and **S**pecific Patches. (our method)
MCPL	only use **M**ulti-scale **C**ommon **P**atches. (our method)
MCSPL	use **M**ulti-scale **C**ommon and **S**pecific Patches. (our method)

5.3.1 Experiments On the Cohn-Kanade Database

The Cohn-Kanade database consists of 100 university students aged from 18 to 30 years old, of which 65% were female, 15% were African-American and 3% were Asian or Latino. Subjects were instructed to perform a series of 23 facial displays, six of which were based on description

of prototypic emotions. For our experiments, image sequences are selected out from 96 subjects, whose sequences could be labeled as one of the six basic emotions. For each sequence, we only use the three peak frames with the most expressions. The faces are detected automatically by Viola's face detector [93], and then they are normalized to 96×96 as in Tian [79] based on the location of the eyes. Figure 5.7 shows some normalized samples with all expressions.

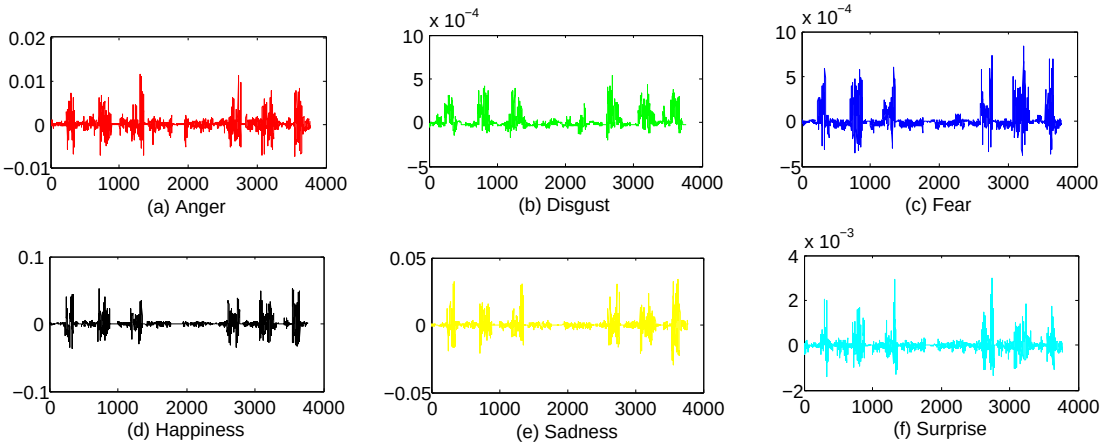
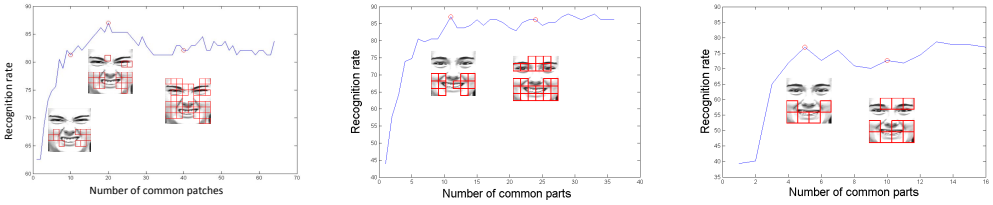


Figure 5.6: Results for six expressions in the coefficient matrix after multi-task sparse learning for learning the common patches. X-axis corresponds to the feature index in the coefficient matrix, where features index are ordered consecutively as group by patches. Y-axis is the weight values for features in each task after multi-task sparse learning. The non-zeros parts are grouped, and matches across all tasks.

Figure 5.7: Example of six basic expressions from the Cohn-Kanade database.(Anger, Disgust, Fear, Happiness, Sadness and Surprise).

To better demonstrate the patch selection strategy and the physical meaning of the selected patches, we first apply our method to only one patch scale ($S8$). The performance of the recognition can be further boosted by combining the patch selection across all different scales.



(a) (b) (c)

Figure 5.8: The expression recognition rate with different number of common patches. (a) The recognition result with selected common patches for the scale $S8$. The patch number for the three faces images marked with selected common patches are 10, 20, 40, respectively. (b) The results for $S6$. Patch number are 11, 24, respectively. (c) The results for $S4$. Patch number are 5, 10, respectively. All results show the most effective patches are around the mouth and the eyes, and using only one third of all the patches can achieve satisfied performance.

Analysis of Common Patches

As described in section 2.1, the proposed multi-task sparse learning aims to select the shared patches instead of the shared features, so we apply the L_1/L_2 norm regularization on the patch level to obtain patch-based group sparsity. Figure 5.6 reports the representation coefficient results for six expression tasks. We can see that the representation coefficients of features are sparse, and show the property of patch-based group sparsity. It is also clear to see the index correspondences for non-zero values across six expressions, which indicates the commonalities among them. So, this result demonstrates the effectiveness of our proposed algorithm in learning the shared common patches for expressions.

Before evaluating the recognition performance of the common patches, we want to inspect the performance when a different number of common patches is selected. Figure 5.8 reports the results with different number of the common patches. We can see that the recognition rate increases quickly with the first leading common patches, and when the number of the selected patches reaches around 20, it will get a recognition rate of 88.42%. If too many common patches are selected, the performance goes down slightly and fluctuates. It means that only some common patches are discriminative for all the expressions. When some patches with little importance are selected as the common patches, they will introduce some noises and influence the discriminative power of the common patches. We set the number of the common patches to be 20 in the following experiments. Figure 5.9 shows the superimposing effect of the selected

common patches over the 10 fold experiments. There are great overlaps between different fold experiments. It indicates that our algorithm is robust to the selection of the training set. The selected common patches are basically around the areas of mouth, eye, and eyebrows, which are consistent with AU-based analysis in FACS [31].

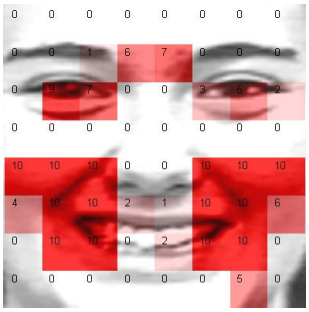


Figure 5.9: The distribution of selected common patches on faces. The darker the red color is, the more times (shown as numbers) the patch has been selected as common patches in 10-fold experiments.

Table 5.4 reports the detailed recognition performance of the common patches on each expression, where the expressions of anger, disgust, fear, happiness, sadness, surprise are denoted as ag, dg, fa, hp, sd, and sp for simplicity. Promising recognition rates are obtained on all the expressions except anger. Anger is often misclassified as sadness. This is because these two expressions have similar appearance variations on the common patches. This problem can be alleviated by adding some specific patches, which will be discussed next.

Analysis of Specific Patches

Although a rewarding recognition result can be obtained by only using the common patches, the performance can be further improved by integrating some specific patches of each expression. Figure 5.10 shows the top three learned specific patches for each expression based on the proposed multi-task learning. We can see the locations of these patches are highly related to expression types. Take surprise for example. The selected specific patches show the characteristics of surprise expression, in which special appearance changes are distributed in opened mouth, on stared eye, and raised eyebrow. In CSPL, the common patches and the specific patches are integrated together, and the experimental results are reported in Table 5.5. Compared to the results of CPL (Table 5.4), we can see that adding specific patches can further

Table 5.2: The confusion matrix of AFL on Cohn-Kanade database.(Measured by recognition rate: %)

	ag	dg	fa	hp	sd	sp
ag	**66.67**	7.5	0	0.83	25	0
dg	6.67	**87.67**	0.67	1.33	3.67	0
fa	3.73	1.43	**77.54**	10.40	6.90	0
hp	1.00	0.33	2.58	**95.42**	0.67	0
sd	10.60	1.25	2.87	0	**84.54**	0.74
sp	0	0	1.73	0	1.25	**97.02**

Table 5.3: The confusion matrix of ADL on Cohn-Kanade database.(Measured by recognition rate: %)

	ag	dg	fa	hp	sd	sp
ag	**64.72**	10.00	1.11	0	24.17	0
dg	5.33	**89.33**	2.00	1.33	2.00	0
fa	3.65	1.43	**78.57**	10.87	5.48	0
hp	1.00	0.33	3.24	**94.76**	0.67	0
sd	11.43	1.25	2.50	0	**84.40**	1.42
sp	0	0	1.73	0	0	**98.27**

improve the performance of the common patches.

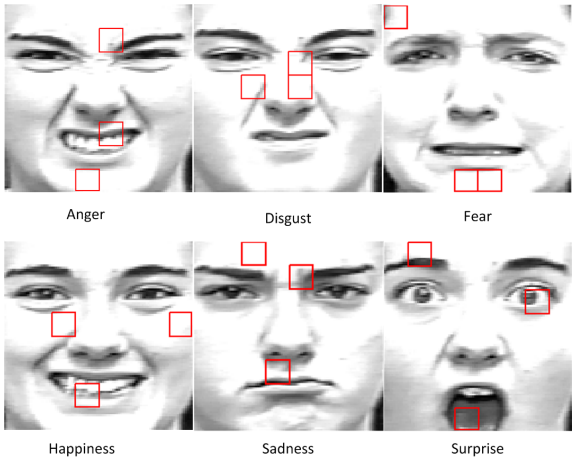


Figure 5.10: The top 3 specific patches for six expressions after eliminating the shared patches on the Cohn-Kanade database.

Comparisons with other methods

To further evaluate the proposed CPL and CSPL, we compare them to ADL and AFL developed in [74]. Table 5.6 lists the F1 measure for every expression and the overall recognition rates of these four methods. AFL gets the recognition rate of 86.94%, which is much worse

Table 5.4: The confusion matrix of CPL on the Cohn-Kanade database.(Measured by recognition rate: %)

	ag	dg	fa	hp	sd	sp
ag	**65.56**	8.33	0	0	25.28	0.83
dg	2.67	**92.67**	0.67	2	2	0
fa	0	1.98	**78.97**	13.25	5.79	0
hp	0.33	0.67	4.24	**94.76**	0	0
sd	6.20	1.67	3.33	0	**87.69**	1.11
sp	0	0	1.25	0	0.48	**98.27**

Table 5.5: The confusion matrix of CSPL on the Cohn-Kanade database.(Measured by recognition rate: %)

	ag	dg	fa	hp	sd	sp
ag	**71.3889**	7.5	0	0.83	19.44	0.83
dg	2.67	**95.33**	0	0	2	0
fa	0	2.46	**81.11**	10	6.43	0
hp	0.33	0.33	3.58	**95.42**	0.33	0
sd	7.45	1.25	2.92	0	**88.01**	0.37
sp	0	0	1.25	0	0.48	**98.27**

than our methods. It shows the importance of selecting discriminative patches. The confusion matrixes of methods, AFL and ADL, can also be found in Table 5.2 and Table 5.3. Although ADL also uses Adaboost to select the patches, it does not take the commonalities among all the expressions into account. ADL gets a recognition rate of 82.26% with the selected patches (highest rate with 20 ± 3 patches), while the recognition rates of CPL and CSPL are 88.42% and 89.89% respectively. It demonstrates that the learned common and specific patches by our proposed two-stage multi-task sparse learning can really improve the performance of expression recognition.

Analysis of multi-scale patches

The selected patches are shown to be effective in expression recognition, even the number of them is limited [118]. However, no prior of the optimal patch size is given. Besides, patches with different scales may represent different information for recognition. So, we can employ the patch selection strategy on different patch sizes, and then combine the selected patches with different scales together to further improve the recognition performance. In Figure 5.11, we show the selected common patches for all the scales for one of the 10-fold experiments.

Table 5.6: Recognition performances and F1 measures per expression for all compared methods(i.e., AFL, ADL, CPL, CSPL, MAFL, MADL, MCPL, MCSPL) on the Cohn-Kanade database.

	Single Scale				Multi Scale			
Expressions	AFL	ADL	**CPL**	**CSPL**	MAFL	MADL	**MCPL**	**MCSPL**
Anger	0.6407	0.6281	**0.7144**	**0.7440**	0.6512	0.6325	**0.7350**	**0.7628**
Disgust	0.8782	0.8776	**0.8927**	**0.9134**	0.8875	0.8796	**0.9105**	**0.9411**
Fear	0.8235	0.8206	**0.8209**	**0.8432**	0.8369	0.8316	**0.8462**	**0.8619**
Happiness	0.9416	0.9381	**0.9305**	**0.9462**	0.9536	0.9467	**0.9512**	**0.9635**
Sadness	0.8204	0.8346	**0.8515**	**0.8619**	0.8269	0.8362	**0.8645**	**0.8829**
Surprise	0.9806	0.9791	**0.9827**	**0.9870**	0.9806	0.9821	**0.9842**	**0.9870**
Recognition Rate	0.8694	0.8226	**0.8842**	**0.8989**	0.8732	0.8324	**0.9034**	**0.9153**

For each scale, only $1/3$ patches are selected out. The results show that for different scales, the effective patches are around the mouth and the eyes. As shown in Table 5.6, the performance of these multi-scale patches can achieve better performance (90.34%) than only using the common patches from a single scale $S8$ (88.42%). The F1 measures for each expression also validate the usefulness of multi-scale strategy. The specific patches are selected out at different scale individually, following the way we already shown for the scale $S8$. Incorporating selected specific patches, we get the best performance (91.53%). Thus, the experiment results show that multi-scale patches could contain more useful statistic information for recognition than one scale with a single division strategy. Our patches selection are effective in selecting the discriminative patches across all patch scales.

Figure 5.11: The distribution of selected common patches with different scales on faces. The darker the red color is, the more times the patch has been selected. To make the results visibly clearer, only one result of the 10-fold experiments is shown. The selected patches for all scale are around mouth and the eyes.

5.3.2 Results on the MMI database

The MMI database includes 30 students and research staff members aged from 19 to 62, of whom 44% are female, having either a European, Asian, or South American ethnic background. In this database, 213 sequences have been labeled with six basic expressions, in which 205 sequences are with frontal face. Different from [74], in which only the experimental data are collected from 99 selected sequences, we conduct our experiments on the data from all the 205 sequences. As in [74], the apex images are extracted from the sequences as the experimental data. Facial image are corpped based on locations of eyes, and resize it to 96×96, same as on Cohn-Kanada database.

MMI is a more challenging database than the Cohn-Kanade database. First, the subjects make expressions non-uniformly. Different people make the same expression in different ways. Second, some subjects wear accessories, such as glasses, headcloth, or moustache. Additionally, in some sequences, the apex frames are not with high expression intensity. All these factors will greatly degrade the recognition performance.

Table 5.7: The confusion matrix of CPL on MMI database.(Measured by recognition rate: %)

(%)	ag	dg	fa	hp	sd	sp
ag	**46.94**	25.83	1.11	1.11	25.00	0.00
dg	24.17	**46.67**	0	16.67	11.67	0.83
fa	2.22	7.78	**49.44**	13.33	22.22	5.00
hp	3.50	7.67	7.83	**70.67**	4.17	6.17
sd	22.50	7.50	7.78	11.39	**50.83**	0.00
sp	0.83	2.50	25.33	2.50	2.50	**66.33**

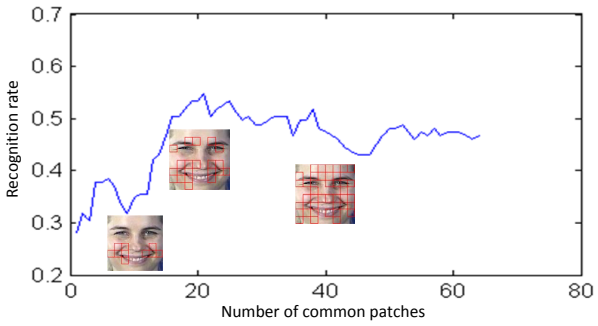


Figure 5.12: Recognition rate with different common patch number. Result of one fold experiment is shown.

Table 5.8: The confusion matrix of CSPL on the MMI database.(Measured by recognition rate: %)

	ag	dg	fa	hp	sd	sp
ag	**50.28**	10.56	5.56	2.50	28.61	2.50
dg	5.50	**79.83**	3.50	2.17	9.00	0
fa	1.67	4.13	**67.14**	15.56	8.97	2.54
hp	2.63	0.67	12.82	**82.91**	0.67	0.30
sd	16.34	2.87	13.98	4.54	**60.28**	1.99
sp	0.42	0	4.94	0.83	5.30	**88.51**

Table 5.9: F1 measures per expression and recognition performances for all compared methods(i.e., AFL, ADL, CPL, CSPL, MAFL, MADL, MCPL, MCSPL) on the MMI database.

	Single Scale				Multi Scale			
Expressions	AFL	ADL	**CPL**	**CSPL**	MAFL	MADL	**MCPL**	**MCSPL**
Anger	0.4584	0.4191	**0.4715**	**0.4949**	0.4851	0.4250	**0.4898**	**0.6568**
Disgust	0.4670	0.3890	**0.4821**	**0.7925**	0.4957	0.4062	**0.4939**	**0.725**
Fear	0.3593	0.3985	**0.4721**	**0.6143**	0.4236	0.4315	**0.4721**	**0.7259**
Happiness	0.6888	0.6572	**0.7022**	**0.8342**	0.7068	0.6984	**0.7365**	**0.8823**
Sadness	0.4771	0.5419	**0.4788**	**0.6311**	0.4874	0.5419	**0.5816**	**0.7109**
Surprise	0.6375	0.5760	**0.7496**	**0.9099**	0.6984	0.6253	**0.7554**	**0.9386**
Recognition rate	0.4774	0.4778	**0.4936**	**0.7353**	0.4924	0.4868	**0.5365**	**0.7739**

We first investigate the performance of the common patches with different patch number, and Figure 5.12 shows the results. It can be seen that the results are similar to the results on the Cohn-Kanade database. About 20 common patches are discriminative for all the expressions, so we set the number of the common patches as 20 on this database too. Table 5.9 lists the F1 measures for each expression and the overall recognition rates of CPL, CSPL, AFL, ADL, AFL, and their corresponding multi-scale methods respectively. Same as on the Cohn-Kanade data, CPL and CSPL are superior to AFL and ADL. However, the performances of all four methods are much lower than that of the Cohn-Kanade database, because this database has several challenging factors as mentioned above. CSPL obtains much better performance than CPL. This is because each expression has a very big variance due to the diversity of the subjects in this database, but the common patches cannot describe these specific variations. Although a much better result of 86.7% is reported in [74], their experimental data are carefully chosen 99 sequences, while we perform the experiments on all the 205 sequences. Besides, they adopt sliding and multi-scale windows to extract much more patches. We only divide the facial image into 64 patches, and we also obtain a recognition rate of 73.53% on more than double size of the

data than [74]. Table 5.7 and Table 5.8 list the confusion matrix of CPL and CSPL, respectively.

We further explore the common and specific patches selection for different patch sizes. The accuracy rate of the baseline method MAFL and MADL increases with more multi-scale patches. The recognition rate of MCPL and MCSPL using more multi-scale patches both achieve better performance than CPL and CSPL with single patch scale. The detail results are also shown in Table 5.9.

The experimental results indicate the location of learned common and specific patches, which confirms the previous knowledge about active facial parts in psychology. The rewarding performances of these patches in facial expression recognition provide a solid basis for patches selection and weight setting in similar applications. Our work opens the road for the researches of utilizing the prior knowledge of facial muscles in psychology, and further improves the performances of existing methods in computer vision.

5.3.3 Experiments On the GEMEP-FERA2011 database

The GEneva Multimodal Emotion Portrayals (GEMEP) is a collection of audio and video recordings [4]. It consists of over 7000 audiovisual emotion portrayals, representing 18 emotions portrayed by 10 actors trained by a professional director. The GEMEP-FERA2011 contains both AU sub-challenge and Emotion sub-challenge. Our method will focus on the Emotion sub-challenge, in which a total of 289 portrayals are selected (155 for training and 134 for testing). The training set included 7 actors with 3 to 5 instances of each emotion per actor. The test set includes 6 actors, half of which were not present in the training set [88, 89]. For each video in training or testing, five frames are evenly extracted from the video. We use the landmark detection from Zhou *et al.* [119] to localize the position of the eyes, based on which we align and crop out the face images. The landmark detection result and cropped images are shown in Figure 5.13.

The emotion recognition challenge involves the classification of the following 5 emotions: anger, fear, joy, relief and sadness. When determining the label of the test video, we first classify the extracted frames individually using a five-way forced strategy, then the emotion class which obtains the highest score will be the winner of the sequence.

Since the subjects in training part of database are quite few, it is quite difficult to learn the

(a) (b) (c)

Figure 5.13: (a) The landmark detection result of [119]. (b, c) Aligned and cropped face image examples from the GEMEP-FERA2011 Database.

common patches and the specific patches on this database itself. In this experiment, we utilize the knowledge of the learned patches from the CK database. It also shows the capability of generalization across different databases. We follow the training/testing partition in the database, and the performances of our methods are compared with some related works in Table 5.10. Our method MCSPL achieve much better recognition rate than the baseline work, i.e., Valstar *et al.* [88] and Chew *et al.* [17]. These results that the patches selected out by our algorithm are discriminative for expression recognition and are also robust across databases. Compared to the 1st prize of 2011 FERA competition, i.e., Senechal *et al.* [72], which gets 83.5% accuracy, our method MCSPL achieves 80.0%. Senechal *et al.* [72] utilize a lot of features, such as LGPB histograms, 2.5 D Active Appearance model to combine appearance and geometry information. It also employs complex classification algorithm, such as, Multi-kernels SVMs, temporal filtering. However, our method just uses the LBP feature and naive SVM. Considering these factors, our patch learning method is proved to be effective, and the results of our method are promising. More detailed comparisons of using different patches are given in Table 5.12.

Table 5.10: The classification rate for emotion detection on the GEMEP-FERA2011 database.

Emotion	baseline [88]	Chew *et al.* [17]	Senechal *et al.* [72]	**MCSPL**
Anger	0.89	0.26	0.963	**0.926**
Fear	0.20	0.40	0.640	**0.640**
Joy	0.71	0.52	0.968	**0.936**
Relief	0.46	0.88	0.846	**0.654**
Sadness	0.52	0.92	0.760	**0.800**
Overall	0.56	0.60	0.835	**0.800**

Table 5.11: Confusion matrix of MCSPL for emotion recognition on the overall test set of GEMEP-FERA2011 database.)

pred truth	Anger	Fear	Joy	Relief	Sadness
Anger	25	4	1	2	3
Fear	2	16	1	0	0
Joy	0	4	29	4	0
Relief	0	0	0	17	2
Sadness	0	1	0	3	20

Table 5.12: F1 measures per expression and recognition performances for all compared methods(i.e., AFL, ADL, CPL, CSPL, MAFL, MADL, MCPL, MCSPL) on the GEMEP-FERA2011 database.

	Single Scale				Multi Scale			
Expressions	AFL	ADL	**CPL**	**CSPL**	MAFL	MADL	**MCPL**	**MCSPL**
Anger	0.5214	0.4628	**0.6942**	**0.7187**	0.5424	0.5024	**0.7858**	**0.8065**
Fear	0.4952	0.4156	**0.5246**	**0.5500**	0.5604	0.4695	**0.6729**	**0.7273**
Joy	0.7608	0.6815	**0.8264**	**0.8529**	0.7428	0.7261	**0.8115**	**0.8529**
Relief	0.7254	0.6358	**0.7592**	**0.7626**	0.7068	0.6424	**0.7188**	**0.7756**
Sadness	0.6892	0.5703	**0.7325**	**0.7600**	0.7451	0.6481	**0.7954**	**0.8163**
Recognition rate	0.6791	0.5824	**0.7241**	**0.7463**	0.6958	0.6251	**0.7751**	**0.8000**

5.4 Conclusions

In this chapter, a new method to analyze facial expressions is proposed. Different from previous work, we aimed at exploring the commonalities among the expressions by discovering the common and specific patches. A two-stage sparse learning model is proposed to learn the locations of these patches based on the prior knowledge of facial muscles and AUs. A multi-scale face division strategy is employed to obtain facial patches with different coverage area and eliminate the side effects from fixed patch size. The effectiveness of these patches are evaluated by facial expression recognition. Extensive experiments show that common patches can generally discriminate all the expressions, and the recognition performance can be further improved by integrating specific patches. More comprehensive patches can also be selected out to achieve better performance by using multi-scale patch division strategy. The learned location information of these patches also confirms the location knowledge of facial muscles in psychology.

Chapter 6

Conclusions and Future Work

This dissertation improves the blind deconvolution methods in deblurring natural images and face images. Because of the unavoidable noises involved in the capturing time in practice, most existing single image deblurring methods fail without considering them. We proposed a directional filter to handle the noise in the deblurring process. Without introducing new side effects from denosing, the directional filter could remove the noise, and also keep the blur information intact in the orthogonal direction. These partially correct blur information can be further used to reconstruct the real 2-D blur kernel. Based on this observation, We proposed the noise-aware kernel estimation algorithm to accurately estimate the motion blur from blurry and noisy image. After the kernel is estimated, we proposed a final non-blind deconvolution method to restore the latent sharp image from severely noisy image. Extensive experiments on both synthetic and real images show that our method outperforms the existing methods on noisy and blurry images.

The faces are the most important areas in images, but quite few research has been done for the face deblurring. Most existing methods are designed mainly for natural images, and they are not suitable for dealing with face images because face images neither contain many salient edges, nor follow the priors learned from natural images. With the facial landmark localization result, we could incorporate the global face structure information, which would provide guidance for the following salient edge detection. The proposed face deblurring method is shown to be robust on blurry face images, and outperforms the existing state-of-the-art methods.

Many face related applications can perform better on restored sharp images than blurry face images, such as face identification and facial expression analysis. To further improve the facial expression recognition performance, we proposed to extract the active facial patches using multitask sparse learning methods. With the learned common and specific facial patches, the

proposed method can achieve better recognition rate on various databases in the experiments.

Our proposed methods also contain some defects, and solving them could be our future research directions. First of all, the proposed single image deblurring method employs heavy noise suppress regulation. So, the final restored images contain much over-smoothing and some color ringing artifacts. In the future, we would explore the solution to handle the noise without heavy smooth regulations.

Secondly, the proposed face image deblurring method relies on the facial landmark localization heavily. When the blur kernel size increases too large, the landmark localization method will fail and provide erroneous information for the deblurring steps. In such situations, our deblurring method will fail as well. As for the future work, our research attention would focus on how to improve the robustness of the facial landmark localization method on images with large blur kernel size, and how to automatically correct the erroneous landmark information in the deblurring steps.

References

[1] http://en.wikipedia.org/wiki/Facial_expression.

[2] Neatimage. http://www.neatimage.com/.

[3] A. Argyriou and T. Evgeniou. Multi-task feature learning. *Neural Information Processing Systems*, 2007.

[4] T. Bänziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *In K. R. Scherer, T. Banziger, E. B. Roesch (Eds.), Blueprint for affective computing: A sourcebook. Oxford, England: Oxford university Press.*, pages 271–294, 2010.

[5] M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 1999.

[6] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. *International Conference on Computer Vision and Pattern Recognition*, 2, 2005.

[7] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. *International Conference on Automatic Face and Gesture Recognition*, 2006.

[8] A. Blake and M. Isard. Active shape models. *Springer*, 1998.

[9] A. Buades, C. B., and J.-M. Morel. The staircasing effect in neighborhood filters and its solution. *Transactions on Image Processing*, 15(6), 2006.

[10] A. Buades, B. coll, and J. Morel. A non-local algorithm for image denoising. *International Conference on Computer Vision and Pattern Recognition*, 2005.

[11] E. Candes, J. romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transaction on Information Theory*, 52(2):489–509, 2006.

[12] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Conference on Computer Vision and Pattern Recognition*, 2012.

[13] R. Caruana. Multi-task learning. *Machine Learning*, 28:41–75, 1997.

[14] Y. Chang, C. Hu, R. feris, and M. Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006.

[15] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[16] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001.

[17] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. *International*

Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pages 915–920, 2011.

[18] H. Cho, J. Wang, and S. Lee. Text image deblurring using text-specific properties. *European Conference on Computer Vision*, 2012.

[19] S. Cho and S. Lee. Fast motion deblurring. *SIGGRAPH ASIA*, 2009.

[20] S. Cho, J. Wang, and S. Lee. Handling outliers in non-blind image deconvolution. *International Conference on Computer Vision*, 2011.

[21] T. S. Cho, S. Paris, B. K. P. Horn, and W. T. Freeman. Blur kernel estimation using the radon transform. *International Conference on Computer Vision and Pattern Recognition*, 2011.

[22] J. F. Cohn. Foundations of human computing: Facial expression and emotion. *International Conference on Multimodal Interfaces*, pages 223–238, 2006.

[23] J. F. Cohn, L. Reed, Z. Ambadar, J. Xiao, and T. Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. *International Conference on Systmes, Man and Cybernetics*, pages 610–616, 2004.

[24] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:1–12, 2004.

[25] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *European Conference on Computer Vision*, 1998.

[26] T. Cootes, M. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. *European Conference on Computer Vision*, 2012.

[27] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 2007.

[28] T. Darrell and K. Wohn. Depth form focus using a pyramid architecture. *International Conference on Computer Vision and Pattern Recognition*, 1988.

[29] T. G. Dietterich, L. Pratt, and S. Thrun. Special issue on inductive transfer. *Machine Learning*, 28(1), 1997.

[30] G. Duchenne. *Mecanisme de la Physionomie Humaine*. 1862.

[31] P. Ekman, W. V. Friesen, and J. C. Hager. Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press*, 1, 2002.

[32] I. Essa and A. Pentland. A vision system for observing and extracting facial action parameters. *International Conference on Computer Vision and Pattern Recognition*, pages 76–83, 1994.

[33] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 2003.

[34] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman. Removing camera shake from a single photograph. *SIGGRAPH*, 2006.

[35] M. Figueiredo, R. Nowak, and S. Wright. gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):587–597, 2007.

[36] G.Obozinski. Multi-task feature selection. *Technical Report. Department of Statistics, UC Berkeley*, 2006.

[37] A. Goldstein and R. Fattal. Blur-kernel estimation from spectral irregularities. *European Conference on Computer Vision*, 2012.

[38] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-pie. *International Conference on Automatic Face and Gesture Recognition*, 2008.

[39] P. Grossmann. Depth from focus. *Pattern Recognition Letters*, 1987.

[40] G. Guo and C. R. Dyer. Learning from examples in the small sample case - face expression recognition. *IEEE Transactions Systems, Man, and Cybernetics- Part B*, 35(3):477 –488, 2005.

[41] Z. Hu and M.-H. Yang. Good regions to deblur. *European Conference on Computer Vision*, 2012.

[42] J. Huang, X. Huang, and D. Metaxas. learning with dynamic group sparsity. *International Conference on Computer Vision and Pattern Recognition*, pages 64–71, 2009.

[43] R. A. Hummel, B. Kimia, and S. W. Zucker. Deblurring gaussian blur. *Computer Vision, Graphics, and Image Processing*, 1987.

[44] C. E. Izard. The face of emotion. *New York: Appleton-Century-Crofts*, 1, 1971.

[45] J. Jia. Single image motion deblurring using transparency. *International Conference on Computer Vision and Pattern Recognition*, 2007.

[46] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time. *International Conference on Automatic Face and Gesture Recognition*, pages 314–321, 2011.

[47] N. Joshi, R. Szeliski, and D. J. Kriegman. Psf estimation using sharp edge prediction. *International Conference on Computer Vision and Pattern Recognition*, 2008.

[48] N. Joshi, C. L. Zitnicky, R. Szeliskiy, and D. J. Kriegman. Image deblurring and denoising using color priors. *International Conference on Computer Vision and Pattern Recognition*, 2009.

[49] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *International Conference on Automatic Face and Gesture Recognition*, 2000.

[50] A. K. Katsaggelos. Digital image resotration. *Springer-Verlag New York, Inc.*, 1991.

[51] B. Kim. *Numerical Optimization Methods for Image Restoration*. PhD thesis, Stanford University, 2002.

[52] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale l1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606–617, 2007.

[53] S. Y. Kim, Y. W. Tai, S. J. Kim, M. S. Brown, and Y. Matsushita. Nonlinear camera response functions and image deblurring. *International Conference on Computer Vision and Pattern Recognition*, 2012.

[54] T. H. Kim, B. Ahn, and K. M. Lee. Dynamic scene deblurring. *International Conference on Computer Vision*, 2013.

[55] D. Krishnan, T. Tay, and R. Fergus. Blind deconvolution using a normalized sparsity measure. *International Conference on Computer Vision and Pattern Recognition*, 2011.

[56] A. Levin, R. Fergus, F. Durand, and W. T. Freeman. Image and depth from a conventional camera with a coded aperture. *SIGGRAPH*, 2007.

[57] A. Levin, Y. Weiss, f. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. *International Conference on Computer Vision and Pattern Recognition*, 2009.

[58] A. Levin, Y. Weiss, f. Durand, and W. T. Freeman. Efficient marginal likelihood optimization in blind deconvolution. *International Conference on Computer Vision and Pattern Recognition*, 2011.

[59] G. Littlewort, M. S. Bartlett, J. S. I. Fasel, and J. Movellan. Dynamics of facial expression extracted automatically from video. *International Conference on Computer Vision and Pattern Recognition*, 2004.

[60] M. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357 –1362, 1999.

[61] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transaction on Signal Processing*, pages 3397–3415, 1993.

[62] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971 –987, 2002.

[63] J. Pan, Z. Hu, Z. Su, and M.-H. Yang. Deblurring face images with exemplars. *European Conference on Computer Vision*, 2012.

[64] M. Pantic and T. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments form face profile image seqences. *IEEE Transactions Systems, Man, and Cybernetics- Part B*, 36(2):433–449, 2006.

[65] M. Pantic and L. J. M. Rothkrantz. Case-based reasoning for user-profiled recognition of emotions from face images. *International Conference on Multimedia and Expo*, pages 391–394, 2004.

[66] A. Rav-Acha and S. Peleg. Two motion-blurred images are better than one. *Pattern Recognition Letters*, 2005.

[67] W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of Ameria*, 1972.

[68] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. D. la Torre, and A. Rossi. Automated facial expression recognition system. *International Carnahan Conference on Security Technology.*, pages 172 –177, 2009.

[69] G. Sandbach, S. Zafeiriou, and M. pantic. Binary pattern analysis for 3d facial action unit detection. *British Machine Vision Conference*, pages 1–12, 2012.

[70] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. Recognition of 3d facial expression dynamics. *Image Vision Comput.*, 30(10):762–773, 2012.

[71] N. Sebe, M. S. Lew, I. Cohen, Y. Sun, T. Gevers, and T. S. Huang. Authentic facial expression analysis. *International Conference on Automatic Face and Gesture Recognition*, 2004.

[72] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Baily, and L. Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):993–1005, 2012.

[73] C. Shan. Smile detection by boosting pixel differences. *IEEE Transactions on Image Processing*, 21(1):431–436, 2012.

[74] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27:803–816, 2009.

[75] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. *SIGGRAPH*, 2008.

[76] C.-T. Shen, W.-L. Hwang, and S.-C. Pei. Spatially-varying out-of-focus image deblurring with l1-2 optimization and a guided blur map. *International Conference on Computer Vision and Pattern Recognition*, 2012.

[77] L. Sun, S. Cho, J. Wang, and J. Hays. Good image priors for non-blind deconvoluton: Generic vs specific. *European Conference on Computer Vision*, 2014.

[78] Y. Tai and S. Lin. Motion-aware noise filtering for deblurring of noisy and blurry images. *International Conference on Computer Vision and Pattern Recognition*, 2012.

[79] Y. Tian. Evaluation of face resolution for expression analysis. *International Conference on Computer Vision and Pattern Recognition*, jun. 2004.

[80] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing upper face action units for facial expression analysis. *International Conference on Computer Vision and Pattern Recognition*, 2000.

[81] Y. Tian, T. Kanade, and J. F. Cohn. Recognizing action unites for facial expression analysis. *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 2001.

[82] P. Toft. *The Radon Transform - Theory and Implementation*. PhD thesis, Technical University of Denmark, 1996.

[83] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. *International Conference on Computer Vision*, 1998.

[84] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):258–273, 2010.

[85] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[86] J. A. Tropp, Anna, and C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transaction on Information Theory*, 53:4655–4666, 2007.

[87] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. *International Conference on Computer Vision and Pattern Recognition*, 2006.

[88] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. *International Conference Automatic Face and Gesture Recognition.*, pages 921–926, 2011.

[89] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics.*, 42(4):966–979, 2012.

[90] M. F. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. *International Conference on Language Resources and Evaluation*, 2010.

[91] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Foundations of human computing: Facial expression and emotion. *International Conference on Multimodal Interfaces*, pages 162–170, 2006.

[92] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 31(1):1743 –1759, 2009.

[93] P. Viola and M. Jones. Robust real-time object detection. *Int. Journal of Computer Vision*, 2001.

[94] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan. Automated drowsiness detection for improved driver safety comprehensive databases for facial expression analysis. *International Conference on Automative Technologies*, 1, 2008.

[95] X. Wang, C. Zhang, and Z. Zhang. Boosted multi-task learning for face verification with applications to web image and video search. *International Conference on Computer Vision and Pattern Recognition*, 2009.

[96] J. Whitehill and C. W. Omlin. Haar features for facs au recognition. *International Conference on Automatic Face and Gesture Recognition*, 2006.

[97] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(2):210–217, 2009.

[98] C. Xi, P. Weike, J. T. Kwok, and J. G. Carbonell. Accelerated gradient method for multi-task sparse learning problem. *International Conference on Data Mining*, 2009.

[99] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. *International Conference on Computer Vision and Pattern Recognition*, 2013.

[100] L. Xu and J. Jia. Two-phase kernel estimation for robust motion deblurring. *European Conference on Computer Vision*, 2010.

[101] L. Xu, C. Lu, Y. Xu, and J. Jia. Image smoothing via l0 gradient minimization. *ACM Transactions on Graphics*, 2011.

[102] L. Xu, S. Zheng, and J. Jia. Unnatural l0 sparse representation for natural image deblurring. *International Conference on Computer Vision and Pattern Recognition*, 2013.

[103] J. Yang, Z. Lin, and S. Cohen. Fast image super-resolution based on in-place example regression. *International Conference on Computer Vision and Pattern Recognition*, 2010.

[104] P. Yang, Q. Liu, and D. N. Metaxas. Exploring facial expressions with compositional features. *International Conference on Computer Vision and Pattern Recognition*, 2010.

[105] S. Yi, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. *International Conference on Computer Vision and Pattern Recognition*, 2012.

[106] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. *International Conference on Automatic Face and Gesture Recognition*, pages 211–216, 2006.

[107] Y. Yitzhaky, I. Mor, A. Lantzman, and N. S. Kopeika. Direct method for restoration of motion-blur images. *Journal of the Optical society of America*, 1998.

[108] L. Yuan, J. Sun, L. Quan, and H. Y. Shum. Progressive inter-scale and intra-scale non-blind image deconvolution. *ACM Transactions on Graphics*, 2008.

[109] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. *International Conference on Computer Vision and Pattern Recognition*, 2010.

[110] S. Zafeiriou and I. Pitas. Discriminant graph structures for facial expression recognition. *IEEE Transaction on Multimedia.*, 10(8):1528–1540, 2008.

[111] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2007.

[112] L. Zhang, A. Deshpande, and X. Chen. Denoising vs. deblurring: Hdr imaging techniques using moving cameras. *International Conference on Computer Vision and Pattern Recognition*, 2010.

[113] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. Metaxas. Automatic image annotation using group sparsity. *International Conference on Computer Vision and Pattern Recognition*, 2010.

[114] S. Zhang, J. Huang, H. Li, and D. Metaxas. Automatic image annotation and retrieval using group sparsity. *IEEE Transactions Systems, Man, and Cybernetics- Part B*, 2012.

[115] S. Zhang, Y. Zhan, M. Dewan, J. Huang, and D. Metaxas. Towards robust and effective shape modeling: Sparse shape compositon. *Medical Image Analysis*, 16(1):265–277, 2012.

[116] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.

[117] L. Zhong, S. Cho, D. Metaxas, S. Paris, and J. Wang. Handling noise in single image deblurring using directional filters. *International Conference on Computer Vision and Pattern Recognition*, 2013.

[118] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. Metaxas. Learning active facial patches for expression analysis. *International Conference on Computer Vision and Pattern Recognition*, pages 2562–2569, 2012.

[119] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional neural network. *ICCV workshop on 300 Faces in-the-Wild Challenge.*, 2013.

[120] D. Zoran and Y. Weiss. From learning models of natural image patches to whole image restoration. *International Conference on Computer Vision*, 2011.