A STUDY ON MOTION BLUR

A Project Report

submitted by

ABHIJEET SHENOI

in partial fulfilment of the requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

May 2017

THESIS CERTIFICATE

This is to certify that the thesis titled **A STUDY ON MOTION BLUR**, submitted by **Abhijeet Shenoi**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. A.N. Rajagopalan
Research Guide
Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai

Date:

ACKNOWLEDGEMENTS

Thank you, first and foremost to Prof. A.N. Rajagopalan, without whose undying support and trust, none of this would have been possible. Prof. Kaushik Mitra's guidance kept me motivated and helped me pursue the area of computer vision and image processing. Also, a special mention to Subeesh Vasu, whose help went a long way in completing much of the work in this thesis.

ABSTRACT

Modern photography is a marvel, capturing moments for posterity, and entire scenes in the form of matrices. With the advent of light weight cameras, camera shake is a problem which plagues most photographic endeavours. In addition to this, capturing of rapidly moving objects is a challenge. Advent of sensors which work at higher frame rates and shutters which open and close in fractions of milliseconds has helped solve these problems, however, in cases where there is not enough light, these approaches are rendered infeasible. Post processing is a cheap and rather effective alternative. Many techniques exist to tackle motion blur. A large fraction of these however consider only translational motion, which is space invariant. Another class of techniques exist which handle space variant blur. In this work, an attempt to handle the problem of motion blur across various domains is made. Motion blur in samples of video which can be used to perform gait based personal identification is one such problem which was explored. Estimating the length and direction of the blur kernel by established methods was dealt with, and a deep learning approach to solve the same problem was also taken up. The work then attempted to perform activity recognition, and deblurring of objects exhibiting articulated motion. Finally, HDR image deblurring with simultaneous super resolution was taken up and a fully functioning algorithm was developed and tested.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

IITM	Indian Institute of Technology, Madras
SR	Super Resolution
LR	Low Resolution
HR	High Resolution
HDR	High Dynamic Range
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
PSF	Point Spread Function
CRF	Camera Response Function

CHAPTER 1

INTRODUCTION

What is motion blur? Motion blur is the apparent streaking of rapidly moving objects in a still image or a sequence of images such as a movie or animation. It results when the image being recorded changes during the recording of a single exposure, either due to rapid movement or long exposure. Unlike defocus blur, whose effects are limited to a single PSF whose scale varies with depth, motion blur can present itself in a large number of ways. The number of PSF's which can result from a camera movement are boundless. Deblurring is therefore, an inherently ill posed problem. Natural image statistics help establish priors which allow for better solutions, but these do not always result in better output. With the advent of deep learning, many new techniques have come to the fore. The problem of deblurring, despite of these new techniques, is still open to better solutions. This work, analyses possible solutions to a small subset of the problems, and works on addressing pertinent issues.

1.1 Contributions

Of the work in this thesis, my contributions include the following: A full and thorough literature survey for gait based personal identification. This was followed by an exploration of the effect of motion blur in the gait identification pipeline. Artificial intelligence was also explored, with a thorough reading of deep learning, and the pertinenet techniques involved. MatConvNet, a package for implementing deep networks was learnt and used. A deep neural network was fully trained and implemented to identify the length and direction of motion blur from the alpha matte of a moving object. Once this was done, a framework for activity recognition was conceived, including a new feature vector, which involves a direction binned histogram. Benchmarking activities for a siamese neural network were carried out on 3 datasets, while ensuring uniform PSNR calculation, and measures to ensure thorough, competent benchmarking. Further, an algorithm was developed and fully implemented to carry out simultaneous super resolution and motion deblurring for images in the irradiance domain.

CHAPTER 2

MOTION BLUR

2.1 An Investigation of Gait Based Human Identification

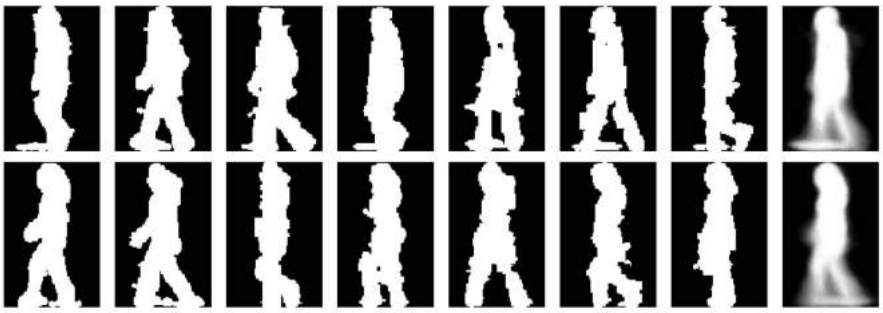


Figure 2.1: Gait sequence in which the rightmost image is the GEI

A large part of initial efforts were devoted to understanding approaches to gait recognition. Man and Bhanu (2006) was a seminal work on the topic, and most works such as Lam *et al.* (2011), Zhang *et al.* (2010), Chen *et al.* (2009), Liu and Zheng (2007) and Wang *et al.* (2012) are more or less modifications of this approach. To succinctly state the general method involved, we can say that the person in question is extracted from the background, either by means of a binary silhouette or an alpha matte as a first step. To augment data for learning, certain operations were carried out on the databasr of silhouettes, such as skewing and translations. The second step involves using these silhouettes to generate a single gait feature, which represents the gait cycle of that individual. These features are subsequently used for classification using a combination of PCA, LDA and SVM's, or other probabilistic classifiers. A deep learning approach has not yet been identified, to the best of my knowledge. In Man and Bhanu (2006), the approach used is a simple averaging of frames. The GEI is hence given by Equation 2.1.

$$G(x,y) = \frac{1}{N}\sum_{t=1}^{N} B_t(x,y) \tag{2.1}$$

where B_t is the binary silhouette. The justification for use of this representation is that the effect of noisy silhouettes is reduced. Once this has been done, a combination of PCA and MDA is applied. A novel approach used for classification here is the generation of a synthetic gait sequence. This is done by deleting the lowest portion of the GEI and resizing the resulting image, to provide robustness. Classification is done as follows: Given a probe sequence P, real gait templates $R_j, j = 1, ..., n_R$ and synthetic gait templates $S_j, j = 1, ..., n_S$ are generated. The transformation matrix obtained by the PCA, MDA combination is denoted by $\hat{T}_r$ and $\hat{T}_S$ for the real and synthetic images respectively. Hence, the feature vectors formed from the probe sequence are given by $\hat{R}_p : \hat{r_j} = T_r R_j, j = 1, ..., n_R$ and $\hat{S}_p : \hat{s_j} = T_r S_j, j = 1, ..., n_S$. The classification is done using a modified nearest neighbour approach. The distance measure is defined in Equation 2.2.

(a) Single blurred frame

(b) Three blurred frames

(c) Five blurred frames

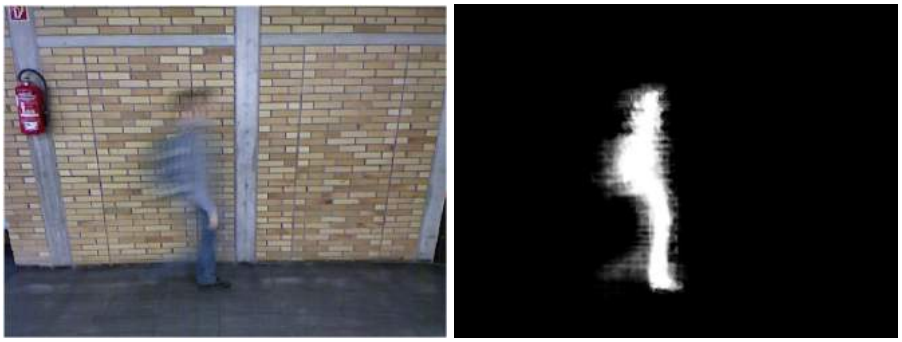
Figure 2.2: Extracted matte from blurred frames

$$D(\hat{R_P}, R_i) = \frac{1}{n_R} \sum_{j=1}^{n_R} \parallel \hat{r_j} - m_{ri} \parallel, i = 1, ..., c \tag{2.2}$$

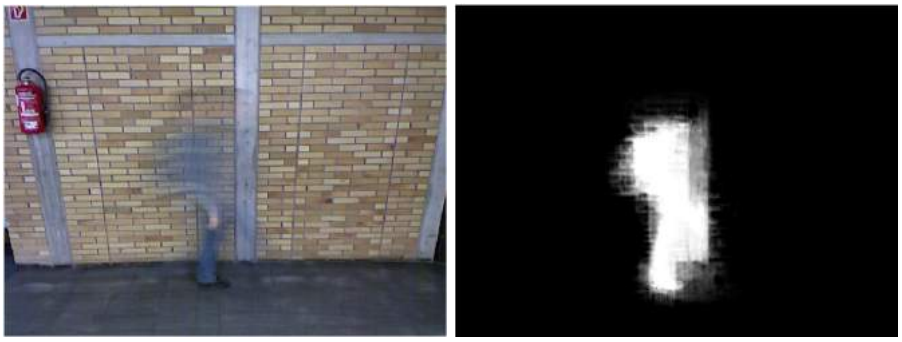
where m_{ri} is the mean of real feature vectors in class i.

(a) Single frame averaged

(b) Three frames averaged



(c) Five frames averaged



(d) Ten frames averaged

Figure 2.3: Matte from blurred frames

A similar measure is defined for the synthetic images. A combination of these two distances is used as a distance metric when using nearest neighbour classification. As the motion information is represented primarily at edges and boundaries, it is logical to check whether all necessary gait information is embedded in the alpha matte of the person. We also wanted to go ahead and see if this problem can be expanded in scope, by using blurred input frames to extract the matte. To generate these blurred frames, temporal averaging was carried out. If $F_i(x, y)$ denotes each frame of the gait cycle, the blurred frames $B_j(x, y)$ were calculated as follows: $B_j(x, y) = \sum_{i=1}^{p} F_i(x, y)$ where p is the number of frames to be averaged. We found that a reliable matte can only be extracted with $20 - 30\%$ of the gait cycle being averaged, on standard databases such as the USF gait database. Matting was carried out using Levin *et al.* (2008). Further averaging led to a poor quality alpha matte, and hardly any discernible blur. These results are shown in Figures 2.4 and 2.3. To further this approach of using blurry frames for gait identification, it seemed pertinent to find the length and direction of motion blur at each contour point of the alpha matte. When focus is restricted to small patches within an image, motion can be approximated to be linear. Hence, in these patches, finding the length and direction of motion is enough to fully evaluate the blur kernel. For this approach, and all subsequent approaches, we only worked with images of sufficiently high resolution. To carry out this kernel estimation, initially, two blind deconvolution techniques were used.Xu and Jia (2010) and Pan *et al.* (2014) were the two approaches which were used.

In Xu and Jia (2010), the energy minimisation carried out is given in Equation 2.3.

$$E(k) = \| \nabla I^s \otimes k - \nabla B \|^2 + \gamma \| k \|^2 \tag{2.3}$$

This enforces the L_2 norm on the kernel. Once this edge map of the latent image is obtained (via k), the edge map is used as a spatial prior to estimate the latent image. Another optimisation is carried out with Equation 2.4.

$$E(I) = \| I \otimes k - B \|^2 + \lambda \| \nabla I - \nabla I^s \|^2 \tag{2.4}$$

The kernel obtained is then refined. Pan *et al.* (2014) was applied wherein the prior is slightly modified to be suited for text applications. The energy fuction to be minimised in this case was $E = \| x \otimes k - y \|_2^2 + \gamma \| k \|_2^2 + \lambda P(x)$ where $P(x)$ is a regularisation

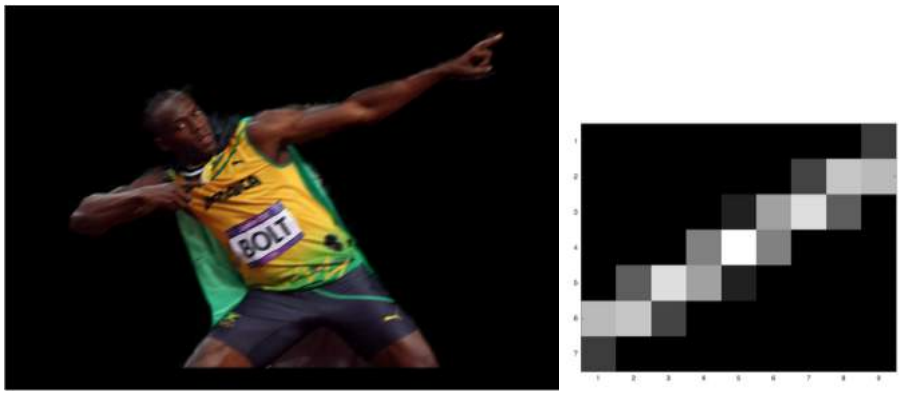


Figure 2.4: Synthetically blurred input with associated ground truth kernel

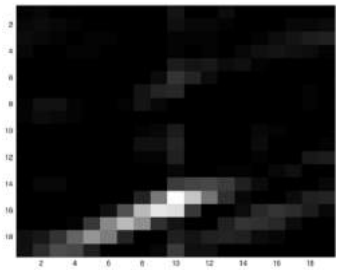


Figure 2.5: Resulting kernel using approach in Xu and Jia (2010)

term. This optimisation was solved using an alternating approach, first optimising with x constant, and then with k constant. These techniques were applied to the alpha matte as well as the foreground itself. Also, a combination of the alpha matte and the foreground was presented to both algorithms. They were tested at different resolutions and for different ground truth kernels. In addition to providing the entire foreground, indi-

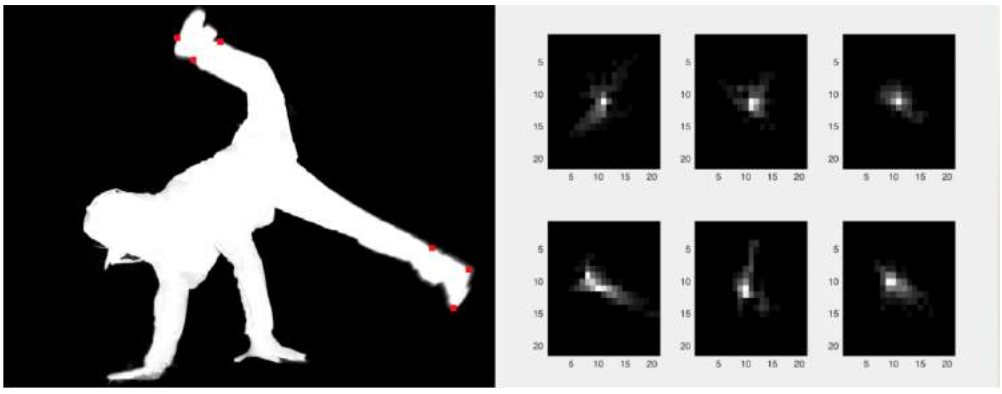
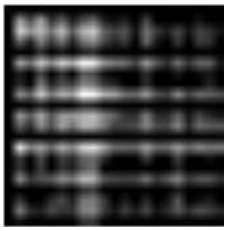


Figure 2.6: Technique from Pan *et al.* (2014) and associated kernels; Kernels for points from left to right (top 3 for higher leg, bottom 3 for lower leg) for a natural image

vidual body parts were also provided. They yielded unsastisfactory results, as shown in Figures 2.5 and 2.6. The estimated latent image presented artifacts, and the estimated kernel was not close to the original. This approach was therefore abandoned, and in-

(a) Input

(b) Estimated latent image



(c) Estimated kernel from Xu and Jia (2010)

Figure 2.7: Failure of blind deblurring methods to estimate kernel

stead a gradient based method was then adopted, wherein the blur was estimated by checking the spread of the alpha matte. Gradients were found using Sobel operator, shown below.

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix}, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \tag{2.5}$$

The amount of blur was estimated to be the distance in pixels to the first zero-value along the gradient. The method was later adapted to include a feature vector at every point, consisting of the distance of the first zero value along the gradient as well as directions inclined away from the gradient. Let l be the line starting at a point p (with co-ordinates (x, y)), with direction cosines c_1 and c_2. Then the magnitude of blur at the point p, is given by B (Equation 2.6).

$$B = \{min(t) \mid I(x + t * c_1, y + t * c_2) = 0\} \tag{2.6}$$

A line search was performed along these directions to estimate the blur. This is shown in Figure 2.8. Instead of this approach, a deep learning architecture was adapted from Sun *et al.* (2015).

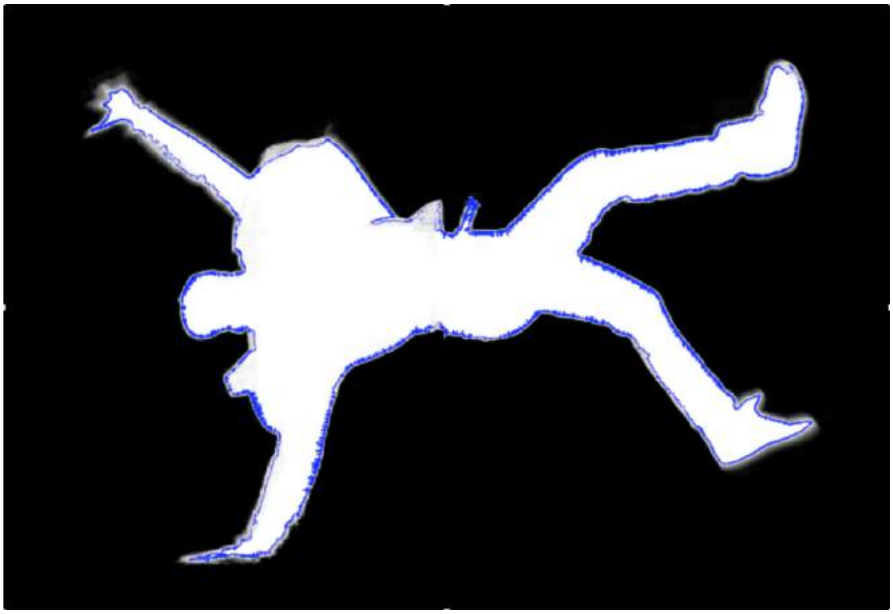


Figure 2.8: Blur estimate using gradients

2.2 Forays into Motion Analysis

Deep learning started out as an extension to artificial neural networks. These neural networks were developed to mimic human neurons. The earliest models of neurons include the perceptron which was developed in the 1950's by Frank Rosenblatt. Today, this has evolved into a different sort of basic unit, called a sigmoid neuron. The way the sigmoid neuron works is as follows. The inputs here are x_i, i ranging from 1 to N. The output then is $f(b + \sum_i w_i x_i)$ where the weights w_i are 'learnt' as a result of training. Here, f is known as the activation function. In sigmoid neurons, this function is the sigmoidal function (2.7).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.7}$$

When a network of these neurons is created, it results in an Artificial Neural Network (ANN). A rough schematic of the same is shown in Figure 2.9. When this framework is extended across more layers, training becomes more difficult, due to a problem known as the vanishing/exploding gradient problem. Hence, architectures with multiple layers are difficult to train, and this led to the establishment of deep learning as a field. One such deep network is shown in Figure 2.10. This network was adapted to our need and hence modified appropriately. The number of output classes was changed to 289. The network was trained using the ground truth silhouettes of the database in Shi *et al.* (2014), blurred with appropriate kernels. Training was carried out using mini-batch

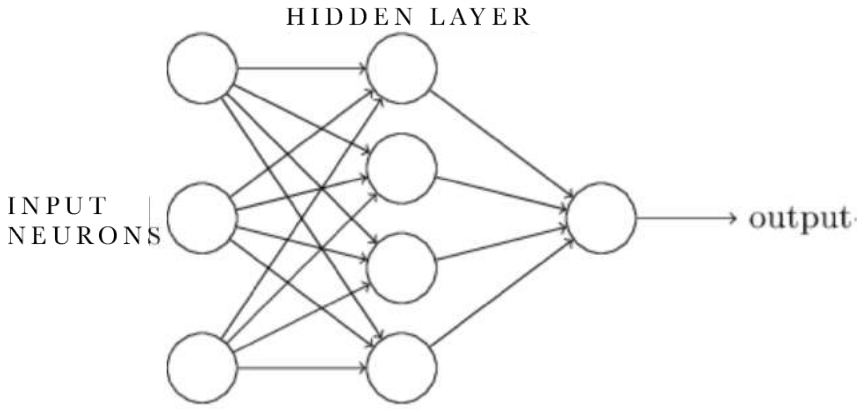


Figure 2.9: Generalised schematic of a single hidden layer neural network

stochastic gradient descent based on back propogation. The kernels used varied from length of 5 to 15 in increments of 1 pixel and from $0°$ to $170°$ with increments of $10°$ degrees. Training error saturated with a top 5 classification error of 5% after 143 epochs. Convolutional neural networks are different from regular networks because

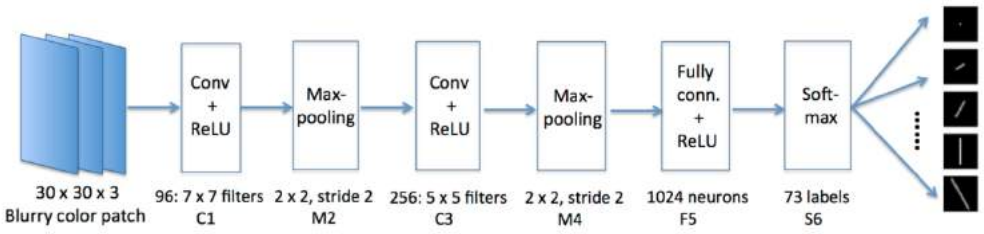


Figure 2.10: Deep learning architecture adapted from Sun *et al.* (2015)

their weights are tied. They are set up as in Figure 2.11. Only parts of the input image are connected to each hidden neuron, as opposed to all of them (as in fully connected layers). This region which is connected to a hidden neuron is called its local receptive field. For every layer, the weights connecting a hidden neuron and its local receptive field are the same. Therefore, the second layer is the result of a convolution between the input and the weights. There are sometimes multiple sets of weights per layer, leading to multiple 'feature maps' in every subsequent layer. To reduce the size of each layer, pooling is done, wherein the elements of a layer are pooled together to result in a form of subsampling operation. Pooling can be done in many ways including, but not limited to, mean pooling, median pooling and max pooling. Max pooling is the most common type. Another layer which is quite common in networks is the ReLU, or the Rectified

Linear Unit. It is of the form given in Equation 2.8.

$$f(x) = \begin{cases} x & x >= 0 \\ 0 & x < 0 \end{cases} \tag{2.8}$$

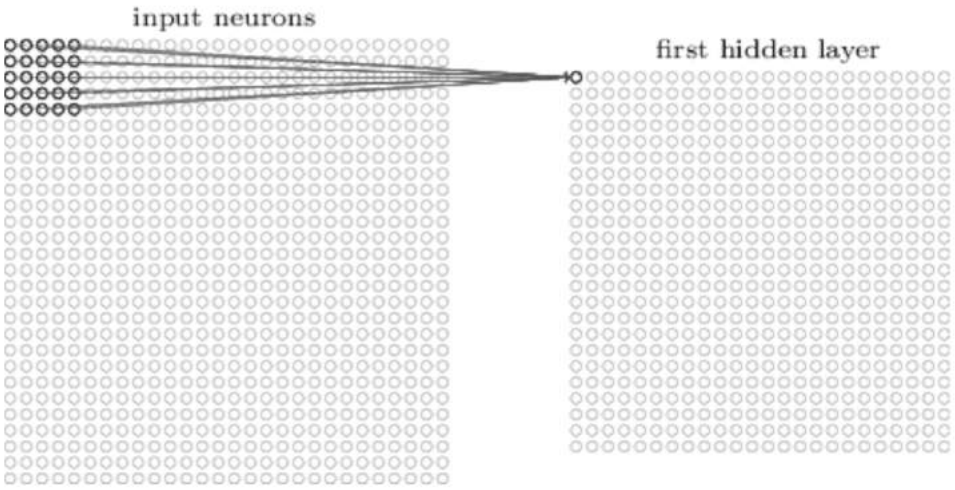


Figure 2.11: Structure of two layers in a typical CNN

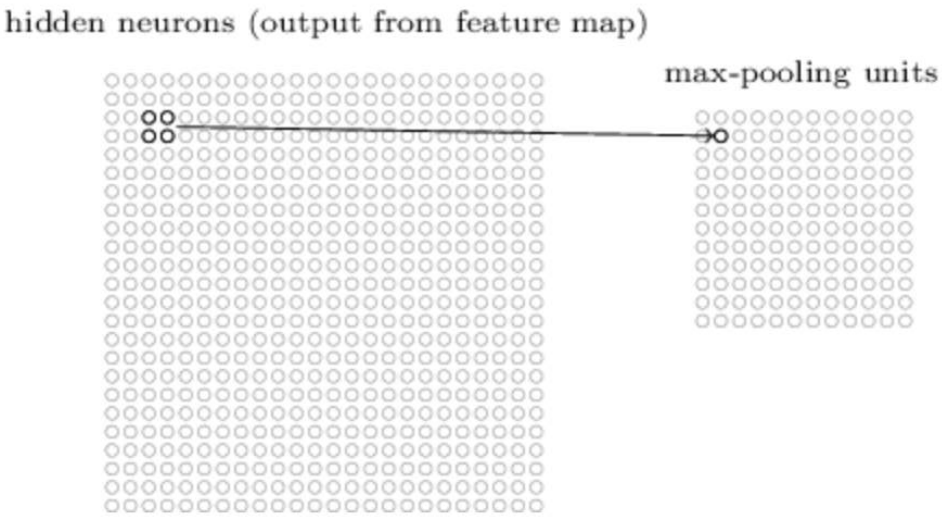


Figure 2.12: Pooling shown with a 2×2 receptive field

Training for this network was done using mini-batch Stochastic Gradient Descent based on back propagation. Once the motion blur is estimated at every point on the contour, kernel evolution with time opens up distinct possibilities. When the estimated motion kernels are analysed on a part by part basis (hand,leg etc.), it is shown that the temporal evolution of these kernels corresponds to the actual movement in space

of these body parts. A histogram of these body parts, therefore, could be used as a feature to detect the activity in a video. Results of this approach are shown in Figure 2.13. In this figure, the alpha matte of a moving person is shown. Along the contour of the alpha matte, the direction of motion blur has been plotted, with the direction being shown by the arrow and the length indicating the degree of blurring. The histogram which is plotted indicates the number of points within the red bounding box which have a particular direction. This is done via voting, wherein each point where blur has been estimated gets a vote weighted by the length of blur. The 360° histogram therefore shows the major direction of movement of a body part. Videos have also been generated showing the correspondence between movement and the histogram obtained.

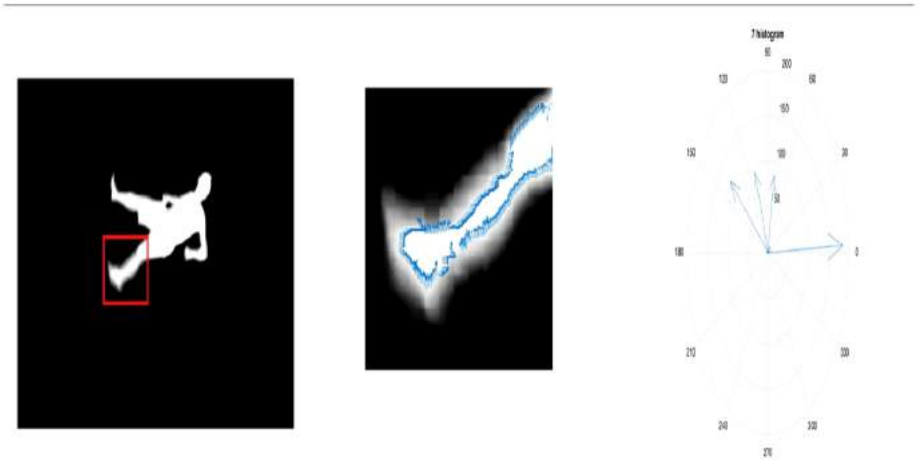


Figure 2.13: Results of binning of motion blur directions for a given body part

Another possible avenue to be explored was body part segmentation. Different body parts can be construed to have similar motion. Segmentation can therefore be done using the length and direction of estimated motion kernels as features. In our preliminary approach, 4 features were used. Each point along the contour had a feature vector F_i given by $F_i = [\alpha f_1, \beta f_2, x, y]$ where f_1 and f_2 are the motion length and direction respectively. The weighting factors α and β are used to weight the features during the actual clustering. x and y are the spatial coordinates of the point to allow for some spatial coherence. A hierarchical framework was adopted to cluster the points. The number of clusters was decided based on the histogram. A histogram was constructed of the motion directions. The bin size employed was 10°. It was found that the number of clusters is equal to the number of peaks in this histogram; i.e. if N denotes the number of clusters to be found, then $N = |\{x|H(x) > H(x-1) \& H(x) > H(x+1)\}|$

where $H(x)$ denotes the bin value of the x^{th} bin. This peak finding is carried out in a circular manner, which means that the last bin and the first bin are assumed to be adjacent. Once N is determined, a hierarchical clustering technique is used. The technique employs the inner squared distance metric (minimum variance algorithm) to perform clustering. Results are shown in Figure 2.15b.

(a) Original blurred image

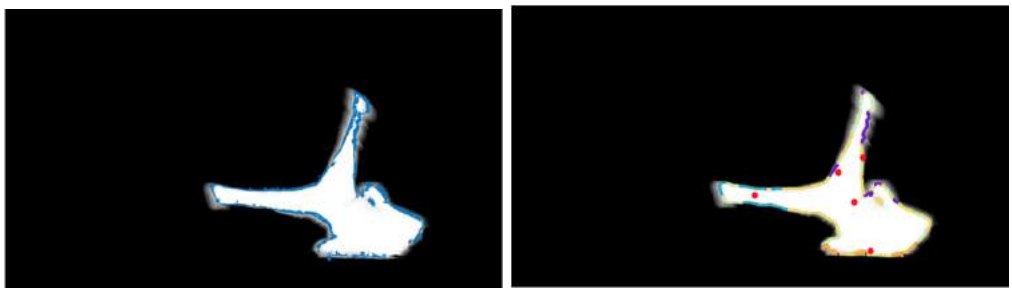
(b) Deblurred image with a patch size of 30×30

(c) Deblurred image with a patch size of 100×100

Figure 2.14: Deblurring for articulated motion

The clustering is jagged and uneven. This is mainly due to the irregular direction estimates and outliers. This can be corrected probably by applying a smoothing oper-

ation such as Markov Random Fields or averaging, either before the clustering step or after.



(a) Estimated directions

(b) Clustering result using $\alpha = 5$ and $\beta = 10$, cluster centroids indicated using red dots

Figure 2.15: Body part segmentation using direction of motion blur

(a) Input 1 (b) Input 2 (c) Input 3

Figure 2.16: Images from the database provided by Köhler *et al.* (2012)

Another possibility is that of deblurring the foreground. Applications can be found in sports photography, or any other form of photography where a high shutter speed is not an option. The deblurring approach used is patch-wise, and uses the non-blind deconvolution proposed in Krishnan and Fergus (2009). In addition, a modified Bartlett-Hanning window (Ha and Pearce (1989)) was applied to each patch, in an effort to remove edge artifacts. The results still present some sort of artifacts in the background, however, significant deblurring is present as shown in Figure 2.14. Two patch sizes were used, 30×30 and 100×100. Further work on deblurring followed. A Siamese network was developed, to deblur a given input image. The input to the network was a pair of images, obtained from the algorithm in Krishnan and Fergus (2009). Each of these was obtained by using a different λ parameter, which controls the weight of the prior applied in the optimisation. These two inputs to the network resulted in an output which did not have the ringing associated with a very low value of λ nor the blurred

(a) Baboon (b) Pepper (c) Zebra

Figure 2.17: Input images used for benchmarking

Image	Lucy-Richardson	Krishnan and Fergus	3x3 FoE	Pairwise MRF	Shrinkage Fields
Baboon	20.226	20.1275	**33.89714**	32.5493	27.4127
Barbara	22.4984	22.9738	**39.8221**	38.2195	31.2494
Bridge	21.2189	21.3735	**34.2744**	33.1093	30.1206
Coastguard	22.1157	22.0846	**35.9256**	35.6890	30.8608
Comic	18.4805	18.8303	28.1743	28.0109	**28.3819**
Face	25.3689	25.7022	**36.4110**	35.1509	34.4171
Flowers	22.1441	22.9614	**32.8632**	32.6740	32.1274
Foreman	22.3996	21.8723	37.6857	**36.7987**	34.5081
Lenna	25.6344	25.7400	43.0887	**40.6222**	35.2952
Man	22.4041	22.9790	**37.5988**	36.7332	31.4494
Monarch	22.8429	24.0706	**39.5834**	37.8078	35.8327
Pepper	25.3659	25.5632	**38.1657**	37.5007	35.4544
ppt3	18.3782	20.1307	**52.3618**	48.5715	29.3832
Zebra	20.0872	20.8863	**34.0405**	32.3988	32.3755

Table 2.1: Benchmarking results for various blind deblurring algorithms

edges associated with a very high value. The structure of the network used is shown in Figure 2.18.

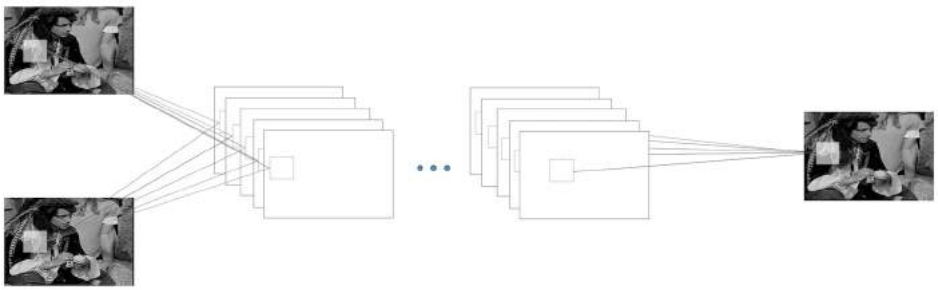
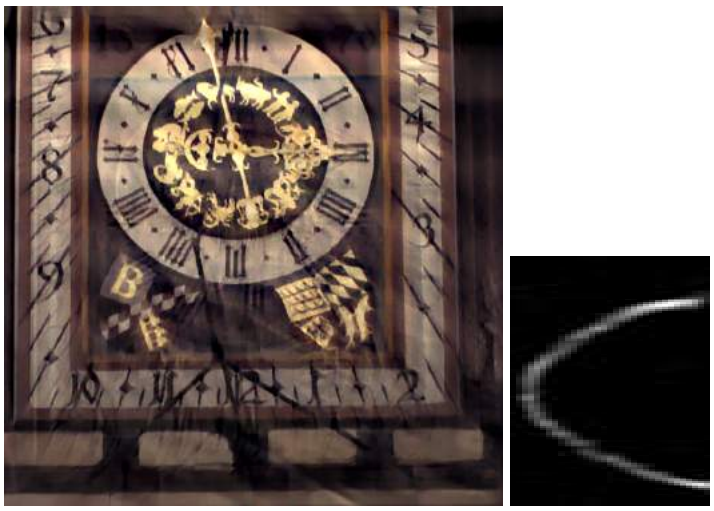


Figure 2.18: Siamese network used for blind deblurring

Benchmarking was then carried out, wherein this algorithm was compared to the performance of standard deblurring algorithms on various datasets.

(a) Output corresponding to first input and estimated blur kernel



(b) Output corresponding to second input and estimated blur kernel

(c) Output corresponding to third input and estimated blur kernel

The deblurring algorithms used were the Lucy-Richardson algorithm, Krishnan and

Fergus' Hyper Laplacian Priors, 3×3 FoE (Regression Tree Fields), Pariwise MRF based deblurring, and a method using shrinkage fields. The result of one such benchmarking activity is shown in Table 2.1, along with some of the images the algorithms were used on. The outputs from one of the deblurring algorithms (Xu et. al) is shown in Figure 2.19c.

CHAPTER 3

HDR based Simultaneous Motion Deblurring and Super Resolution

(a) Input image

(b) PSF's at corresponding locations of the input image

Figure 3.1: An example of the blurred images which serves as inputs

3.1 Introduction

Handheld photography is a burgeoning practice, with the proliferation of light-weight, high quality imaging devices. Images from these sources are especially prone to camera shake. Traditional methods for super-resolution (SR) and high dynamic range (HDR) imaging are designed to work well when the input images are free of blur. However, in many real applications, blur in input images is a common occurrence, leading to failure of these methods. In this paper, a unified approach to perform SR and HDR from multiple non-uniformly motion blurred input images is proposed. Unlike existing approaches, our approach attempts to harness the complementarity present in terms of the sensor exposure and blur to yield a high quality image which has both higher spatial resolution as well as dynamic range. Our method minimizes a regularized energy function defined in terms of the desired image, input images and the blur introduced by the camera shake, where we use regularization on the image and camera motion to ensure the convergence of the algorithm. Experiments have been carried out on several real and synthetically generated images to validate the efficacy of our framework.Capturing

Figure 3.2: A High Dynamic Range Image

high quality images is one of the, if not the most important goals of modern photography. However, there exist multiple hurdles to achieving this goal. Natural scenes typically have irradiances spanning multiple orders of magnitude. This far exceeds the range of the commonly used camera sensors and often leads to pictures containing over-exposed/under-exposed regions. Although HDR cameras are commercially available they are very expensive. High dynamic range (HDR) imaging is concerned with extraction of latent scene irradiance. HDR imaging techniques attempts to algorithmically fuse the information contained in multiple low dynamic range (LDR) images captured at different exposure settings. The idea behind HDR imaging is the fact that, if we can capture LDR images such that all the scene points are well exposed in atleast one image, one can try to solve for a single image which is devoid of over-exposed/under-exposed regions. Based on this idea, several algorithmic approaches for scene irradiance estimation Mitsunaga and Nayar (1999); Ward (2003); Debevec and Malik (2008); Sen *et al.* (2012) have been proposed, all, however, assuming availability of blur-free input images. Increasing spatial resolution of images is a well known classical problem. While a direct way to achieve this is to increase the number of sensor elements per unit area, the cost of producing such sensor arrays may not be commensurate with the gains. Pixel size reduction also results in the increased effect of shot noise in the captured images. An alternative approach that is commonly used is the use of signal processing techniques to improve spatial resolution. Multi-image based super-resolution techniques

try to exploit the extra information present in different LR images to improve the spatial resolution Park *et al.* (2003). Such a scheme attempts to model the imaging process between an unknown high-resolution image and multiple low-resolution observations and try to solve the inverse problem. However, most existing methods are designed to work well for the limited scenario where all the captured images will be blur free.

The fact that both SR and HDR imaging (hereafter referred to as HDR-SR) require multiple input images opens up the possibility to employ them for simultaneous SR-HDR. Although many approaches for SR-HDR have been proposed, all of them assume the input images being blur-free. With the introduction of light-weight, hand-held, photographic devices, images are now increasingly susceptible to camera shake due to unsteady hands. Unlike the most commonly used model for space-invariant blur, camera-shake often leads to spatially varying blur in the captured images Köhler *et al.* (2012). This paper aims at obtaining a single SR-HDR image free of blur from multiple spatially varying motion blur affected input images. However, the problem in its own form is ill-posed since we also need to estimate the camera motion which resulted in the blur in input images. Previous studies Whyte *et al.* (2014); Pan *et al.* (2016) on motion deblurring have shown that a careful handling of saturation regions is important for reliable camera motion and latent image estimation. Since we work with different exposure input images, it is important to carefully model the image formation mechanism so that the right information get fused at the final image while keeping our estimation process robust enough to handle the saturation regions. We propose an alternating minimization scheme to simultaneously solve for the the camera motion at high-resolution as well as a blur-free SR-HDR image.

3.2 Prior Work

Many prior works exist, which are relevant to the development of the technique in this chapter. The first domain that we must discuss is the estimation of scene irradiance. Works such as Mann and Picard estimate the response curve through the use of multiple points across images and curve fitting a parametric model of the CRF. Ng *et al.* (2007) also assumes a parametric form for the CRF, in terms of polynomial exponents and estimates the same by the use of geometric invariants. Many other approaches exist,

but since this work assumes that the input is in the form of irradiance images, we will not delve into it further. Methods exist to effectively fuse differently exposed pictures to obtain an image without saturation and/or darkness. In general, a weighting scheme is proposed, and the images are suitably combined to obtain the final HDR-equivalent LDR image. For example, Mertens *et al.* (2009) uses a weight which takes into account local contrast as well as well-exposedness of a pixel. Raskar *et al.* (2005) and Fattal *et al.* (2002) both analyse the differnetly exposed images in the gradient domain. All of these approaches to HDR imaging, however, do not explore the possibility of camera motion.

Motion deblurring in itself is a problem which has held the attention of researchers time and time again. Fergus *et al.* (2006) achieves this by using a prior which they developed, on the image. However this only accounts for spatially invariant blurring. Many approaches to tackle non uniform motion exist. Gupta *et al.* (2010) Tai *et al.* (2011) and Whyte *et al.* (2012) all work on a TSF like approach wherein the image is modelled as an average of images captured at various camera poses. These approaches take a single image as input, and do not leverage multiple images in situations which present themselves in this light. Vio *et al.* (2004) and Schulz (1993) both propose methods to perform multi-image blind deconvolution. A more recent work, Zhang *et al.* (2013) implements a novel Bayesian approach based prior to address the multi-image deblurring problem.

Motion blur naturally lends itself to the super resolution problem, which has been a central goal of many research groups over the years. The desire to ahieve, by signal processing, an increase in the resolution of images, without an increase in cost, or a change in hardware has led to extensive work being done in the domain. The earliest works which combines multiple low resolution images, Elad and Feuer (1997) and Nguyen *et al.* (2001), establish an observational model which involves blurring of images (often with sub-pixel movements) and subsequent decimation. Once this model is established, the reconstruction algorithm employed varies from work to work. Ur and Gross (1992) uses nonuniform interpolation by using the generalised multichannel sampling theorem. Frequency domain approaches have also been proposed such as in Bose *et al.* (1993) and Rhee and Kang (1999). A more recent work, Li *et al.* (2010), uses the same framework but imposes additional regularisation terms to better define the inherently ill-posed problem of super resoltuion.These approaches are all fundamentally limited

in the sense that the type of blur experienced by every image is assumed to be the same. Approaches to simultaneously tackle both motion blur and super resolution have also been made such as Harmeling *et al.* (2010), and most notably in Sroubek *et al.* (2007), however this has not been adapted to an application for HDR imaging. However, none of these approaches models the effect of a large dynamic range of the input images.

Approaches to attempt HDR and super resolution have also gained traction. In Schubert *et al.* (2009), two input sequences are required, one with differently exposed images, and one with slightly perturbed views of the scene. The optimal number of images needed as well as the exposure times to ensure a good HDR-SR output has been explored in Traonmilin and Aguerrebere (2014). The work in Haraldsson *et al.* (2007) uses a modified image acquisition system, to obtain a multisampled image capable of giving a HDR latent image. Bayesian and Graphical model based approaches to the problem have also been proposed, such as in Zhang and Carin (2014) and Whyte *et al.* (2012). Working in a transform domain, which accounts for the appearance to the human visual system, Bengtsson *et al.* (2012) works on minimising a tailored cost function. Attempts to perform SR interpolation have also been carried out, such as in Rad *et al.* (2007). A requirement for all of these methods however is that the images must be registered and motion blur is not present, except in Zhang and Carin (2014) wherein HDR has not been tackled.

The main contributions of this work is simultaneous super resolution, HDR imaging and motion blur removal for the first time in literature

3.3 Formulation

To begin, we will illustrate the image formation model adopted, and arrive at the model we have used in this chapter.

We will begin by assuming that the scene, in the irradiance domain, has been referred to as l. Let us say that this latent image is of resolution $fM \times fN$ where f is the super resolution (down-sampling) factor. In a situation without any camera motion, the

final image is an integrated function of the irradiance. This is shown in Equation 3.1.

$$E = f(\int_0^t l dt + n) \tag{3.1}$$

where n is additive noise, f is the CRF and the scene irradiance is assumed to be constant within the exposure time of t. The mapping from irradiance to pixel values is done by the CRF and varies from camera to camera. Many techniques exist to estimate this function, such as Mann and Picard and Ng *et al.* (2007). This work however, uses the approach illustrated in Debevec and Malik (2008). In the case when the camera undergoes motion, the pose in which the camera sensor exists at every time instant results in the latent image l being appropriately modified by means of a homography. At every time instant, let the camera pose induce a homography on the captured image, referred to as Γ_t. The final intensity image obtained is then an average of the scene, as viewed from these various camera poses. This is shown in Equation 3.2.

$$E = f(\int_0^t \Gamma_t(l) dt + n) \tag{3.2}$$

Finally, to account for downsampling, we assum that the real world scene (of resolution $fM \times fN$) is averaged and then decimated by the sensor (of resolution $M \times N$). Let us call the matrix which averages this image as H and the matrix which decimates i.e. downsamples the image as M. The final matrix which accomplishes both these tasks is referred to as D. The final image captured by the camera, henceforth referred to as B is then obtained. This can be seen in Equation 3.3

$$E = f(D \int_0^t \Gamma_t(l) dt + n) \tag{3.3}$$

Despite the camera trajectory being continuous, for the sake of tractability, the final image can be viewed as the weighted average of a few dominant camera poses, effectively discretising the camera's motion. The weight is determined by the fraction of exposure time that the camera is in that pose. We can thus replace the integral in Equations 3.1 3.2 and 3.3 by a summation over these dominant poses, hence forth referred to as the pose space. The final model, which summarises the image capture process is shown in Equation 3.4.

$$E = f(D \sum_0^k h_k(\Gamma_k)\Gamma_k(l) + n) \tag{3.4}$$

This section contains the theoretical backing for the proposed approach, any derivations which might be needed, as well as the outline of the work presented. A rough schematic is provided in Algorithm 1.

Algorithm 1 Rough outline of the algorithm implemented in our work

Require: All Input Images in Irradiance Domain by Inverse Mapping with CRF

1: **while** not converged **do**
2: Solve $h_i = \min_{h_i} E(h_i, \hat{l})$, minimising for TSF
3: $\hat{l} = \min_{\hat{l}} E(h_i, \hat{l})$, minimising for latent image
4: Pose perturbation
5: **end while**
6: Perform Final Optimisation for Colour Image
return l, $h_i's$

3.3.1 Images in the irradiance domain

The image formation model involves various sub parts. This chapter works in the irradiance domain, and hence an understanding of how the capture of images is modeled is required. Let l refer to the high resolution, latent image in the irradiance domain. Let the hypothetically captured, clean, high resolution image be E. Then, E and l are related by the Camera Response Function (CRF) as in Equation 3.5

$$E = f(E \times t_e) \tag{3.5}$$

where t_e is the exposure time and f is the CRF. The CRF takes into account the entire camera pipeline, and relates the final pixel values to the scene irradiance. Standard techniques exist to estimate this CRF, which is a smooth monotonous function. We can also model this light capture as an integration, given in Equation 3.6.

$$P = \int_0^{t_e} l dt + n \tag{3.6}$$

where P is the captured image and n is the added noise. A typical CRF is shown in Figure 3.3. To effectively illustrate the algorithm adopted in this paper, we will assume that the CRF has been estimated and the input images to our algorithm, referred to as B_i's, are low resolution, blurred images, in the irradiance domain i.e. they have been inverse mapped using the CRF.

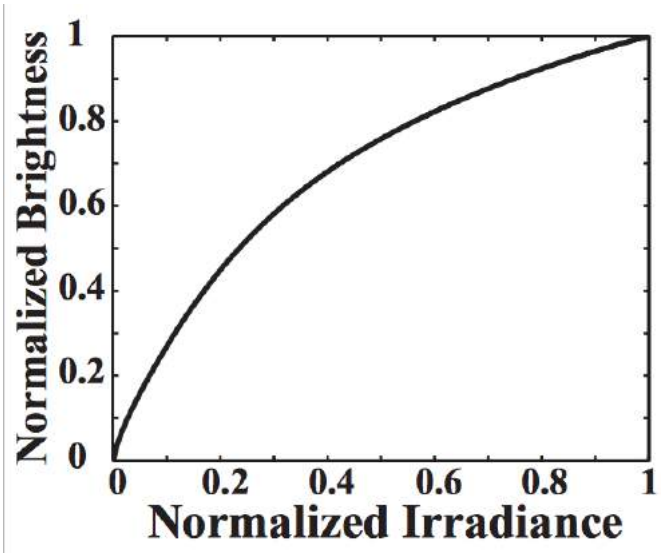


Figure 3.3: A typical CRF

3.3.2 Blurring

Working in the irradiance domain, we first assume that the high resolution latent image has been subjected to capture in the presence of camera shake. A camera has 6 degrees of freedom, namely translation along t_x, t_y and t_z and also rotation along the three axes ϕ_x, phi_y and phi_z. We assume that the scene has minimal depth variations, and so the camera shake can be modelled as an average of various homographies of the latent image. With this constraint, the out of plane rotations can be modelled as translations, and hence only 3 degrees of freedom are considered in this chapter, namely t_x, t, y and phi_z. The camera pose space is then discretised appropriately. We adopt the TSF

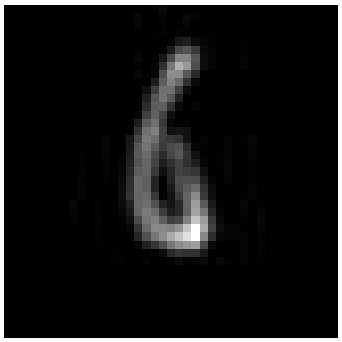


Figure 3.4: A typical PSF obtained due to camera shake

approach, wherein the intermediate, high resolution, blurred images are referred to as $M_i's$. The relationship between these $M_i's$ and l is as given in Equation 3.7.

$$M_j = \sum_k h_j(\Gamma_k)\Gamma_k(l) \tag{3.7}$$

where each Γ_k is a transformation which applies a homography (weighted by a real, non-negative number denoted by $h_j(\Gamma_k)$) to the latent image l. Each of the j input images is formed by a separate, possibly distinct TSF, hence the subscript. The function h_j is defined in Equation 3.8 and is referred to as the TSF. It denotes the fraction of the exposure the time that the camera remained in the corresponding pose. Note that since the blurring is assumed to act on the high resolution image, the final output of our algorithm is the high resolution TSF.

$$h : \Gamma \rightarrow [0, 1] \mid \sum_k h(\Gamma_k) = 1 \tag{3.8}$$

3.3.3 Decimation

The blurred high resolution images are then modelled to be decimated to form blurred, low resolution images. The decimation operation first applies a blur to the HR image, and then selects pixels so as to form a low resolution image. For example, with a super-resolution factor of 4, every alternate pixel is selected (per row and column). The final decimation matrix which acts on the $M_i's$ is denoted by D. Hence, we arrive at Equation 3.9.

$$B_i = DM_i \tag{3.9}$$

3.3.4 Putting it all together

As illustrated so far, the complete pipeling from ground truth scene irradiance, to the blurred, low resolution irradiance images which are input to our algorithm is summarised in Equation 3.10.

$$B_i = D\left(\sum_k h_i(\Gamma_k)\Gamma_k(l)\right) \tag{3.10}$$

3.3.5 Optimisation

The optimisation is carried out to estimate two unknowns, namely the high resolution (HR) TSF's h_i corresponding to each blurred image B_i and the HR, latent irradiance image l. To do so, an alternating optimisation approach is adopted. The first of the two

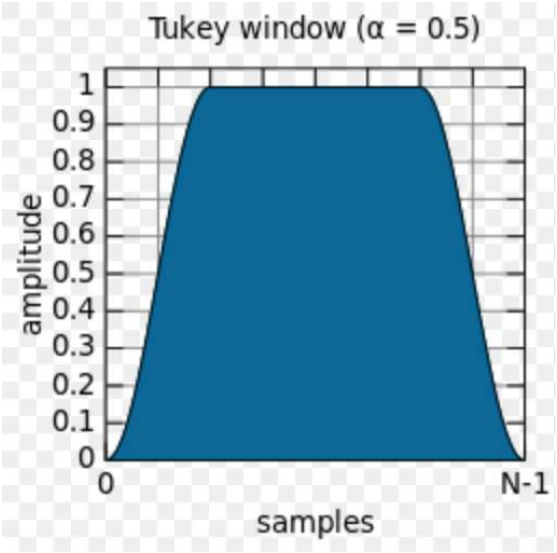


Figure 3.5: Tukey window. In our case N is 256

steps involves the TSF estimation by assuming the latent image to be known. Additional terms, such as the $l1$ norm to enforce sparsity on the TSF have also been added. The energy term to be minimised in this first optimisation is shown in Equation 3.11.

$$\underset{h_i}{argmin} \| B_i \; - \; D\left(\sum_k h_i(\Gamma_k)\Gamma_k(l)\right)\|^2 \; + \; \|\nabla B_i \; - \; D\left(\sum_k h_i(\Gamma_k)\Gamma_k(\nabla l)\right)\|^2 \tag{3.11}$$

Additional terms to enforce smoothness on the latent image, such as the TV norm are also included. An added term which has shown to give better results, involves the gradients of the images involved. The energy function to be minimised then takes the form of Equation 3.12 for every input image B_i

$$\hat{l} = \underset{l}{argmin} \sum_i \|B_i - D\left(\sum_k h_i(\Gamma_k)\Gamma_k(l)\right)\|^2 + \|\nabla B_i - D\left(\sum_k h_i(\Gamma_k)\Gamma_k(\nabla l)\right)\|^2 + \|\nabla l\|^2 \tag{3.12}$$

This optimisation is carried out using Preconditioned Conjugate Gradients. The Conjugate Gradient approach is based on the assumption that the contour is convex. However, it is in general applicable to non-linear, non convex surfaces as well, when applied iteratively. The founding equations of the conjugate gradient method are shown in Equation

??.

$$p_k = r_k - \sum_{i<k} \left(\frac{p_i^T A r_k}{p_i^T A p_i} p_i \right) \tag{3.13}$$

$$x_{k+1} = x_k + \alpha_k p_k \alpha_k = \frac{p_k^T r_k}{p_k^T A p_k} \tag{3.14}$$

The function to minimise is

$$f(x) = \frac{1}{2} x^T A x - x^T b \tag{3.15}$$

and r_k is the residual given by $r_k = b - Ax_k$ and p_0 is set to be an arbitrary direction, often the gradient. The function to be optimised is quadratic, and hence this method is applied directly.The minimisation is carried out to obtain the latent image estimate $\hat{l}$ and the process is iteratively repeated. Since our images are in converted to the irradiance domain in a preprocessing step, saturation or extremely low values may result in inaccurate values. Hence, a mask is applied to ignore those pixels which are saturated or have a value too low. This is done in the following way: if the energy function to be optimised is $\|B_i - l\|^2$, this is the same as $\|mB_i - ml\|^2$ where m is the mask. This is similarly extended to our objective function.

3.3.6 Pose Space Modification

As in Punnappurath *et al.* (2014) the pose space is modified at every iteration to limit its size. This is done by randomly sampling in and around the most dominant poses at every iteration, by using a gaussian distribution as the sampling distribution. Clever choosing of this pose space often leads to good results in relatively shorter periods of time.

3.4 Experiments and testing

Experiments were carried out mainly with synthetic images. The images used are shown in Figure 3.6. The following methodology was used to generate the input images. The input images were all HDR images, wherein the scene irradiance is recorded. Different

(a) Exposure time $30ms$

(b) Exposure time $100ms$

(c) Exposure time $1000ms$

(d) Exposure time $4000ms$

Figure 3.6: Clean images at various exposures

exposure times were chosen and were each multiplied with the ground truth irradiance images. These differently 'exposed' images were then subjected to blurring by using a TSF. The TSF's used were different for every exposure time. Once done, these images were then forward transformed using the CRF to give the input images. These input images were RGB 8 bit images. Images are then blurred by way of TSF's which have been synthesised. Examples of these blurred input images can be seen in Figure 3.7. One of the HDR outputs can be seen in Figure 3.8. Notice that there are no regions

(a) Exposure time: $1000s$

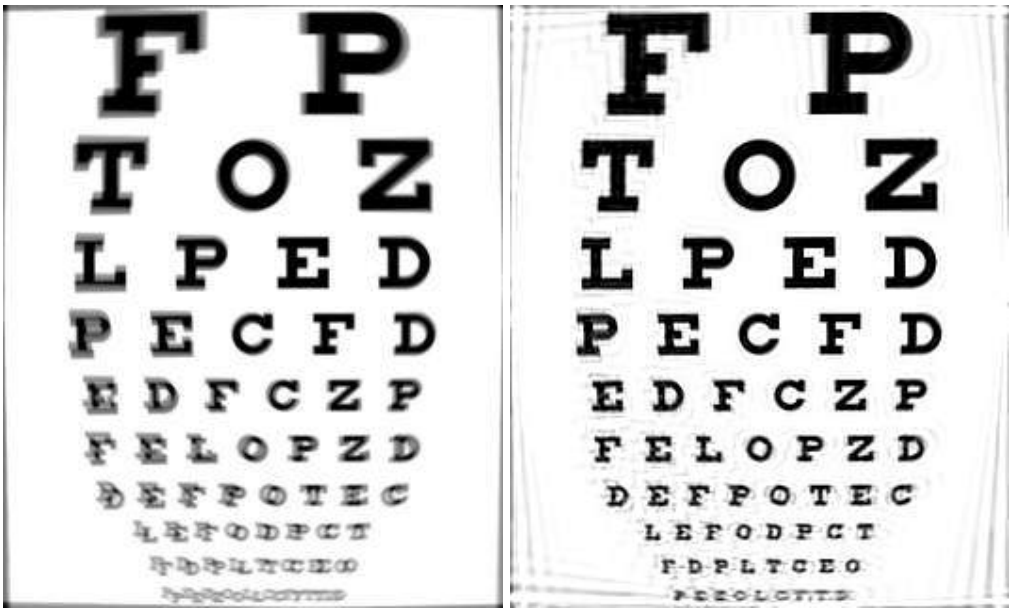
(b) Exposure time: $4000s$

Figure 3.7: Synthetically generated blurred images simulating camera shake

of saturation/darkness. Also, the blur has been significantly reduced when compared to Figure 3.7. However, there is colour saturation along the bright edges, and also a ghosting effect, typical of HDR algorithms. Further parameter tuning must be carried

Figure 3.8: Output of our algorithm

out under other examples. There are 3 parameters to be tuned. The expected kernel size, the coefficient of the TV prior in the latent image estimation step, and the weight of the $l1$ prior in the TSF estimation step all effect the quality of the output. Their interdependencies are intricate and are to be further examined.



(a) One of the input images (b) Deblurred output

Figure 3.9: Example of deblurring without HDR

To demonstrate the efficacy of the deblurring our algorithm performs, a non-HDR example is shown in Figure 3.9. It is clear that the algorithm performs significant de-

blurring. The 7^th row of text has been nearly totally deblurred, and is legible in the output. The edges of the image however show grey lines, symptoms of the zero padding used while computing homographies of the latent image.

3.5 Conclusion

A further extension to this work is the modelling depth variations. However, for the case of the flat scene, we have shown the efficacy of our algorithm. Methods to speed up the running time, as well as careful selection of the camera pose space are issues which pose a threat to the effectiveness of the algorithm.

REFERENCES

17. **Bengtsson, T.**, **I. Y.-H. Gu**, **M. Viberg**, and **K. Lindström**, Regularized optimization for joint super-resolution and high dynamic range image reconstruction in a perceptually uniform domain. *In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012.

2. **Bose, N.**, **H. Kim**, and **H. Valenzuela**, Recursive implementation of total least squares algorithm for image reconstruction from noisy, undersampled multiframes. *In Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 5. IEEE, 1993.

3. **Chen, C.**, **J. Liang**, **H. Zhao**, **H. Hu**, and **J. Tian** (2009). Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, **30**(11), 977–984.

4. **Debevec, P. E.** and **J. Malik**, Recovering high dynamic range radiance maps from photographs. *In ACM SIGGRAPH 2008 classes*. ACM, 2008.

5. **Elad, M.** and **A. Feuer** (1997). Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, **6**(12), 1646–1658.

6. **Fattal, R.**, **D. Lischinski**, and **M. Werman**, Gradient domain high dynamic range compression. *In ACM Transactions on Graphics (TOG)*, volume 21. ACM, 2002.

7. **Fergus, R.**, **B. Singh**, **A. Hertzmann**, **S. T. Roweis**, and **W. T. Freeman**, Removing camera shake from a single photograph. *In ACM Transactions on Graphics (TOG)*, volume 25. ACM, 2006.

8. **Gupta, A.**, **N. Joshi**, **C. Lawrence Zitnick**, **M. Cohen**, and **B. Curless** (2010). Single image deblurring using motion density functions. *Computer Vision–ECCV 2010*, 171–184.

9. **Ha, Y. H.** and **J. A. Pearce** (1989). A new window and comparison to standard windows. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**(2), 298–301.

10. **Haraldsson, H. B.**, **M. Tanaka**, and **M. Okutomi**, Reconstruction of a high dynamic range and high resolution image from a multisampled image sequence. *In Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*. IEEE, 2007.

11. **Harmeling, S.**, **S. Sra**, **M. Hirsch**, and **B. Schölkopf**, Multiframe blind deconvolution, super-resolution, and saturation correction via incremental em. *In Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010.

12. **Köhler, R.**, **M. Hirsch**, **B. Mohler**, **B. Schölkopf**, and **S. Harmeling**, Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. *In European Conference on Computer Vision*. Springer, 2012.

13. **Krishnan, D.** and **R. Fergus**, Fast image deconvolution using hyper-laplacian priors. *In Advances in Neural Information Processing Systems*. 2009.

14. **Lam, T. H.**, **K. H. Cheung**, and **J. N. Liu** (2011). Gait flow image: A silhouette-based gait representation for human identification. *Pattern recognition*, **44**(4), 973–987.

15. **Levin, A.**, **D. Lischinski**, and **Y. Weiss** (2008). A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(2), 228–242.

16. **Li, X.**, **Y. Hu**, **X. Gao**, **D. Tao**, and **B. Ning** (2010). A multi-frame image super-resolution method. *Signal Processing*, **90**(2), 405–414.

17. **Liu, J.** and **N. Zheng**, Gait history image: a novel temporal template for gait recognition. *In Multimedia and Expo, 2007 IEEE International Conference on*. IEEE, 2007.

18. **Man, J.** and **B. Bhanu** (2006). Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, **28**(2), 316–322.

19. **Mann, S.** and **R. Picard** (). On being âĂŸundigitalâĂŹwith digital cameras: Extending dynamic range by combining differently exposed pictures, 7 pages.

20. **Mertens, T.**, **J. Kautz**, and **F. Van Reeth**, Exposure fusion: A simple and practical alternative to high dynamic range photography. *In Computer Graphics Forum*, volume 28. Wiley Online Library, 2009.

21. **Mitsunaga, T.** and **S. K. Nayar**, Radiometric self calibration. *In Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1. IEEE, 1999.

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

23. **Nguyen, N.**, **P. Milanfar**, and **G. Golub** (2001). Efficient generalized cross-validation with applications to parametric image restoration and resolution enhancement. *IEEE Transactions on image processing*, **10**(9), 1299–1308.

24. **Pan, J.**, **Z. Hu**, **Z. Su**, and **M.-H. Yang**, Deblurring text images via l0-regularized intensity and gradient prior. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

25. **Pan, J.**, **Z. Lin**, **Z. Su**, and **M.-H. Yang**, Robust kernel estimation with outliers handling for image deblurring. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

26. **Park, S. C.**, **M. K. Park**, and **M. G. Kang** (2003). Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine*, **20**(3), 21–36.

27. **Punnappurath, A.**, **A. Rajagopalan**, and **G. Seetharaman**, Blind restoration of aerial imagery degraded by spatially varying motion blur. *In SPIE Defense+ Security*. International Society for Optics and Photonics, 2014.

28. **Rad, A. A.**, **L. Meylan**, **P. Vandewalle**, and **S. Süsstrunk**, Multidimensional image enhancement from a set of unregistered and differently exposed images. *In Electronic Imaging 2007*. International Society for Optics and Photonics, 2007.

29. **Raskar, R.**, **A. Ilie**, and **J. Yu**, Image fusion for context enhancement and video surrealism. *In ACM SIGGRAPH 2005 Courses*. ACM, 2005.

30. **Rhee, S.** and **M. G. Kang** (1999). Discrete cosine transform based regularized high-resolution image reconstruction algorithm. *Optical Engineering*, **38**(8), 1348–1356.

31. **Schubert, F.**, **K. Schertler**, and **K. Mikolajczyk**, A hands-on approach to high-dynamic-range and superresolution fusion. *In Applications of Computer Vision (WACV), 2009 Workshop on*. IEEE, 2009.

32. **Schulz, T. J.** (1993). Multiframe blind deconvolution of astronomical images. *JOSA A*, **10**(5), 1064–1073.

33. **Sen, P.**, **N. K. Kalantari**, **M. Yaesoubi**, **S. Darabi**, **D. B. Goldman**, and **E. Shechtman** (2012). Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, **31**(6), 203.

34. **Shi, J.**, **L. Xu**, and **J. Jia**, Discriminative blur detection features. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

35. **Sroubek, F.**, **G. Cristóbal**, and **J. Flusser** (2007). A unified approach to superresolution and multichannel blind deconvolution. *IEEE Transactions on Image Processing*, **16**(9), 2322–2332.

36. **Sun, J.**, **W. Cao**, **Z. Xu**, and **J. Ponce**, Learning a convolutional neural network for non-uniform motion blur removal. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.

37. **Tai, Y.-W.**, **P. Tan**, and **M. S. Brown** (2011). Richardson-lucy deblurring for scenes under a projective motion path. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **33**(8), 1603–1618.

38. **Traonmilin, Y.** and **C. Aguerrebere** (2014). Simultaneous high dynamic range and superresolution imaging without regularization. *SIAM Journal on Imaging Sciences*, **7**(3), 1624–1644.

39. **Ur, H.** and **D. Gross** (1992). Improved resolution from subpixel shifted pictures. *CVGIP: Graphical Models and Image Processing*, **54**(2), 181–186.

40. **Vio, R.**, **J. Nagy**, **L. Tenorio**, and **W. Wamsteker** (2004). A simple but efficient algorithm for multiple-image deblurring. *Astronomy & Astrophysics*, **416**(1), 403–410.

41. **Wang, C.**, **J. Zhang**, **L. Wang**, **J. Pu**, and **X. Yuan** (2012). Human identification using temporal information preserving gait template. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(11), 2164–2176.

42. **Ward, G.** (2003). Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *Journal of graphics tools*, **8**(2), 17–30.

43. **Whyte, O.**, **J. Sivic**, and **A. Zisserman** (2014). Deblurring shaken and partially saturated images. *International Journal of Computer Vision*. URL `http://dx.doi.org/10.1007/s11263-014-0727-3`.

44. **Whyte, O.**, **J. Sivic**, **A. Zisserman**, and **J. Ponce** (2012). Non-uniform deblurring for shaken images. *International journal of computer vision*, **98**(2), 168–186.

45. **Xu, L.** and **J. Jia**, Two-phase kernel estimation for robust motion deblurring. *In European Conference on Computer Vision*. Springer, 2010.

46. **Zhang, E.**, **Y. Zhao**, and **W. Xiong** (2010). Active energy image plus 2dlpp for gait recognition. *Signal Processing*, **90**(7), 2295–2302.

47. **Zhang, H.** and **L. Carin**, Multi-shot imaging: joint alignment, deblurring and resolution-enhancement. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.

48. **Zhang, H.**, **D. Wipf**, and **Y. Zhang**, Multi-image blind deblurring using a coupled adaptive sparse prior. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013.