Thesis for the Degree of Licentiate of Engineering

Towards Joint Super-Resolution and High Dynamic Range Image Reconstruction

Tomas Bengtsson

Department of Signals and Systems
Chalmers University of Technology
Göteborg, Sweden 2013

Towards Joint Super-Resolution and High
Dynamic Range Image Reconstruction
Tomas Bengtsson

Technical report number: R011/2013
ISSN 1403-266X

Department of Signals and Systems
Chalmers University of Technology
SE–412 96 Göteborg
Sweden
Telephone: +46 (0)31 – 772 1000

Typeset by the author using LaTeX.

Chalmers Reproservice
Göteborg, Sweden 2013

We are all in the gutter, but some of us are looking at the stars.
- Oscar Wilde

Abstract

The main objective for digital image- and video camera systems is to reproduce a real-world scene in such a way that a high visual quality is obtained. A crucial aspect in this regard is, naturally, the quality of the hardware components of the camera device. There are, however, always some undesired limitations imposed by the sensor of the camera. To begin with, the dynamic range of light intensities that the sensor can capture in its non-saturated region is much smaller than the dynamic range of most common daylight scenes. Secondly, the achievable spatial resolution of the camera is limited, especially for video capture with a high frame rate. Signal processing software algorithms can be used that fuse the information from a sequence of images into one enhanced image. Thus, the dynamic range limitation can be overcome, and the spatial resolution can be improved.

This thesis discusses different methods that utilize data from a set of multiple images, that exhibits photometric diversity, spatial diversity, or both. For the case where the images are differently exposed, photometric alignment is performed prior to reconstructing an image of a higher dynamic range. For the case where there is spatial diversity, a Super-Resolution reconstruction method is applied, in which an inverse problem is formulated and solved to obtain a high resolution reconstruction result. For either case, as well as for the optimistic and promising combination of the two methods, the problem formulation should consider how the scene information is perceived by humans. Incorporating the properties of the human vision system in novel mathematical formulations for joint high dynamic range and high resolution image reconstruction is the main contribution of the thesis, in particular of the published papers that are included. The potential usefulness of high dynamic range image reconstruction on the one hand, and Super-Resolution image reconstruction on the other, are demonstrated. Finally, the combination of the two is discussed and results from simulations are given.

Keywords: Super-resolution, Dynamic range, Image reconstruction, Inverse problem, Human visual system, Digital camera system, Spatial alignment, Photometric alignment

Preface

It gives me pleasure to present this licentiate thesis. The thesis has been organized in two parts. In the first part, the topic is introduced, taking on a broader view as compared to the second part, in which the published papers are appended. The layouts of the papers have been revised to fit the thesis format. Having spent most of my life trying to understand how things work, now is a time when I try to be extremely humble in the face of all the things I do not understand. I am inspired to make this work available under ⓪ PUBLIC DOMAIN. Thanks to "Sita sings the blues" for taking this stance, and to my dear friend Tilak Rajesh Lakshmana for reminding me that knowledge should be free.

This thesis is in partial fulfillment of the requirements for the degree of Licentiate of Engineering at Chalmers University of Technology, Göteborg, Sweden. The work has been supported in part by VINNOVA (the Swedish Governmental Agency for Innovative Systems) within the project Visual Quality Measures (2009-00071). Other project partners have been Volvo Car Corporation, Fraunhofer Chalmers and Epsilon.

Acknowledgements

I would like to thank my supervisor Professor Tomas McKelvey for giving me the opportunity to work in an inspiring environment. In particular, thank you for your kind support and guidance, and for the social evenings, including everything from the nice dinners to horseback riding. Also, thanks to my former supervisor, Professor Irene Gu, you introduced me to the field of research and it was a most interesting learning experience. To Doctor Konstantin Lindström at Volvo Car Corporation, thank you for your thoughtful input to my research and for a very innovative and memorable breakfast meeting in your home. Thanks also to Professor Mats Viberg, group leader when I applied for the Ph.D. position, for the delightful, regular wine tastings that resonate well with your outlook that life should be thoroughly enjoyed.

I am inspired to go to work in the mornings, mainly thanks to my colleagues, at the group level and at the department level. Some special mentions go out to Abu, thanks for your kind wishes at various special occasions, and not least for bringing me not one, but two Indian kurtas. Thanks also to Lennart, for your open door policy which I have gladly taken advantage of on several occasions, and for stressing the importance of good research practice. To Oskar, Nina, Malin and Lars, thank you for the numerous coffee break conversations and for the everyday support that is necessary in order to keep a Ph.D. student happy and motivated.

Thanks to friends, old and new. You know who you are. I hope I get the chance to continue to thank you for a long time still. From my childhood friends, to my study mates at Chalmers, to you who share the interest of Africa and the poor state of the world, to my dance friends who provide a place of joyfulness. To my mother Boel, my father Tage and my brother Martin, for holding everything together, thank you. I love you all. My gratitude also extends to friends yet to be met. To sources of inspiration everywhere.

Carpe diem!

Tomas Ben

Tomas Bengtsson
Göteborg, April 2013

List of publications

This thesis is based on the following two appended papers:

Paper 1

T. Bengtsson, 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, *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, Kyoto, Japan.

Paper 2

T. Bengtsson, T. McKelvey and I. Y-H. Gu, Super-Resolution Reconstruction of High Dynamic Range Images in a Perceptually Uniform Domain, *SPIE, Journal of Optical Engineering, Special Issue on High Dynamic Range Imaging* , October 2013.

Other publications

T. Bengtsson, T. McKelvey and I. Y-H. Gu, Super-Resolution Reconstruction of High Dynamic Range Images with Perceptual Weighting of Errors, *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, Vancouver, Canada.

Contents

I Introductory chapters

II Included papers

Part I

Introductory chapters

Chapter 1

Introduction

Prehistoric cave paintings are testament to the longstanding human fascination of making images of the world. The relatively modern technique of photography, which has enabled us to record realistic looking images in an instant, first saw light about 200 years ago. Earlier variants of cameras date back much further, to ancient times. Nowadays, it is safe to say that the technology has matured significantly, however much is expected still in the development of the modern digital camera technology. For instance, so called High Dynamic Range (HDR) image capture is currently emerging as a new functionality of camera devices. For a single image with a fixed exposure duration, the camera sensor hardware has a dynamic range which is often insufficient. As a result, certain images areas are either over- or underexposed. Thus, in order to produce an HDR image, information from multiple differently exposed images is combined [1]. In overcoming the dynamic range limitation, reliable HDR functionality should actually be seen as quite revolutionary. In order to fuse multiple images robustly, the images first need to be aligned to compensate for camera movement and possible movement within the image. If the pose of an object has changed from one image to the next, that has to be accounted for in order to avoid reconstruction artifacts in the fused image.

A somewhat related field of research to HDR image reconstruction is that of Super-Resolution Reconstruction (SRR) [2], which is used in order to enhance spatial resolution by utilizing several images. Thus, both techniques attempt to combine information from an image set of the same real-world scene, in order to produce a single image of high visual quality. In particular, these respective techniques may help to provide images with higher contrasts, owing to the enhanced dynamic range, and improved clarity of visible details, thanks to a higher spatial resolution. The extension of these techniques from producing a single output image to full video sequences is straightforward. A sliding window approach on the frames of the video sequence may be used to enhance each frame separately. Thus, all the discussed methods applied to reconstruct a still image could be used on video data, by simply repeating the same method for each frame. The terms image as well as video frame will be used interchangeably as seen appropriate. Furthermore, input images to SRR are referred to as a Low Resolution (LR) images, and the reconstructed image of enhanced resolution are referred to as a High Resolution (HR) image.

In the image reconstruction methods discussed throughout this thesis, the aim is to capture as much meaningful data about the original scene as possible, or as necessary. The next step, if we consider a full camera system, is concerned with how to code the raw data (all the observations of the scene), in order to for example visualize it on a display device, or for storage. Image (and video) formats that are widely used today are designed for the hardware that has been available over the last several decades. That essentially means that, due to the relatively Low Dynamic Range (LDR) of both capture and display devices historically, modeling of the Human Visual System (HVS), that serves as the basis for image coding, is lacking for high dynamic range scenarios. HDR technology was not around to influence standardization of these earlier formats, but with HDR technology now becoming more common, so is work on HDR coding for use in standardization.

SRR techniques may also be subject to future use in image coding. For example, it has been suggested for use in image compression [3]. Another area for SRR is the case where a video sequence of a given resolution should be displayed on a device with a higher resolution. This is to date typically achieved with simple interpolation. Thirdly, in terms of hardware, having a small pixel size comes at the cost of increasing the exposure duration [4], which can cause undesired effects such as motion blur. Thus, under such circumstances, the size of the pixels could be kept larger, while instead using SRR to achieve the same total resolution. Custom sensor equipment has been proposed to accommodate this [5].

1.1 Aim of the thesis

The main topic of this thesis is about the answer to the following question. Given a set of related images of the same real-world scene, how can the information in the respective images best be utilized in order to produce one enhanced image representation that is perceived to have a high resemblance with reality? Specifically, the text aims to achieve the following:

- Present a unified survey of reconstruction methods based on multiple input images. This provides a broad view of the research area, in which the contributions of the included papers are placed.
- Discuss the potential for joint image reconstruction of high resolution, high dynamic range images.
- Highlight the impact of the human visual system in the problem formulation for high dynamic range image reconstruction.

1.2 Thesis outline

This thesis is divided into two parts. In Part I, the research area of image reconstruction based on multiple images is discussed, providing a background for the two papers that are included in Part II of the thesis. The presentation of the related literature is not exhaustive. Rather, it is a selection of work which is relevant to the methods used, and in particular to the proposed method of joint HR, HDR image reconstruction (Chapter 5). In Chapter 2, an introduction to digital camera systems is given, including certain properties of human visual perception. The mathematical model for the camera that is used in the formulation of the image reconstruction methods is also presented. Chapter 3 treats reconstruction of high dynamic range images from differently exposed LDR input images. Reconstruction of images with enhanced spatial resolution, by the use of a Super-Resolution method, is discussed in Chapter 4. In Chapter 5, SRR of HDR images is discussed. First, a generic method is outlined, where similar to all previous work on joint SR, HDR, reconstruction is formulated in an unsuitable image domain. Then, an alternative method is presented, in which perceptual characteristics of human vision is taken into account in the mathematical formulation. A summary of the included papers (in Part II) is given in Chapter 6, and concluding remarks are given in Chapter 7.

Chapter 2

Introduction to digital camera systems

A digital camera is used to capture still images or video for digital reproduction of a real-world scene. The intended application may vary. In this thesis, the application is to output visually pleasing images (or video). In contrast, the camera could for example be a part of an automatic image analysis system, for which the design objectives of the camera differ. The remainder of this chapter is structured as follows. First, in the context of digital image processing, characteristics of natural scenes are discussed, as well as what is required for an image of a scene to be of high visual quality. The human eyes has certain properties when it comes to how light is perceived. These properties, as part of the Human Visual System (HVS), are discussed in the following section with relation to its relevance for camera design and digital image processing. Finally, a mathematical model for the camera is introduced.

2.1 Digital image processing overview

To reproduce an image of a natural scene, the entire digital camera system must be considered, from the characteristics of the scene itself to the final step, the observer. The critical aspects in order to enable high visual quality of the output image should then be analyzed and addressed. An overview of a general digital camera system is presented in Figure 2.1. To the left of the figure is a real-world scene, which may be observed either directly by a human observer, or on a display device as an image which has been captured and processed digitally. The intermediate steps, divided into three steps here, impact the characteristics of the output image. First, there is the camera, the capture device which collects data from the scene. Secondly, the captured data is coded suitably (in the camera itself or in a computer), such that it retains the essential information of the scene, and outputs that data to the third and final step, the display device, in a suitable format. In summary, the overall objective of the system is to enable to visualize, on some display device, a high quality image of the original scene.

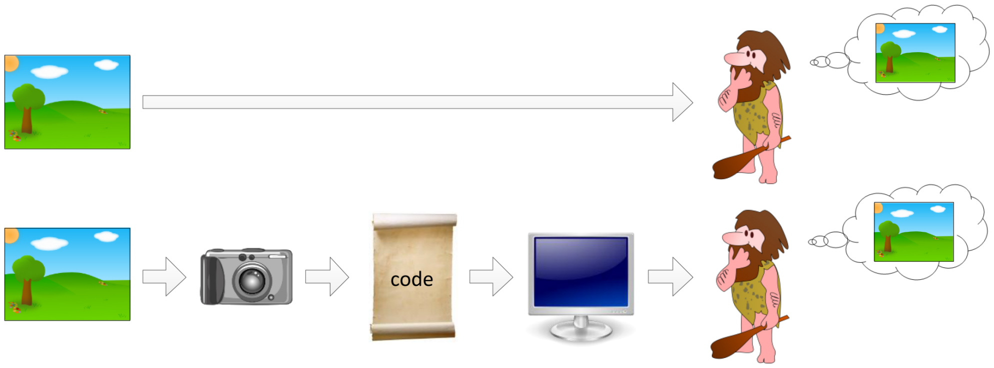

Figure 2.1: A digital camera system. Data of an original scene is captured with a camera, coded with some algorithm and visualized on a display device. The ambition is that the produced image should be perceptually similar to directly observing the scene.

A scene to be imaged is perceived as it is due to the light reflectance properties of its contained objects. An incident spectrum of light from the scene passes the lens of an eye or a camera and is registered by the cone cells in the retina of the eye or the pixel elements of the camera sensor respectively, producing a visual sensation or an image respectively. The spectral response of the sensor determines what fraction, as a function of wavelength, of the incoming light that is registered. In mathematical terms, the registered light is the inner product of the incident light spectrum and the spectral response of the sensor, thus producing a scalar output value [6], that may or may not be in the operational range of the sensor. In the case of the camera, these scalar outputs from the pixel elements is the raw data from a single image that is available for image coding.

An important question that arises is, how is image quality assessed? The question can be posed in the context of comparing an image to the underlying real-world scene, and in that case, first of all, relates to the acquisition of data. The captured image data should have a sufficient dynamic range, and it should provide a high spatial resolution with crisp (not blurred) image content, in order to be of high visual quality. Quality assessment can also be framed as comparing a degraded image (as a general example, this could for instance be a compressed image) to an original image. This has to do with how the specific available image data is coded, in order to maintain fidelity of colors, contrasts and to provide a natural looking images. The image coding aspects, of course, are equally important for the case of quality assessment with regard to the underlying scene. Some objective image quality measures, that are used at later stages of this thesis, are presented in Section 2.1.4.

The motivation for this work essentially stems from the limitations imposed by the sensor of the camera, in terms of dynamic range of registered light, as well as spatial resolution, two concepts that are discussed in the following subsections. By using the camera in Figure 2.1 to capture multiple images of the scene, the total information captured enables to produce and display an image that is free from over- and underexposure, and has a high spatial resolution, both crucial properties for a high perceived visual quality.

2.1.1 Dynamic range

For some arbitrary positive quantity Q, the dynamic range is defined as the ratio of the largest and smallest value that the quantity can take, that is

$$DR(Q) = Q_{max}/Q_{min}. \tag{2.1}$$

For analogue signals that contain noise, this definition is too vague. Thus, consider a signal Q that is the input signal to a sensor, with the logarithm of Q plotted against the (normalized) output in Figure 2.2. At low signal

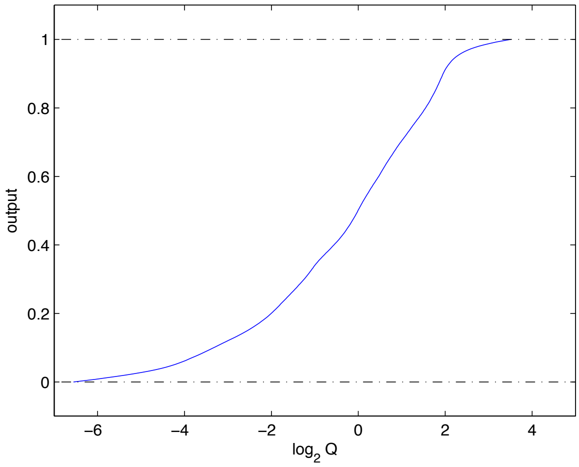

Figure 2.2: The input-output relationship for a signal Q to a sensor.

levels, the signal is drowned in electrical noise. At some level, denoted Q_{min}, the signal becomes statistically distinguishable from the noise. Similarly, at signal levels above Q_{max}, the signal saturates the sensor. These definitions are thus used in (2.1). If Q is digitized, Q_{min} and Q_{max} are fixed as the lowest and highest quantization levels.

The dynamic range of an image of a real-world scene refers to the light, in the unit of illuminance[1], that is incident on each individual sensor pixel element,

$$X = \int_{-\infty}^{\infty} I(\lambda)V(\lambda)d\lambda, \tag{2.2}$$

where $I(\lambda)$ is the incident light spectrum on the surface of the sensor element and $V(\lambda)$ is the spectral response of the sensor element, specifically of its color filter layer. Let $\mathbf{X}$ be an image which consists of the illuminance values, given as in (2.2), of all pixels of the camera sensor. Then, the dynamic range of a given, pixelated scene is $DR(\mathbf{X}) = max(\mathbf{X})/min(\mathbf{X})$.

As such, a general image $\mathbf{X}$ has no dynamic range restrictions. However, for an image generated from a single camera exposure, things are different. Depending on the brightness level of the scene, the camera sensor is exposed for an appropriate duration Δt. Thus, the sensor exposure is

$$E = \int_{t_0}^{t_0+\Delta t} X(t)dt. \tag{2.3}$$

For the mathematical modeling of the camera, however, it is assumed that $X(t)$ is constant over the time interval of the exposure, thus $E = \Delta t X$. A camera sensor element has a fixed interval $[E_{min}, E_{max}]$ of absolute exposure values that provides a signal-to-noise ratio that is deemed to be satisfactory (a design choice). The dynamic range of the camera sensor is then $DR(E) = E_{max}/E_{min}$. Unfortunately, this sensor dynamic range is often lower than that of real-world scenes, which causes the sensor to be either over- or underexposed. However, by varying Δt between different images (or alternatively, varying the aperture setting), diverse scene content in terms of illuminance values can be captured, and the information fused into one HDR image $\mathbf{X}$.

Direct sunlight corresponds to an illuminance in the order of 10^5 Lux, while a clear night sky is on the order of 10^{-3} Lux [7]. These conditions are naturally never experienced simultaneously. However, common real-world scenes, such as an indoor scene with a sunlit window, or a daytime outdoor environment containing shadow areas, have a dynamic range that often

[1] If $V(\lambda)$ is the luminous efficacy curve, X is a photometric *illuminance* value. In this thesis, however, the term illuminance is used for X as long as $V(\lambda)$ approximately mimics the human perception.

greatly exceeds that of the camera sensor of professional cameras. Table 2.1 presents an illustrative example of the dynamic range for the different parts of the digital camera system portrayed in Figure 2.1. A scene may, not uncommonly, contain a dynamic range of about 10^5, which is about the level that the HVS can perceive at a given adaptation level. The HVS is able to adapt to illuminance differences up to ten orders of magnitude, under varying conditions. A camera typically only captures a dynamic range on the order of 10^3 in each image. In the field of photography, the dynamic range of a camera is typically expressed in the base-2 logarithm, as the number of *Stops* $S = \log_2(DR)$, in the unit Exposure Value (EV).

	Dynamic Range	Stops
Original real-world scene	10^5	16.6
Camera (capture device)	10^3	9.97
LCD monitor (display device)	10^3	9.97
Human visual system (observer)	10^5	16.6

Table 2.1: An example with representative dynamic range values, where the real-world scene has a high dynamic range.

Typically, to visualize HDR content on a display device, such as an LCD monitor, a dynamic range restriction is presented yet again, due to that display devices have a low dynamic range. This issue is, however, practically overcome by *tonemapping* the HDR image information to an LDR image in such a way that it, to the HVS, is perceived similarly as the original image that it was created from [8]. An *LDR image* in this context means that the image is coded in a device independent format (for which the concept of dynamic range is somewhat unspecific) that is appropriate for output on LDR devices.

2.1.2 Spatial resolution

In a digital camera, a scene is imaged by a sensor that consists of a discrete set of pixel elements in a planar array. The number of pixels horizontally times the number of pixels vertically is the pixel resolution of the sensor. This typically exceeds the pixel resolution of digital display devices, which then determines the spatial resolution of the full system in terms of pixels per inch (PPI). If a digital image is to be printed on a paper, the dots per inch (DPI), a term related to but with a slightly different meaning than PPI, should be relatively high to obtain a high quality of a print of relatively large size. Thus, for that purpose, a high pixel resolution of the image is required.

The term *spatial resolution* refers to pixels per unit length. However, it is also often used, in a non-strict manner, as a term for the pixel resolution of a digital image, and in doing so effectively gives a distinction from the related temporal resolution of video frames. To emphasize the spatial dimension, spatial resolution is used with its wider meaning throughout this thesis.

For a fixed size of the sensor chip, the natural way to increase the pixel resolution is to reduce the size of the pixel elements. However, reducing the size of a pixel also reduces its light sensitivity. Thus, in order to reach the same Signal-to-Noise Ratio (SNR) in the sensor element, the exposure duration Δt needs to be increased [4]. That is, there is a tradeoff between two desired properties. An increase in the pixel resolution gives a requirement for a longer exposure duration, which reduces the temporal resolution that is essential for video capture, and makes images more susceptible to motion blur. Additionally, to manufacture sensors with smaller pixel elements comes with a higher cost. Generally speaking, increasing the size of the image sensor helps to improve image quality. Even so, enlarging the sensor size is not feasible for devices that are required to be compact. The above tradeoff, as well as the cost benefit, serves as a motivation for Super-Resolution techniques to be used.

2.1.3 Color properties of the camera sensor

The standard digital camera is equipped with a so called *Bayer filter*, which is an array of color filters, on top of its sensor elements. Only the light that passes through the filter is converted to electrical signals in the sensor elements. Figure 2.3 shows the mosaic pattern of the Bayer filter on top of the sensor elements, displayed in grey.

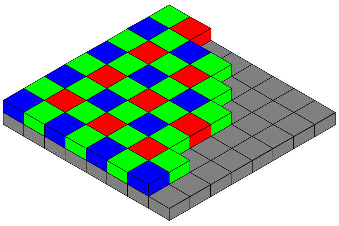

Figure 2.3: The color filter array of the Bayer pattern.

The color filter elements are designed so that they roughly match the average human eye. Thus, red, green and blue (RGB) color primaries are used, although their spectral responses may differ between different vendors (thus, there are numerous RGB color spaces). The signal at each sensor element, that was presented in (2.2), can now be specificed further as

$$X^c = \int_{-\infty}^{\infty} I(\lambda) V^c(\lambda) d\lambda, \tag{2.4}$$

where $V^c(\lambda)$ is the spectral response for either of the red, green or blue filters, $c = \{r, g, b\}$, in the Bayer pattern. Each pixel only has information about one of these color channels. To obtain values for the two missing color components, an interpolation process called *demosaicing* is performed. The demosaicing could alternatively be formulated within the Super-Resolution framework, as discussed by Farsiu et al. in [9]. Commonly, however, the SRR is performed on demosaiced images. Thus, the color filter process, which registers different color spectra for the same scene content depending on how the images are shifted relative to each other, is not modeled. This is the approach taken in this thesis. Greyscale images, sometimes used for experimental simulations, are given as a function of the r,g,b-values of demosaiced images.

2.1.4 Image quality measures

Image quality assessment is a delicate matter, much due to the perception of the HVS. Proposed objective quality measures are thus tested and assessed for how well they correlate with quality scores from extensive subjective test procedures on human subjects. Even for the use of more established objective quality measures, the evaluated images should be presented alongside to enable visual inspection.

Objective image quality measures can be categorized in the two classes of reference quality measures and no-reference quality measures. The former, where an image of interest is assessed with relation to a second image, a reference image, is (by far) the most common. No-reference quality assessment is only practically applicable for the case where the type of degradation is known, for instance a JPEG compressed image could be assessed without the uncompressed original at hand. Other criteria for no-reference quality assessment could be to estimate the sharpness of an image, or the proportion of saturated image areas. No-reference image measures can be used to determine the respective weights when fusing multiple images by weighted average, for example in order to give saturated image areas less weight.

For the case of reference image quality assessment, the Mean Structural Similarity (MSSIM) index provides relatively reliable results [10]. Unlike the Peak Signal-to-Noise Ratio (PSNR), which is useful in many applications of signal processing, but at best provides a crude benchmark for image processing, the MSSIM method compares image structure rather than individual pixels by themselves. In fact, the MSSIM is a product of a mean intensity comparison (for image blocks), a constrast comparison and a structure comparison. For more details on MSSIM (and its superiority to PSNR), refer to the original paper by Wang et al. [10]. MSSIM, and several other quality measures, treats each color channel individually, and thus says nothing about the quality of how colors are perceived. Color fidelity, instead, relies on the use of a proper color space.

2.2 The Human Visual System

So far, an image has mainly been referred to as a discrete set of pixel values in the illuminance domain. However, digital images are typically stored or processed in standardized pixel value domains of a relatively low bit depth. How, then, are these images related to the discussed illuminance images? The answer to that question stems from the properties of the Human Visual System, some of which are discussed here.

To begin with, the human visible spectrum is, roughly, light of wavelengths $\lambda \in [400, 700]$ nm. Furthermore, the spectral sensitivity of the eye differs depending on the wavelength within the visible spectrum, as a consequence of the composition and properties of the three different types of cone receptor cells (responsible for daytime vision) in the eyes. In combination, the spectral responses of each cone type determine both how colors are perceived as well as perceived brightness. For greyscale vision, which is conceptually simpler, the luminous efficacy curve describes what fraction of light at each wavelength that contributes to greyscale illuminance.

The registered illuminance is in turn interpreted by the brain in a highly nonlinear manner. Perceived brightness as a function of illuminance is approximately logarithmic, although more accurate models are used in practice. The key feature is that the eye is more sensitive to differences in illuminance at low levels than at high absolute illuminance levels [6]. To accommodate this feature, the exposure of a camera image (proportional to the illuminance) is *gamma compressed* by a nonlinear concave function before it is quantized to a lower bit depth. This is the case, for example, in standard 8-bit LDR formats. The visual sensation is additionally influenced by the brightness of the area surrounding a viewed object on different scales, both by the immediate surround but also by the overall brightness level of the background.

As for color vision, different light spectra can produce the same perceived color. Furthermore, the same visual sensation can be expressed using different sets of three basis functions, referred to as color primaries. In color science, several subjective terms are defined and objectified as standardized units, in order to quantify effects of image processing. To exemplify, some color spaces aim to define a basis of color primaries in which color, as perceived by the HVS, is uniformly distributed, some aim to orthogonalize perceived brightness on the one hand and color sensation on the remaining two basis functions. The property of color uniformity are not well fulfilled by r,g,b-spaces (among other), which may lead to a loss of color fidelity as a result of image processing in the r,g,b-space.

As far as this thesis is concerned, the proposed image reconstruction method in Chapter 5 addresses the nonlinear relation of illuminance to perceived visual brightness. This is the property that will otherwise cause the most severe reconstruction artifacts, should it not be considered, due to that small reconstruction errors in terms of illuminance have a large perceptual impact in dim image areas.

2.2.1 Perceptually uniform image domains

In the traditional LDR case, image processing is performed in various perceptually uniform image domains. For example, gamma compressed r,g,b spaces (often denoted r',g',b') are approximately perceptually uniform with respect to brightness, although no special care has been taken to assure color fidelity is maintained when manipulating the image in that domain. For the *L*a*b** color space, the L*-component is essentially the cube root of the greyscale illuminance (which is in turn a linear function of the r,g,b-values), and thus an approximation for subjective brightness, sometimes denoted *Lightness*. The a* and b* components are so called *color opponent* dimensions, that express the color sensation in a way which is perceptually orthogonal to the lightness dimension. Conventional color spaces such as L*a*b* are however not directly applicable to HDR data, because they are typically designed based on modeling of the HVS for a lower dynamic range. Thus, the modern HDR capabilities should serve as a motivation to advance new HDR formats.

Hence forth, any image domain that attempts to approximate the nonlinear behavior of the HVS, in particular the nonlinear response of perceived brightness as a function of illuminance, will be denoted a Perceptually Uniform (PU) domain. Objective quality measures, such as the ones discussed in Section 2.1.4, should be applied in a PU domain.

2.3 Camera model

This section presents a mathematical model of a digital camera, which is later used to derive formulations of image reconstruction methods. Specifically, the images that the camera delivers will be used as input to methods that aim to enhance their dynamic range, spatial resolution, or both. Consider a sequence of high quality digital images, $\{\mathbf{X}_k\}, k = 1, \ldots, K$, each of size (resolution) $X_1 \times X_2$, that are in the illuminance domain. These images are merely a modeling construction, representing undegraded versions of the actual available images, $\{\mathbf{Y}_k\}, k = 1, \ldots, K$, as depicted in Figure 2.4. The $\mathbf{Y}_k$ images are observations of the $\mathbf{X}_k$ images, according to the camera model introduced shortly in this section. Both $\mathbf{Y}_k$ and $\mathbf{X}_k$ are images, of different quality, of an underlying real-world scene.

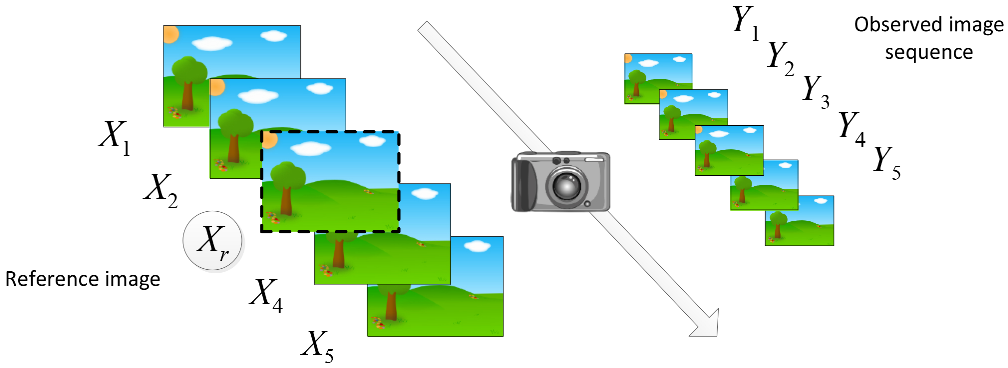

Figure 2.4: An example of $K = 5$ observed images $\mathbf{Y}_k$, that could be used to reconstruct a reference image $\mathbf{X}_r$ of, for example, a higher resolution or a higher dynamic range.

Because images are assumed to be taken in a sequence, for instance with a single hand-held camera, the $\mathbf{X}_k$ will generally differ, both due to camera movement and due to motion within the scene. To express the relation between the $\mathbf{X}_k$, let $\mathbf{X}_r$ denote a reference image, that should later be reconstructed from $\{\mathbf{Y}_k\}$. Assuming brightness constancy of scene objects, let the other images be related to the reference according to

$$\mathbf{X}_k(i,j) = \mathbf{X}_r(i + D^x_{k,r}(i,j), j + D^y_{k,r}(i,j)) \tag{2.5}$$

where (i,j) is the pixel location in the image array and $D^x_{k,r}(i,j)$ and $D^y_{k,r}(i,j)$ denote respectively the horizontal and vertical components of the displacement field

$$\mathbf{D}_{k,r}(i,j) = (D^x_{k,r}(i,j), D^y_{k,r}(i,j)) \tag{2.6}$$

that describes the (local) motion of each pixel in image k relative to the reference image. Note that, due to occlusion, there may be pixels for which no displacement vector exists. Furthermore, since pixel indexes are integer numbers, the displacements, with this formulation, are limited to be integer numbers. Hence forth, a matrix-vector representation is used to represent images and image operations. Using $\mathbf{x}_k = vec(\mathbf{X}_k)$, of size $(X_1X_2) \times 1 \triangleq n \times 1$, equation (2.5) is re-expressed as

$$\mathbf{x}_k = \mathbf{T}\{\mathbf{D}_{k,r}\}\mathbf{x}_r, \tag{2.7}$$

where $\mathbf{T}\{\mathbf{D}_{k,r}\}$ is a matrix of size $n \times n$, parameterized by the $X_1 \times X_2 \times 2$ displacements $\mathbf{D}_{k,r}(i,j)$, that relate $\mathbf{x}_k$ and $\mathbf{x}_r$ through a warping operation. With this formulation, non-integer displacements (with respect to the pixel grid of $\mathbf{x}_r$) are straight-forwardly expressed in $\mathbf{T}\{\mathbf{D}_{k,r}\}$, thus relating multiple pixels in $\mathbf{x}_r$ to a certain pixel in $\mathbf{x}_k$. For further details as well as an example of relating images with the displacement field, refer to the book by Katsaggelos et al. [3].

The camera model that provides observations $\mathbf{y}_k = vec(\mathbf{Y}_k)$, of size $(n/L^2) \times 1$, is

$$\mathbf{y}_k = f(\Delta t_k \mathbf{R}\mathbf{C}\{\mathbf{H}_k\}\mathbf{x}_k + \mathbf{n}_k) + \mathbf{q}_k. \quad k = 1, \ldots, K \tag{2.8}$$

For each of the multiple observations, $\mathbf{C}\{\mathbf{H}_k\}$ of size $n \times n$ represents 2d convolution on the vectorized HR image $\mathbf{x}_k$ with the convolution kernel $\mathbf{H}_k$ of support $H_1 \times H_2$. Different assumptions are made for $\mathbf{H}_k$, with respect to what it models and what its parametrization is, depending on the reconstruction method employed, as discussed further in the next couple of sections. The downsampling matrix $\mathbf{R}$, of size $(n/L^2) \times n$, decimates the spatial resolution a factor L in the x- and y-direction, and Δt_k is the exposure duration. The noise in the camera sensor is modeled by $\mathbf{n}_k$ and $\mathbf{q}_k$ represents quantization noise, both are of size $(n/L^2) \times 1$.

The exposure on the camera sensor is $\mathbf{e}_k = \Delta t_k \mathbf{R}\mathbf{C}\{\mathbf{H}_k\}\mathbf{x}_k + \mathbf{n}_k$. For each pixel $i \in \{1, \ldots, n/L^2\}$, the exposure $[\mathbf{e}_k]_i$ is mapped by the pixelwise, nonlinear Camera Response Function (CRF),

$$f(E) = \begin{cases} 0 & , E \leq E_{min} \\ f_{op}(E) & , E_{min} \leq E \leq E_{max} \\ 1 & , E \geq E_{max} \end{cases}, \tag{2.9}$$

where f_{op} is a concave mapping to quantized 8-bit pixel values, $Y \in \{0, \ldots, 1\}$, in the PU (LDR) image domain of $\mathbf{y}_k$. The CRF has an operational range of exposure values, $[E_{min}, E_{max}]$, which does not cause over- or underexposure. Exposure values outside of this interval are clipped by the CRF and cannot be recovered (from that single image). This is what causes the observed images to be of low dynamic range. For example, $[E_{min}, E_{max}] = [0.01, 10]$ gives a sensor dynamic range of 10^3, as in the fictive example of Table 2.1. The CRF is made up of several nonlinear components of the physical camera capture process [6]. On top of that, it is adjusted in the design process to achieve the purpose of mapping the sensor exposure data to a PU output domain. For simulation purposes, $f_{op}(E)$ in the CRF may be modeled as a parametric function, for example

$$f_{op}(E) = \left(\frac{E - E_{min}}{E_{max} - E_{min}}\right)^{\gamma_{LDR}}, \tag{2.10}$$

where the choice of $\gamma_{LDR} = 1/2.2$ is the same exponent as often used for gamma correction applications. This description of $f_{op}(E)$ helps to contextualize the design of a similar concave mapping to a PU domain in the HDR scenario, for instance to be used in the formulation of image reconstruction methods, as is discussed in Chapter 5.

Quantization of the input signal takes place twice. First, the Analog-to-Digital (A/D) converter digitizes the exposure data to a relatively high bit depth, typically 12-14 bits. This effect takes place before the CRF, and is thus taken to be part of $\mathbf{n}_k$. Then, after the mapping by $f(\cdot)$, the image is quantized to the 2^8 uniformly spaced quantization levels. In a device independent interpretation, the quantization levels are commonly referred to as pixel values in the (integer) set $\{0, \ldots, 255\}$.

In summary, the observed images $\mathbf{y}_k$, generated by (2.8), are related to $\mathbf{x}_r$ due to (2.7). An overview of the generative process is shown in Figure 2.5. A spectrum of light from an original scene is incident on a pixel grid, included in the figure to stress that no attempt is made to include demosaicing, discussed in Section 2.1.3, in the model. Then, the image $\mathbf{x}_r$, which may be thought of as a single channel greyscale image, or to contain (demosaiced) r,g,b information, may be warped, blurred and downsampled, as decided by

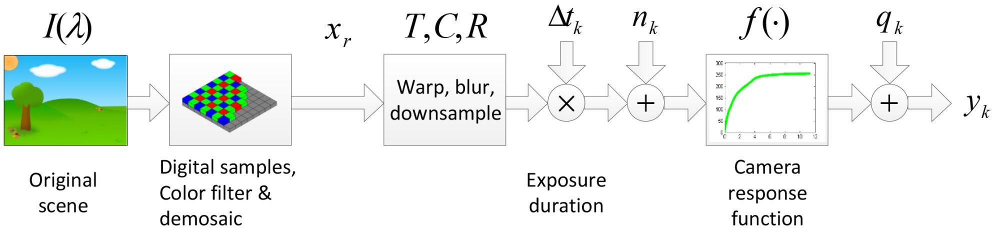

Figure 2.5: The generative camera model.

the scenario of interest to model. The exposure image is then mapped by the CRF and finally quantized to produce $\mathbf{y}_k$.

In the following chapters, image sets $\{\mathbf{y}_k\}$ are used to reconstruct images of enhanced dynamic range (Chapter 3), spatial resolution (Chapter 4), and of both enhanced dynamic range and spatial resolution (Chapter 5). Ultimately, the ambition is to reconstruct (estimate) a HR, HDR image $\mathbf{x}_r$, but the more restrictive reconstruction methods are treated along the way.

Chapter 3

High Dynamic Range Image Reconstruction

This chapter discusses how an HDR image can be reconstructed from a set of images, $\{\mathbf{y}_k\}$. For this reconstruction method, as for the ones presented in the next chapters, specific assumptions are made with the respect to the operators in the generative model for $\mathbf{y}_k$, which is presented in general terms in (2.8). Here, no downsampling is included, which means that no attempt is made to enhance the spatial resolution. In terms of the model in (2.8), $\mathbf{R} = \mathbf{I}$, where $\mathbf{I}$ is the Identity matrix. The blur matrix $\mathbf{C}\{\mathbf{H}_k\}$ is excluded as well. That is not to say that there is no blur in the images, it is just not modeled.

Based on the above, assume that there is an HDR image $\mathbf{x}_1$ (the reference image), observed through the differently exposed LDR images

$$\begin{aligned} \mathbf{y}_1 &= f(\Delta t_1 \mathbf{x}_1 + \mathbf{n}_1) + \mathbf{q}_1, \\ \tilde{\mathbf{y}}_2 &= f(\Delta t_2 \mathbf{T}\{\mathbf{D}_{2,1}\}\mathbf{x}_1 + \tilde{\mathbf{n}}_2) + \tilde{\mathbf{q}}_2, \end{aligned} \tag{3.1}$$

where $\Delta t_1 < \Delta t_2$ is a short exposure duration that results in underexposure in dim image areas, and Δt_2 is a longer exposure duration that causes bright image areas to be overexposed. The two images have a high combined dynamic range, which should ideally be larger than the dynamic range of the original scene, in order to completely avoid over- and underexposure in $\mathbf{x}_1$.

The first step, in order to reconstruct $\mathbf{x}_1$, is to align the observed images to a reference image. In this case, $\tilde{\mathbf{y}}_2$ is first spatially aligned to $\mathbf{y}_1$ by a reverse warping to yield

$$\mathbf{y}_2 = \mathbf{T}\{\mathbf{D}_{1,2}\}\tilde{\mathbf{y}}_2. \tag{3.2}$$

If the displacement fields between $\mathbf{y}_1$ and $\mathbf{y}_2$ adhere to a global translational model, that is $D^x_{2,1}(i,j) = D^x_{2,1}, D^y_{2,1}(i,j) = D^y_{2,1}, \forall(i,j)$, and the translational shifts are integer numbers of pixels, it follows that, neglecting the image boundaries that are shifted out of the image, $\mathbf{T}\{\mathbf{D}_{1,2}\}\mathbf{T}\{\mathbf{D}_{2,1}\} = \mathbf{I}$. Furthermore, because $f(\cdot)$ is a pixelwise function,

$$\begin{aligned} \mathbf{y}_1 &= f(\Delta t_1 \mathbf{x}_1 + \mathbf{n}_1) + \mathbf{q}_1, \\ \mathbf{y}_2 &= f(\Delta t_2 \mathbf{x}_1 + \mathbf{n}_2) + \mathbf{q}_2. \end{aligned} \tag{3.3}$$

Thus, $\mathbf{y}_1$ and $\mathbf{y}_2$ are two differently exposed, spatially aligned observations of $\mathbf{x}_1$. If, on the other hand, the translational shifts are non-integer numbers, $\mathbf{y}_1$ and $\mathbf{y}_2$ will not be perfectly aligned as suggested by (3.3). This is because, in that case, interpolation is included in $\mathbf{T}$, and thus $\mathbf{T}\{\mathbf{D}_{1,2}\}\mathbf{T}\{\mathbf{D}_{2,1}\} \neq \mathbf{I}$. Rotation, change of scale or more complex local motion all likewise give rise to interpolation in $\mathbf{T}$. Furthermore, because the reverse warp operator, $\mathbf{T}\{\mathbf{D}_{1,2}\}$, is applied outside of $f(\cdot)$, another small imperfection occurs. These effects are in practice always the case, since the subpixel displacements are arbitrary in an uncontrolled environment. If occlusion occurs, some image parts are not possible to align at all. Imperfections in the alignment are not desired, however they may not be crucial for this application, since, on average, adjacent pixels (that incorrectly spill over due to alignment errors) have similar pixel values.

Given a set of K spatially aligned images $\mathbf{y}_k$, for instance $K = 2$ as above, or a larger number, photometric alignment should be performed in order to obtain $\mathbf{x}_1$. This is achieved by mapping the $\mathbf{y}_k$ images with the approximate inverse of the CRF, denoted by $g(\cdot)$ ($\simeq f^{-1}(\cdot)$, barring quantization and saturation effects in $f(\cdot)$), and dividing the resulting exposure values with their respective exposure durations to retrieve the (estimated) illuminance values.

In order to estimate the (non-parametric) function $g(Y)$, using the method of Debevec and Malik [1], a set of P pixel positions are selected at random, to provide sample points from each of the $\mathbf{y}_k$. If some image areas were not possible to align spatially, these should be avoided in the selection of the sample points. The $g(Y)$ function is estimated for all input values it can take, $Y \in \{Y_{min}, \dots, Y_{max}\} = \{0, \dots, 255\}$, jointly with the unknown illuminance values $[\mathbf{x}_r]_i$ of the P sample point pixel positions $i \in \mathbf{p}$, by minimizing

$$\begin{aligned} &\sum_{i \in \mathbf{p}} \sum_{k=1}^{K} \{w([\mathbf{y}_k]_i)[\ln(g([\mathbf{y}_k]_i)) - \ln([\mathbf{x}_r]_i) - \ln(\Delta t_k)]\}^2 + \\ &+ \alpha \sum_{Y=Y_{min}+1}^{Y_{max}-1} w(y) g''(Y)^2, \end{aligned} \tag{3.4}$$

where

$$w(Y) = \begin{cases} Y & ,Y \leq 127 \\ 255 - Y & ,Y > 127 \end{cases} \tag{3.5}$$

is a function that is designed to give a higher weight to image data in the middle of the exposure range, which typically exhibits the best SNR. As seen in (3.4), the minimization is performed in the logarithmic domain, which is much closer to perceptual uniformity than linear illuminance. A smoothness term with weight parameter α is used to enforce a slowly changing slope of $g(Y)$ in the solution. The second derivative can for example be implemented as $g''(Y) = g(Y-1) - 2g(Y) + g(Y+1)$. The objective is easily re-written in a matrix formulation, and the optimum is obtained by solving a standard Least Squares problem in a matrix formulation, see [1] for details. The total number of unknowns are $256 + P$. Thus, disregarding the influence of the smoothness term, P and K should be chosen to fulfill $(P-1)K > 256$. More points can readily be used for a more robust estimator.

In Figure 3.1, an estimated $g(Y)$ function is shown. The relation between pixel values $Y \in \{0, \ldots, 255\}$ to the exposure $E \in \{g(0), \ldots, g(255\} = \{E_{min}, \ldots, E_{max}\} = \{0.0106, \ldots, 11.383\}$ is depicted in Figure 3.1 (a). The dynamic range of the camera is thus $DR(E) = 1.07 \cdot 10^3$. Figure 3.1 (b) shows the operational range of illuminance values, $[E_{min}, E_{max}]/\Delta t_k$, plotted in the base-2 logarithmic domain, for each of the $K = 4$ differently exposed images. That is, the horizontal axis shows $\log_2 E$ shifted by $-\log_2(\Delta t_k)$, for each of the exposure durations. The dashed green line is $\log_2 E$ itself (equivalent in values to illuminance, should $\Delta t_k = 1$). The exposure durations used in the example are $\{\Delta t_1, \Delta t_2, \Delta t_3, \Delta t_4\} = \{3.2, 0.8, 0.25, 0.0167\}$. The combined dynamic range captured is,

$$2^{([\log_2(E_{min}) - \log_2(\Delta t_1)] - [\log_2(E_{max}) - \log_2(\Delta t_4)])} = 2.06 \cdot 10^5.$$

Generally speaking, if the exposure durations are selected with appropriate care, as few as 2 images $\mathbf{y}_k$ are often sufficient to capture HDR scenes. At the least, 2 images give a substantial improvement compared to a single image, in terms of overcoming dynamic range limitations of the camera. An alternative to estimating $g(\cdot)$ as in the method of Debevec and Malik, discussed above, is to use a parametric approach. For example, Choi et al., use a third degree polynomial parameterization of $g(\cdot)$, as the inverse of f_{op} in (2), and estimate the polynomial coefficients [11].

With the estimated $g(\cdot)$ at hand, the illuminance information of the LDR images is obtained as

$$\mathbf{i}_k = g(\mathbf{y}_k)/\Delta t_k, \tag{3.6}$$

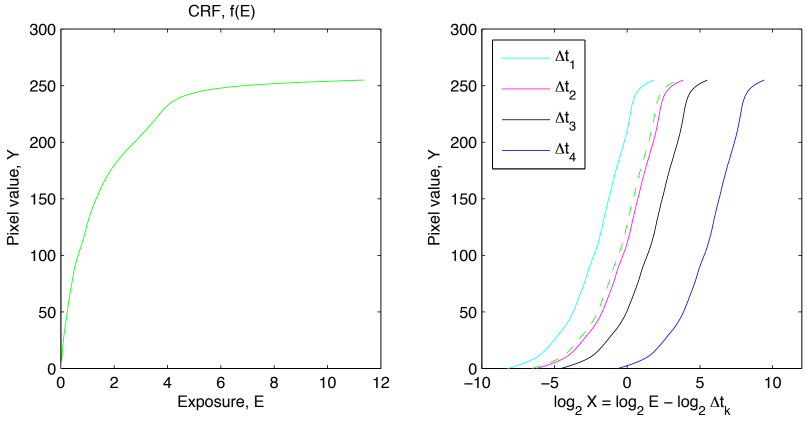

Figure 3.1: From left to right: (a) Results of estimating the inverse CRF from the four LDR images in Figure 3.2, using the method of Debevec and Malik [1]. (b) A plot that shows the combined dynamic range of the LDR observations.

such that they become photometrically aligned in the same domain. In the method of Debevec and Malik, the $\mathbf{i}_k$ images are fused by pixelwise weighted average in the logarithmic (PU) illuminance domain [1]. That is, the pixels values of the reconstructed HDR image $\mathbf{x}_r$ are given as

$$[\mathbf{x}_r]_i = \exp\left(\frac{\sum_{k=1}^{K} w([\mathbf{y}_k]_i)(\ln g([\mathbf{y}_k]_i) - \ln \Delta t_k)}{\sum_{k=1}^{K} w([\mathbf{y}_k]_i)}\right). \tag{3.7}$$

Note that a zero weight ($w(Y)$ as in (3.5)) is given to pixels valued 0 or 255, that are likely to be saturated. To exemplify the reconstruction of a HDR image, consider the set of $K = 4$ spatially aligned, differently exposed images

$$\mathbf{y}_k = f(\Delta t_k \mathbf{x}_r + \mathbf{n}_k) + \mathbf{q}_k, \tag{3.8}$$

taken with the same exposure durations as above. Such an image set, as shown in Figure 3.2 (a)-(d), is often referred to as an *Exposure stack*. It is used here to reconstruct $\mathbf{x}_r$ according to (3.7).

In order to display the reconstructed HDR image, which has a dynamic range that exceeds that of typical LDR display devices, such as commercial digital monitors or printers, it is tonemapped to an LDR format suitable for visualization. Figure 3.2 (e) and (f) show two different tonemapped results, using the simple tonemapping function in Matlab (e) and the more sophisticated tonemapping function of iCAM06 [8], which is able to better preserve a natural look of colors. The next section gives an overview of existing tonemapping operators.

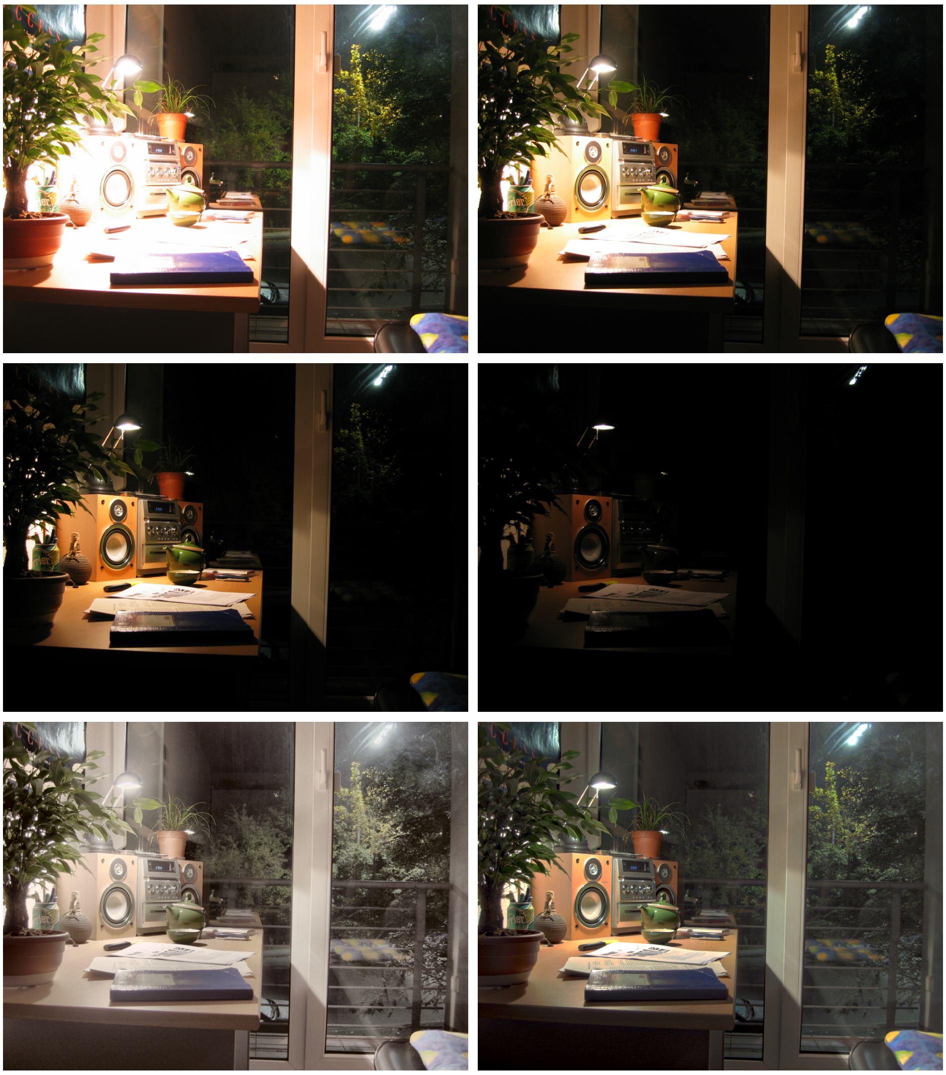

Figure 3.2: Top and middle rows: (a)-(d), Differently exposed LDR input images. Bottom row: Tonemapped HDR result, using the method of Matlab to the left (e), and iCAM06 [8] to the right (f).

3.1 Tonemapping of High Dynamic Range images

Whether an image is LDR or if it contains HDR content, it is typically stored in a computer in a device-independent format, commonly with three 8-bit color channels. The discrete pixel values, $\{0, \ldots, 255\}$, are interpreted by the display device's driver files, and thus mapped to appropriate output luminance values depending on the dynamic range of the device. For conventional LDR images, the mapping from raw sensor data to pixel values is done using standard, well established mappings, that include some form of gamma compression to a PU domain.

HDR images, such as the $\mathbf{x}_r$ discussed above in this chapter, should also be stored in some device-independent format. However, due to the higher dynamic range, the 256 quantization levels that 8-bit formats offer is a bit more restrictive, and higher bit depths may be desirable. At this stage, however, much research is done on how to tonemap HDR images to a PU 8-bit domain, for visualization on LDR devices. The simplest tonemapping operators (TMO) simply compress the HDR data linearly by a pixelwise, global function, however in a PU domain rather than directly in the illuminance domain. The Matlab TMO does just this, with the compression of the dynamic range taking place in the L*a*b* domain. As was seen in the result of Figure 3.2 (e), the Matlab TMO does not preserve colors well, hinting that compressing the dynamic range in the L*a*b* domain for a HDR image may not be the best choice.

More sophisticated methods perform various kinds of local processing, depending on the surrounding image content. For example, the iCAM06 TMO separates the image into a base layer (low-pass filtered image) and a detail layer, and performs different operations on each layer [8]. Contrasts are compressed only for the base layer, that is, across different image segments, rather than on the details within image segments. This method also takes into account background light conditions, and furthermore compensates for various other (peculiar) effects of perception. The various operations in the iCAM06 TMO, in addition, are implemented in a number of different color spaces.

To judge how well a TMO performs its task, subjective evaluation is used for a set of essential perceptual attributes. For a survey of this sort, see for example the work by Cadik et al. [12]. A conclusion that is drawn by the authors from their survey is that, while local processing or multiresolution decompositions may be of use, the most essential part in order to obtain good perceptual results is how the actual dynamic range compression is performed (globally). That is, it is crucial to select a color space (more generally denoted as image domain) that is perceptually uniform, both with regard to brightness and color sensation.

Chapter 4

Super-Resolution Image Reconstruction

In this chapter, a set $\{\mathbf{y}_k\}$ of low resolution images are used to reconstruct, by the Super-Resolution method, an image $\mathbf{x}_r$ of a higher resolution. All LR images are assumed to be taken with the same exposure duration. Thus, the reconstructed image will be a low dynamic range image as well. What makes SRR work is that each of the $\mathbf{y}_k$ images provides new information of $\mathbf{x}_r$, as depicted in the example of Figure 4.1. The images thus need to be shifted relative to each other by non-integer subpixel level shifts, or blurred by different (known or estimated) blur functions. The LR image $\mathbf{y}_r$ in Figure 4.1 (b) provides information about $\mathbf{x}_r$ in Figure 4.1 (a), but it is not sufficient to determine, for example based on the upper-left pixel value, what all four pixel values should be in the corresponding location of $\mathbf{x}_r$. Taking into consideration more observations, such as those in Figure 4.1 (c) and (d), additional information about $\mathbf{x}_r$ is given.

For the discussion on SRR in the traditional case where the $\mathbf{y}_k$ have the same exposure setting, we divert from the camera model presented in (2.8) and alternatively use the camera model

$$\begin{aligned} \mathbf{y}_k &= \mathbf{RC}\{\mathbf{H}_k\}\mathbf{T}\{\mathbf{D}_{k,r}\}f(\Delta t\mathbf{x}_r) + \mathbf{n}_k = \\ &= \mathbf{RC}\{\mathbf{H}_k\}\mathbf{T}\{\mathbf{D}_{k,r}\}\mathbf{z}_r + \mathbf{n}_k, \qquad k = 1, \ldots, K \end{aligned} \tag{4.1}$$

where the quantization noise term is left out of the expression, instead considered to be included in $\mathbf{n}_k$. The HR image $\mathbf{z}_r = f(\Delta t\mathbf{x}_r)$ is estimated directly in the pixel domain, due to $\Delta t_k = \Delta t, \forall k$. This is the camera model that has been used traditionally in the literature on SRR of LDR images. It was first when differently exposed images were considered, primarily in the last couple of years, that authors on the topic of SRR for HDR images adopted the model in (2.8), which is more natural considering the physics of the camera, see for example Gevrekci and Gunturk [13]. It is possible that

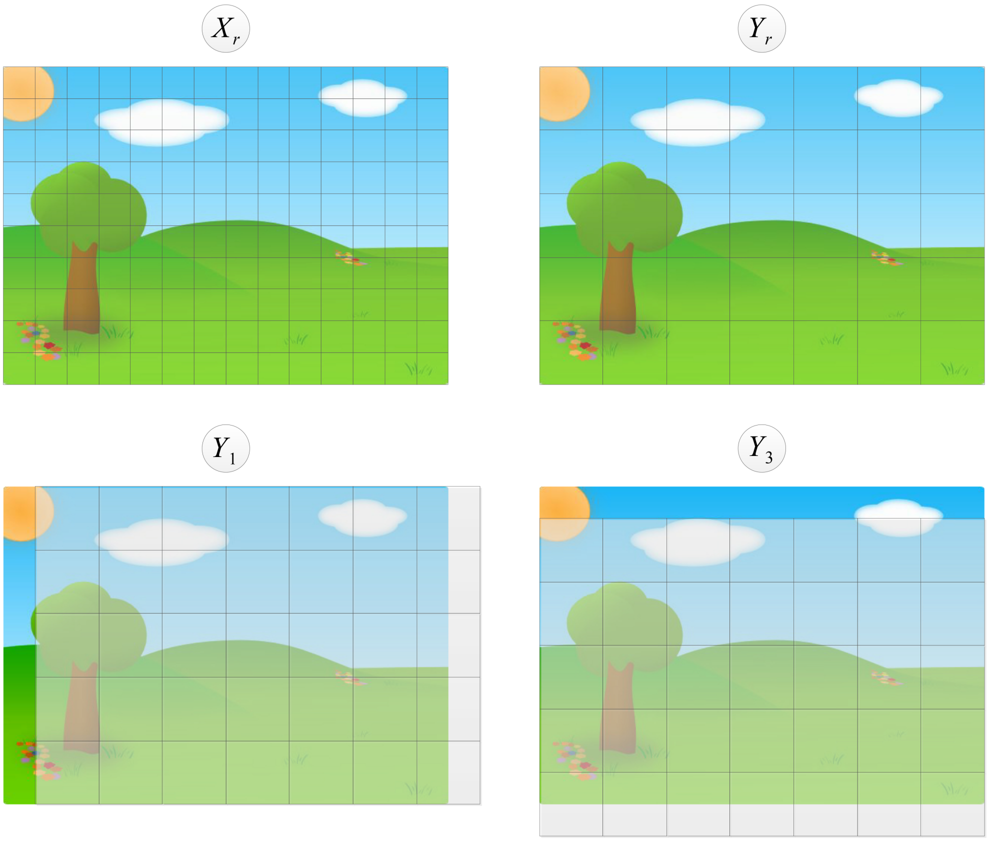

Figure 4.1: Top row, from left to right: (a) The HR reference image. (b) The LR observation of the reference image. Bottom row: (c)-(d) Two more LR observations, that provide additional information by sampling the $\mathbf{x}_r$ pixels using different basis functions. Each square in the grids correspond to the a pixel in the respective image.

the camera model in (4.1) is adopted regardless partially due to its pleasant linear formulation.

A convenient notation for the model is obtained by stacking the LR observations in a vector $\mathbf{y} = [\mathbf{y}_1^T, \ldots, \mathbf{y}_K^T]^T$ and introducing the noise vector $\mathbf{n} = [\mathbf{n}_1^T, \ldots, \mathbf{n}_K^T]^T$, both of size $(nK/L^2) \times 1$, and defining the system matrix

$$\boldsymbol{\mathcal{H}} \triangleq [(\mathbf{RC}\{\mathbf{H}_1\}\mathbf{T}\{\mathbf{D}_{1,r}\})^T, ..., (\mathbf{RC}\{\mathbf{H}_K\}\mathbf{T}\{\mathbf{D}_{K,r}\})^T]^T \tag{4.2}$$

of size $(nK/L^2) \times n$, such that

$$\mathbf{y} = \boldsymbol{\mathcal{H}}\mathbf{z}_r + \mathbf{n}. \tag{4.3}$$

In order to obtain a unique solution to $\mathbf{z}_r$ given $\mathbf{y}$, and for a given downsampling factor L, the number of observed images K should satisfy $K \geq L^2$, otherwise the system of equations is underdetermined.

To show the possible usefulness of SRR, before proceeding to the discussion of (some of) its challenges, an example on $K = 3$ images is presented that compares SRR using the inverse problem formulation with two interpolation approaches. The model in (4.3) is used to generate $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$, where the original $\mathbf{z}_r$ is a pixel valued image normalized to $\{0, \ldots, 1\}$, and the noise $\mathbf{n}$ consists of zero-mean Gaussian components with variance $\sigma_n^2 = 10^{-4}$. In this example, $\mathbf{R}$ performs downsampling by a factor $L = 2$, the $\mathbf{H}_k$ represent a mean operator on an image patch of $L \times L$ pixels (averaging the illuminance, which is an intensity measure) to model a simple, idealistic point spread function (PSF) of the camera sensor. The (global) subpixel shifts used are $(D^x_{1,r} = 0.5, D^y_{1,r} = 0)$ and $(D^x_{3,r} = 0, D^y_{3,r} = 0.5)$. Both the PSF and the subpixel shifts in this example match the illustrations in Figure 4.1 (b)-(d). Perfect knowledge about the operators in $\boldsymbol{\mathcal{H}}$ is assumed in the reconstruction.

The results of the comparative example are shown in Figure 4.2. Figure 4.2 (a) displays the original image $\mathbf{z}_r$. Figure 4.2 (b) shows $\mathbf{y}_r$ upsampled by a factor $L = 2$ using bicubic interpolation. The second upsampling approach, shown in Figure 4.2 (c), is the average of the three upsampled and aligned observations. For that case, zero-order hold (ZOH) interpolation was used for the upsampling of the $\mathbf{y}_k$, as it gave a better MSSIM score than when using bicubic interpolation on the three $\mathbf{y}_k$. Finally, in Figure 4.2 (d), the result from the SRR with the regularized inverse problem formulation

$$\hat{\mathbf{z}}_r = \arg\min_{\mathbf{z}_r} \|\boldsymbol{\mathcal{H}}\mathbf{z}_r - \mathbf{y}\|_2^2 + \lambda \|\mathbf{\Gamma}\mathbf{z}_r\|_2^2 \tag{4.4}$$

is shown. Each color channel in $\mathbf{z}_r$ is treated separately, by solving the minimization problem three times with the corresponding color channel in $\mathbf{y}$. If not for the linear regularization term $\mathbf{\Gamma}\mathbf{z}_r$, of weight $\lambda = 10^{-3}$, the minimization for the given example would not provide a unique solution, due to the nullspace of $\boldsymbol{\mathcal{H}}$. The nullspace exists because only $K = 3 < L^2 = 4$ images are available. The matrix $\mathbf{\Gamma}$, of size $n \times n$ in this example represents 2d convolution on the vectorized image $\mathbf{z}_r$ with a 3×3 Laplacian convolution kernel that penalizes the second derivative in order to enforce a smooth solution. Table 4.1 presents MSSIM image quality scores of the respective greyscale versions of the results from the three approaches.

In the remainder of this chapter, some of the challenges for SRR based on the inverse problem formulation are presented. The next section gives a review of image alignment strategies. Then, the objective function in the SRR minimization of (4.4) is analyzed in more general terms, with respect to the properties of the system matrix $\boldsymbol{\mathcal{H}}$, the choice of norm function for the data residual and the choice of regularization function. Finally, the full SRR algorithm is outlined.

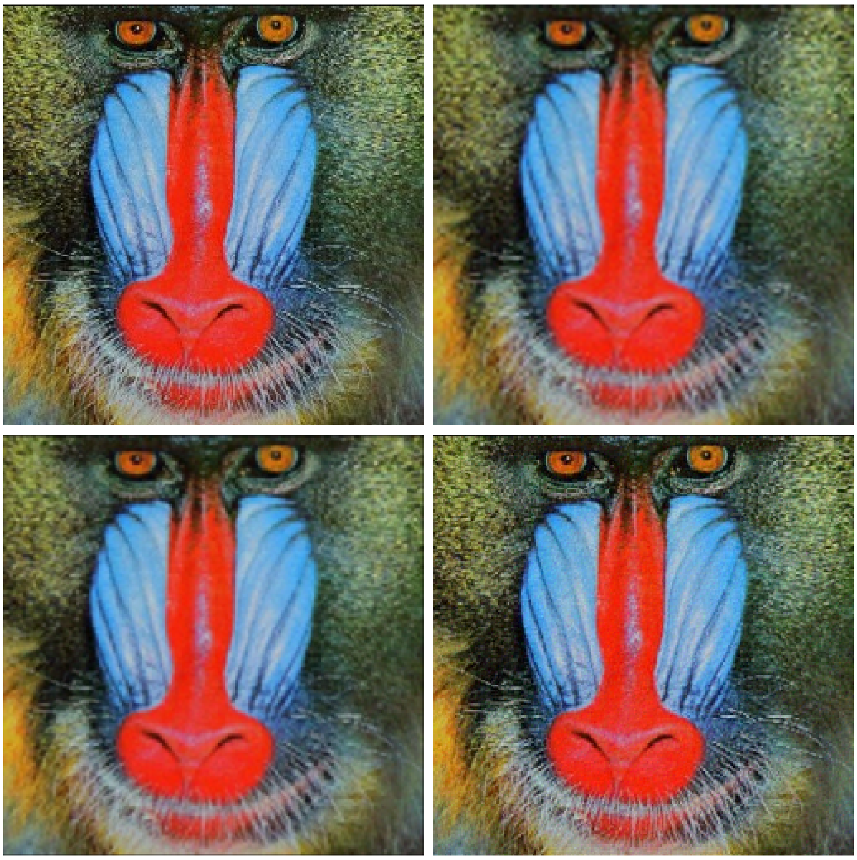

Figure 4.2: Top row, from left to right: (a) Original image $\mathbf{z}_r$. (b) Bicubic interpolation of $\mathbf{y}_2$. Bottom row: (c) Average of the zero-order hold interpolated $\mathbf{y}_k$ images. (d) Result of solving the SRR problem in (4.4).

Method	MSSIM [10]
1. Bicubic Interpolation of $\mathbf{y}_r$	0.7416
2. Upsampled (ZOH interpolation) average	0.8035
3. SRR using the inverse formulation (4.4)	0.9396

Table 4.1: MSSIM results that show the superiority of solving the inverse SRR problem compared to interpolation methods, for the example in Figure 4.2.

4.1 Estimation of displacement fields

If images are taken with, for instance, a handheld camera, as is commonly the case, camera movement will cause the images to be shifted relative to each other. This shift is typically well described by a planar global motion model, for example with affine motion parameters. Furthermore, regardless if the images are taken with a tripod, most scenes contain moving objects that are displaced with relation to the other images in an image sequence. This motion is described as local motion within the image. For reconstruction of an HR image from LR images captured under real-world conditions, the displacement fields $\mathbf{D}_{k,r}$ (contained in $\boldsymbol{\mathcal{H}}$) should therefore be estimated, using a suitable model. For a high quality SRR result, the precision of the displacement estimates is critical. The matter is further complicated by the fact that only downsampled LR images are available for estimating the displacement field, which should be expressed with relation to the HR pixel grid.

To estimate $\mathbf{D}_{k,r}$, several authors of SRR literature assume a global motion model and use low-dimensional parameterizations for the displacements, thus not attempting to model motion within the scene. A global motion model may be a good description for the majority of the image content, the static parts of the scene, which can be useful in itself for some applications. For instance, if the global method is combined with a method which detects where the motion estimation is accurate and forms an image mask containing those areas, the image enhancement method can be applied there, while areas in $\mathbf{z}_r$ for which motion estimation is unreliable can be reconstructed with a simple upsampling method from a single $\mathbf{y}_k$ image. Examples of global alignment strategies include the popular Scale Invariant Feature Transformation (SIFT) method [14], that estimates affine transformation parameters, and frequency domain approaches, for instance as proposed by Vandewalle et al. [15], that estimate planar translation and rotation.

A class of methods that estimate non-parametric displacement fields, in order to model local motion, are termed *optical flow* methods. Seminal papers by Horn-Schunk [16], for global optical flow models, and Lucas-Kanade [17], for local optical flow models, have been the basis for developing optical flow methods for SRR applications. For instance, Baker and Kanade extend the Lucas-Kanade optical flow method to the specific application of SRR. Because $\mathbf{D}_{k,r}$ has two unknown flow components for each pixel, there is not enough information in the images to estimate a non-parametric displacement field without adding additional constraints. Local methods address this by adding local spatial or spatiotemporal constancy constraints, while global methods add a regularization term on the displacement field, to enforce a flow solution that is (piecewise) smooth. Local and global methods have their strengths and weaknesses, when it comes to robustness to noise or to estimate flow fields within homogenous objects. Bruhn and Weickert investigate these properties and propose a combination of the local and global approaches in [18].

Moving objects in the images are referred to as being either rigid or non-rigid (deformable) objects, where a swaying tree or a moving wave are examples of the latter category. These presents larger challenges for flow estimation, and consequently for multi-image reconstruction methods in general. Thus, similarly as for occluded objects that always cause invalid motion estimates, detection of non-rigid motion should be included in an implementations of image alignment methods, and accounted for in subsequent image reconstruction [3].

An alternative, or rather complementary, approach to perform subpixel scale image alignment is that of Blind Super-Resolution (BSR) [19]. The method is similar to Multichannel Blind Deconvolution (MBD), with the extension of downsampling. Both the unknown image and (non-parametric) kernels of a fixed support, one for each input image, are estimated, typically by alternating minimization. Both subproblems are convex in their standard formulations, however the problem is unfortunately non-convex in the kernels, $\{\mathbf{H}_k\}$, and the image jointly. Prior to performing BSR reconstruction, the input images are approximately aligned by a conventional method. Then, the alignment is fine-tuned by the estimation of the blur kernels, that include both the blur kernels as well as small-scale spatial shifts.

4.2 The inverse SRR problem

Earlier in this chapter, the SRR problem was posed in an example as solving the minimization problem (4.4), in order to obtain an estimate of the HR image $\mathbf{z}_r$, given observed image data $\mathbf{y}$. The specific objective function contained in (4.4) is a special case of the more general formulation,

$$\hat{\mathbf{z}}_r = \arg\min_{\mathbf{z}_r} \rho_1(\boldsymbol{\mathcal{H}}\mathbf{z}_r - \mathbf{y}) + \lambda\rho_2(\boldsymbol{\psi}(\mathbf{z}_r)), \tag{4.5}$$

where $\rho_1(\boldsymbol{\mathcal{H}}\mathbf{z}_r - \mathbf{y})$ is the data term, $\rho_2(\boldsymbol{\psi}(\mathbf{z}_r))$ is a regularization term of weight λ, and $\rho_1(\cdot), \rho_2(\cdot)$ are norm-like functions (not necessarily norms in the strict sense). Naturally, the HR image should match the observed data. That is, the residual $\boldsymbol{\mathcal{H}}\mathbf{z}_r - \mathbf{y}$ should be small in $\rho_1(\cdot)$, which should preferably be a function that makes the residual robust, both to noise in the observations $\mathbf{y}$, and to errors in the system matrix $\boldsymbol{\mathcal{H}}$, due to model mismatch or estimation errors in the model parameters. Robust norm functions are discussed in several papers on SRR. The L1-norm has been proposed as an improvement over the L2-norm, for its ability to better handle errors in the model parameters, for instance related to the motion estimation [20]. The *Lorentzian norm*, which acts as the L2-norm for small residuals and as the L1-norm for large residuals, has shown promising results for various noise assumptions [21].

If the minimization of the data term by itself is underdetermined, due to insufficient observations $K < L^2$, there is an infinite number of solutions to the problem, and thus a regularization term, $\rho_2(\boldsymbol{\psi}(\mathbf{z}_r))$, must be added to enforce a solution of desired properties. Even if $\boldsymbol{\mathcal{H}}$ is a full rank matrix, regularization is typically used to improve the otherwise poor condition number of the overall problem, (4.5), at the cost of fidelity of the data term. Note that, if the minimization problem is non-linear, the condition number refers to linearizations of the objective function, that are used in order to solve the problem iteratively.

Generally speaking, a common type of regularization is to penalize the norm of the unknown vector, such that the minimal-norm solution is obtained from the set of solutions. However, it does not make sense in this context to penalize the norm of the image $\mathbf{z}_r$. Instead, because images are known to be relatively smooth (they contain mostly low frequencies), the first or second derivative may be penalized to enforce a smooth solution. Several authors adopt nonlinear regularization functions that are designed not to penalize strong image edges between different image segments, noting that images are somewhat better described as *piecewise* smooth [20, 21]. The use of regularization function can similarly be thought of in a Bayesian framework, where it would represent a prior density on the HR image, and (variational) Bayesian inference could then be used in order to perform the SRR [22, 23].

4.3 The SRR algorithm

Up until now, the two main ingredients of the full SRR algorithm, that is, estimating the displacement fields, as well as the HR image, have been discussed separately. A high level SRR algorithm, in which displacement field- and HR image estimation may be iterated until some stop condition is met, is presented in Table 4.2.

First, the image displacements $\mathbf{D}_{k,r}$ are estimated for a selected motion model. In the initial estimation, this is (typically) done on $\mathbf{y}_k$ images that are upsampled by interpolation to the higher resolution. The current estimate of $\mathbf{z}_r$ may then be used in the subsequent iterations of the displacement field estimation, if the estimation process is iterated. Next, $\mathbf{z}_r$ is reconstructed by solving a minimization problem of the form in (4.5). If a nonlinear objective function is adopted, or if the dimension of the problem is so large, such that an iterative minimization method must be used, the estimate may be initialized using an upsampled version of $\mathbf{y}_r$. The warp-

```
SRR Algorithm
while ~stopflag
    1: $\{\hat{\mathbf{D}}_{k,r}\}$   ← estimate the displacement fields,
    2: $\hat{\mathbf{z}}_r$           ← solve (4.5) to reconstruct the HR image,
    3: stopflag                       ← check if stop condition is met,
end
```

Table 4.2: A high level SRR algorithm consisting of two main estimation steps.

ing of $\mathbf{z}_r$ by $\mathbf{D}_{k,r}$, contained in $\boldsymbol{\mathcal{H}}$, to the spatial location of $\mathbf{y}_k$ includes interpolation onto the HR grid. The impact of this interpolation is not well established theoretically in the SR literature. If BSR is included in the SRR algorithm, an extra step

$$1b: \ \{\hat{\mathbf{H}}_k\} \leftarrow \text{estimate kernels that represent blur and small-scale shifts}$$

is added. In the BSR case, the SRR algorithm should necessarily be iterated in order for the estimates to converge. Choices of a stop condition could be a fixed number of iterations, or a threshold value for some minimum difference on the updated estimates compared to that of the previous iteration. If BSR is not included (which it seldom is), it is quite often the case in the literature that only one iteration is performed, thus estimating displacement fields and the HR image in a sequence.

To finish off this chapter, it is noted that the SRR methods discussed can be straightforwardly applied to color images by solving (4.5) for each color channel. The displacement fields may be estimated jointly for all color channels, benefiting the estimation accuracy.

Chapter 5

Super-Resolution Reconstruction for differently exposed images

Similarly to the two previous chapters, an image set $\{\mathbf{y}_k\}$ is used here to reconstruct a single image, which can benefit from all the information in the multiple observations. In this chapter, the $\mathbf{y}_k$ provide both spatial diversity and differently exposed observations of an underlying HDR scene. Thus, a HR, HDR image $\mathbf{x}_r$ may be reconstructed.

To begin with, corresponding illuminance domain images, $\mathbf{i}_k$, are obtained from the $\mathbf{y}_k$ as in (3.6). Using the model (2.8) for $\mathbf{y}_k$, it follows that

$$\begin{aligned} \mathbf{W}_k\mathbf{i}_k &= \mathbf{W}_k(g(\mathbf{y}_k)/\Delta t_k) = \\ &= \mathbf{W}_k(\mathbf{DC}\{\mathbf{H}_k\}\mathbf{T}\{\mathbf{D}_{k,r}\}\mathbf{x}_r + \mathbf{n}_k), \qquad k = 1, \ldots, K \end{aligned} \tag{5.1}$$

where $\mathbf{W}_k$ is a diagonal weight matrix of size $(n/L^2) \times (n/L^2)$. It gives zero weight to pixels in $\mathbf{i}_k$ that are over- or underexposed, that is, pixels that have an exposure value outside the operational range of $f(\cdot)$ in (2.8). This clipping in the $\mathbf{y}_k$ is not invertible by $g(\cdot)$, and thus the impact of the resulting erroneous information, with respective to the HDR information to be reconstructed in $\mathbf{x}_r$, is excluded by $\mathbf{W}_k$. The introduction of $\mathbf{W}_k$ leads to that the second equality in (6) holds. All the pixel exposures that are in the operational range are given the same weight of one, although downweighting the low and high extremes would likely improve performance in a real case. For mathematical convenience, the impact of the quantization noise $\mathbf{q}_k$ is neglected in the inverse problem formulation (quantization is nevertheless used when generating $\mathbf{y}_k$), as it typically is small in relation to other sources of reconstruction errors, such as the image alignment.

Introducing the notation, $\mathbf{i} = [\mathbf{i}_1^T, ..., \mathbf{i}_K^T]^T$, $\mathbf{v} = [\mathbf{n}_1^T/\Delta t_1, ..., \mathbf{n}_K^T/\Delta t_K]^T$, both of size $(nK/L^2) \times 1$, and $\mathbf{W} = diag(\mathbf{W}_1, ..., \mathbf{W}_K)$, of size $(nK/L^2) \times$

(nK/L^2), a compact equivalent form of (6) is

$$\mathbf{W}\mathbf{i} = \mathbf{W}(\boldsymbol{\mathcal{H}}\mathbf{x}_r + \mathbf{v}), \tag{5.2}$$

where $\boldsymbol{\mathcal{H}}$ is the same system matrix as in Chapter 4. Now, somewhat analogously to the reconstruction of a HR image in Chapter 4, which was achieved by minimizing (4.5), one could solve

$$\hat{\mathbf{x}}_r = \arg\min_{\mathbf{x}_r} \rho_1(\mathbf{W}(\boldsymbol{\mathcal{H}}\mathbf{x}_r - \mathbf{i})) + \lambda\rho_2(\boldsymbol{\psi}(\mathbf{x}_r)) \tag{5.3}$$

in order to obtain a reconstruction of a HR, HDR image, based on the information in $\{\mathbf{y}_k\}$. Similarly to in Chapter 4, the functions $\rho_1(\cdot), \rho_2(\cdot)$ are norm-like functions and $\boldsymbol{\psi}(\cdot)$ is a regularization function. There is a subtle difference, however. Traditionally, SRR is performed on similarly exposed LDR pixel valued images, as was the case in Chapter 4. Whereas the pixel value domain is perceptually uniform, the illuminance domain of $\mathbf{i}$ and $\mathbf{x}_r$ in (5.3) is not. On the contrary, residuals $\rho_1(\mathbf{W}(\boldsymbol{\mathcal{H}}\mathbf{x}_r - \mathbf{i}))$ have a higher perceptual impact for low absolute illuminance levels of $\mathbf{x}_r$.

The published work, so far, on HDR SRR has in common that the reconstruction takes place in the illuminance domain. For example, see the papers by Choi et al., Schubert et al. and Zimmer et al. [11, 24, 25]. An objective function of the form of (5.3) is minimized in order to obtain the resulting HR, HDR image. In the last section of this chapter, the objective function is altered in such a way that the residual vector is expressed in a perceptually uniform domain. First, however, consider the HDR SRR algorithm presented in Table 5.1. It is similar to Table 4.2 with the difference that, unlike the case in Chapter 4, the $\mathbf{y}_k$ here are differently exposed, which adds the step of photometrical alignment. The most common approach, to align the input images both spatially and photometrically, is to first estimate the displacement fields. The displacement field estimates are used to warp the $\mathbf{y}_k$ such that they are aligned spatially. Then, for photometric

```
HDR SRR Algorithm
while ~stopflag
    1:  $\{\hat{\mathbf{D}}_{k,r}\}$   ← estimate the displacement fields,
    2:  $\hat{g}(\cdot)$              ← estimate the mapping from pixel value to exposure,
    3:  $\hat{\mathbf{x}}_r$          ← estimate the HR, HDR image,
    4:  stopflag                       ← check if stop condition is met,
end
```

Table 5.1: A high level SRR algorithm for differently exposed images.

alignment, $g(\cdot)$ is estimated, and used to retrieve the illuminance domain information images $\mathbf{i}_k$. Having aligned the LR, LDR observations both spatially and photometrically, the HR, HDR image is finally estimated.

5.1 Spatial and photometric alignment

To align the $\mathbf{y}_k$ images both spatially and photometrically is somewhat more challenging than the case in which the images are taken with the same exposure settings. Gevrekci and Gunturk discuss different approaches as to go about with the task [26]. The most common approach, which Gevrekci and Gunturk also adopt, is to first align the differently exposed images spatially, and then perform photometric alignment. An alternative approach is to first estimate $g(\cdot)$ based on, for example, a histogram-based approach, followed by spatial alignment of images that have been photometrically aligned.

Various approaches have been proposed to robustly align differently exposed images spatially. For example, the SIFT algorithm ([14]) has been modified specifically for the purpose [27]. There, the SIFT key-points (image features) are obtained from a contrast domain representation of the images. Thus, accurate global motion estimation can be performed to compensate for camera movement. Contrary to the SIFT-based method, which can not handle motion within the scene, Zimmer et al. include in their HDR SR method an optical flow alignment strategy that can handle local motion within the scene [25]. The flow method employed is also their own work, and includes some sophisticated elements. Ultimately, a displacement field between two images is computed by minimizing an energy functional in a gradient image domain, that includes robust penalization functions for outlier handling, due to, for instance, occlusion [28]. At the time of publication, it was reported to be the top ranked method at the Middlebury benchmark[2] for evaluations of optical flow methods, but new methods by other authors now show improved results.

To increase the robustness of image reconstruction methods (that rely on the estimated displacement fields), algorithms that detect troublesome image areas with regard to accurate displacement field estimation should be used. For Example, Hu et al. propose a method for this purpose, that includes a routine for detection of non-rigid motion [29]. These areas then receive special treatment in the image reconstruction methods, typically by the use of some less ambitious reconstruction method.

Once the motion between image frames has been established, image

[2] available at *http://vision.middlebury.edu/flow/eval/results/*

warping can be performed to align the images. Then, a method for photometric alignment, that typically maps pixel valued images to the HDR illuminance domain, is applied [1,30].

5.2 Proposed objective function for SRR of HDR images

In this section, which leads up to the included papers of this thesis, that are summarized in the next chapter, an alternative objective function to that of (5.3) is proposed. The illuminance domain formulation of the minimization problem in (5.3) is thus generalized to

$$\hat{\mathbf{x}}_r = \arg\min_{\mathbf{x}_r} \rho_1(\mathbf{r}_{\text{data}}(\mathbf{x}_r)) + \lambda \rho_2(\boldsymbol{\psi}(\mathbf{x}_r)), \tag{5.4}$$

where $\mathbf{r}_{\text{data}}(\mathbf{x}_r)$ is a residual vector related to the data term, and $\boldsymbol{\psi}(\mathbf{x}_r)$, as before, is a regularization function. If $\rho_1(\cdot)$ and $\rho_2(\cdot)$ are confined to be the L2-norm, (5.4) can be expressed as

$$\hat{\mathbf{x}}_r = \arg\min_{\mathbf{x}_r} \|\mathbf{r}(\mathbf{x}_r)\|_2^2 = \arg\min_{\mathbf{x}_r} \left\| \begin{bmatrix} \mathbf{r}_{\text{data}}(\mathbf{x}_r) \\ \sqrt{\lambda}\boldsymbol{\psi}(\mathbf{x}_r) \end{bmatrix} \right\|_2^2 . \tag{5.5}$$

Unless the data is completely noise-free and the system parameters of $\boldsymbol{\mathcal{H}}$ are estimated perfectly, any choice of objective function will result in some reconstruction errors. Consider the task of minimizing the data term residual $\mathbf{W}(\boldsymbol{\mathcal{H}}\mathbf{x}_r - \mathbf{i})$ for the case where a unique solution exists, that is, $K \geq L^2$ and $rank(\mathbf{W}) > X_1X_2$. If the relative motions between the observed images are small, such that they can be completely included in the support of $\mathbf{H}_k$ (thus, $\mathbf{T}(\mathbf{D}_{k,r})$ is the Identity matrix), and if in addition the elements of $\mathbf{H}_k$ are 0 except for a single element which is 1, denoted *delta sampling* here, then $cond(\boldsymbol{\mathcal{H}}) = 1$, where $cond(\cdot)$ is the condition number of a matrix. The unique solution will differ from $\mathbf{x}_r$ due to noise, but noise will be suppressed rather than amplified.

However, as soon as resolution enhancement is attempted in the reconstruction, which means that $L > 1$, delta sampling (which would still allow $cond(\boldsymbol{\mathcal{H}}) = 1$) is no longer a realistic point spread function. An idealistic PSF, as modelled by $\mathbf{H}_k$, would rather be an $L \times L$ mean filter (at some position within the support of $\mathbf{H}_k$). Along these lines, Baker and Kanade report that, for any PSF that is a reasonable model of the camera sensor, be it an $L \times L$ square PSF or for example a Gaussian PSF of support equal to or greater than $L \times L$, the condition number always grows at least quadratically with L [31]. Furthermore, $cond(\boldsymbol{\mathcal{H}})$ increases linearly with the size of the image vector $\mathbf{x}_r$. Thus, ill-conditioning is a severe problem for SRR. Reconstruction errors are largest near image edges. This is because the noise amplification when solving the inverse problem is large for high frequency components, due to the low-pass characteristics of the forward camera model. Adding more observations (increasing K) somewhat improves the condition number of the problem, but even so, a regularization term is typically required to further improve the conditioning, and thus limit the noise-amplification.

If the general problem (5.4) is taken as the illuminance domain formulation of (5.3), even small reconstruction errors, of the type discussed above, will cause clearly visible edge artifacts in the dim region across image edges. The numerical errors are of the same magnitude on both sides of the edges, but the perceived impact of the reconstruction errors will be much larger at low illuminance regions. To alleviate this issue, the illuminance data, $\mathbf{i}$, is first normalized to [0,1] (the same notation, $\mathbf{i}$, is kept). The data residual is then taken to be $\mathbf{r}_{\text{data}}(\mathbf{x}_r) = \mathbf{W}(\tilde{f}(\boldsymbol{\mathcal{H}}\mathbf{x}_r) - \tilde{f}(\mathbf{i}))$, where $\tilde{f} = (\cdot)^{\gamma_{HDR}}, \gamma_{HDR} < 1$ is a concave, pixelwise function. An interpretation of $\tilde{f}$ is that it is a global tonemapping operator. It maps illuminance values at each pixel to a PU image domain. Note that $\mathbf{W}(\tilde{f}(\boldsymbol{\mathcal{H}}\mathbf{x}_r - \mathbf{i}))$ would not correspond to the perceived size of the error, as the absolute illuminance level is lost when taking the difference.

As a regularization function, $\boldsymbol{\psi}(\mathbf{x}_r) = \boldsymbol{\Gamma}_{\boldsymbol{\mathcal{L}}}\tilde{f}(\mathbf{x}_r)$, where $\boldsymbol{\Gamma}$ is a matrix that represents 2d convolution on a vectorized image with the Laplacian kernel

$$\boldsymbol{\mathcal{L}} = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{5.6}$$

may be used. A smooth solution is thus enforced by penalizing the second derivative. The larger the regularization weight λ, the better the condition number of the overall problem, albeit this comes at the cost of less fidelity of the data term. It is crucial that the regularization term is chosen such that it enforces a structure in $\mathbf{x}_r$ that corresponds to natural image statistics. For this purpose, a piecewise smooth solution is typically preferred, which can be implemented using an *edge-preserving* regularization function. Learning-based methods could also be used to avoid penalizing some common image textures, but these are not considered in this thesis.

If the (norm) function $\rho_2(\cdot)$ is selected appropriately, the penalization of strong image edges can be downgraded. For example, the Lorentzian norm, which acts as the L2-norm for small values and as the L1-norm for large values (as set by a threshold parameter), can be used [21]. The Lorentzian-Laplacian norm then effectively fulfills the similar purpose as the often used, nonlinear, edge-preserving Bilateral Total Variation (BTV) regularization function [20]. Better experimental results than the BTV are reported by [21]. Zimmer et al., in their work on optical flow and HDR SRR methods, use an amended regularization method, based on the work of Sun et al., [32], that only includes smoothing constraint along image edges, and not across image edges [25, 28]. This same function is also used for regularization of displacement fields, where it avoids blurring flow discontinuities that are present around image edges. Nagel and Enkelmann presented the theoretical foundation for the method employed by Zimmer et al., and derive a method to obtain the local image orientations [33].

At this stage, a PU domain has been formulated for the HDR SRR problem. For the remaining discussion in this chapter, consider the L2-norm of the proposed data- and regularization term, contained in the minimization problem

$$\hat{\mathbf{x}}_r = \arg\min_{\mathbf{x}_r} \left\| \begin{bmatrix} \mathbf{W}(\tilde{f}(\boldsymbol{\mathcal{H}}\mathbf{x}_r) - \tilde{f}(\mathbf{i})) \\ \sqrt{\lambda}\boldsymbol{\Gamma}_{\boldsymbol{\mathcal{L}}}\tilde{f}(\mathbf{x}_r) \end{bmatrix} \right\|_2^2 . \tag{5.7}$$

Numerical reconstruction errors are of the same magnitude for any choice of γ_{HDR} in the expression of $\tilde{f}(\cdot)$, but the large perceptual impact in low illuminance regions is avoided thanks to the PU domain which is achieved for a suitable choice of γ_{HDR}. The value which should be used is not entirely clear. As a comparison, the value for γ_{LDR} that is used in gamma correction for common LDR formats is 1/2.2. For the HDR case, a value as low as $\gamma_{HDR} = 1/6$ is necessary to achieve a residual function $\mathbf{r}_{data}(\mathbf{x}_r)$ that is perceptually uniform with respect to the HVS. This value is based on empirical experiments and coincides with the value used in the work by Fairchild and Johnson on the image appearance model *iCAM* [34]. To perform tonemapping with their updated model, iCAM06, an (gamma) exponent of 1/3 is used to encode the illuminance component of a low-pass filtered base layer of the image, followed at a later step by a further exponent in the range of $[0.6, 0.85]$ (depending on the viewing condition) in the r,g,b-space, for an overall gamma (somewhat loosely speaking, since different color spaces are mixed, and additional manipulation is also made) in the range of $[1/5, 1/3.53]$ [8]. The importance of exact perceptual uniformity as well as color fidelity in $\tilde{f}(\cdot)$ is not as crucial as for the TMO that is used to visualize HDR images. Rather, a function that gives a mathematically sound problem formulation should perhaps be seen as satisfactory for the HDR image reconstruction procedure.

Chapter 6

Summary of included papers

This chapter provides a brief summary of the two papers that are included in Part II of the thesis. The papers have been reformatted to comply with the layout of the thesis, but the content has otherwise not been changed. Both papers address SRR of HDR images, and the formulated models incorporate the perception of the HVS.

Paper 1

> T. Bengtsson, 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, *Proc. of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, Kyoto, Japan.

In this paper, HDR SRR is performed, not in the illuminance domain, but in the L*a*b* domain, which is designed to be perceptually uniform, albeit for LDR image data. Thus, the PU approximation is somewhat crude. Although the gamma of 1/3 which is included in the conversion to L*a*b* is not the most suitable for visualization of HDR images, it gives a reasonable image domain for the image reconstruction procedure. Experiments contained in the paper show, by displaying image results as well as MSSIM quality maps, that the artifacts caused by HDR SRR in the illuminance domain are avoided in the PU L*a*b* domain. Furthermore, reconstruction quality, in terms of MSSIM values, is shown for varying numbers of input images, K.

The formulation of the objective function in Paper 1 is, however, based on a heuristic which breaks down for certain scenarios of the system parameters. This is because the actual observations, $\mathbf{y}_k$, are generated according to (1), while the corresponding LR, LDR representation of the unknown

HR, HDR image is modeled according to (4.1) in the objective function of the inverse problem formulation. With this approach, the minimization problem becomes linear with respect to the unknown image in the L*a*b* domain, thus it is a simple formulation, but not stringent. There is an interesting analogy for this heuristic, to the case of traditional SRR for real (non-simulated) data. Should (4.1) be used in the formulation of traditional SRR with similarly exposed images, while (1) proves to be, in fact, closer to the physical reality, the same issue is attained as in the objective function in this paper.

Paper 2

> T. Bengtsson, T. McKelvey and I. Y-H. Gu, Super-Resolution Reconstruction of High Dynamic Range Images in a Perceptually Uniform Domain, *SPIE, Journal of Optical Engineering, Special Issue on High Dynamic Range Imaging* , October 2013.

This paper presents a more thorough body of work, as compared to Paper 1. A nonlinear objective function of the form in (5.4) is proposed. Other choices for $\rho_1(\cdot)$ and $\rho_2(\cdot)$, than the L2 norm in (5.7), are discussed at greater lengths than in Chapter 5. To a certain extent, Paper 2 is a continuation of Chapter 5, that elaborates on the mathematics of the PU formulation of the HDR SRR problem. Three different objective functions are derived and experimentally evaluated. Furthermore, the choices of iterative solution strategies are discussed. Finally, reconstruction results are visualized and presented alongside the objective quality measures PSNR and MSSIM. These results demonstrate the benefit of using a PU domain formulation, such as (5.7), as compared to the illuminance domain formulation (5.3).

Chapter 7

Concluding remarks

Part I of this thesis has presented and discussed methods for reconstructing images of high visual quality, based on a set of lower quality images. In particular, methods for High Dynamic Range image reconstruction as well as Super-Resolution image reconstruction have been treated, first separately, and then jointly in Chapter 5. The respective algorithms have been outlined, rather than covered in detail. Different existing methods have been touched upon, for instance various approaches to image alignment, as well as for the choice of a regularization function in the inverse formulation of the SRR problem. For the case of differently exposed images, it has been stressed that the reconstruction problem should be formulated in an image domain which is perceptually uniform to human perception of brightness.

For successful image reconstruction of a high quality image, it is crucial that the image alignment is performed with high precision. In a real case with complex motion within the scene, this is a challenging task. The closer in time that the multiple observations of the scene are taken, the better it is with regard to estimation performance of the relative motion between the images. HDR image reconstruction, without resolution enhancement, has already been introduced to common users as a setting in consumer cameras. Whether SR will become more widely used, looking ahead, is determined mainly by the success of image alignment algorithms on downsampled images. Furthermore, the SR algorithms to be implemented have to meet critical demands of computational feasibility.

References

[1] P. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 369–378. [Online]. Available: http://dx.doi.org/10.1145/258734.258884

[2] S. Park, M. Park, and M. Kang, "Super-resolution image reconstruction: a technical overview," *Signal Processing Magazine, IEEE*, vol. 20, no. 3, pp. 21 – 36, May 2003.

[3] A. Katsaggelos, R. Molina, and J. Mateos, *Super resolution of images and video*, ser. Synthesis Lectures on Image, Video, and Multimedia Processing. Morgan & Claypool, 2007. [Online]. Available: http://decsai.ugr.es/vip/files/books/SRbook.html

[4] J. Farrell, F. Xiao, and S. Kavusi, "Resolution and light sensitivity tradeoff with pixel size," in *Proceedings of SPIE*, vol. 6069, 2006, pp. 211–218.

[5] M. Angelopoulou, C.-S. Bouganis, P. Cheung, and G. Constantinides, "Robust real-time super-resolution on fpga and an application to video enhancement," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 2, no. 4, pp. 22:1–22:29, 2009. [Online]. Available: http://doi.acm.org/10.1145/1575779.1575782

[6] E. Allen and S. Triantaphillidou, *The Manual of Photography and Digital Imaging.* Taylor & Francis, 2012.

[7] B. Hoefflinger, *High-dynamic-range (HDR) vision : microelectronics, image processing, computer graphics.* Berlin : Springer, 2007.

[8] J. Kuang, G. Johnson, and M. Fairchild, "icam06: A refined image appearance model for hdr image rendering," *J. Vis. Comun. Image Represent.*, vol. 18, pp. 406–414, Oct. 2007.

[9] S. Farsiu, M. Elad, and P. Milanfar, “Multiframe demosaicing and super-resolution of color images,” *Image Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 141 –159, jan. 2006.

[10] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600 –612, april 2004.

[11] J. Choi, M. Park, and M. Kang, “High dynamic range image reconstruction with spatial resolution enhancement,” *Comput. J.*, vol. 52, pp. 114–125, Jan. 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1521033.1521042

[12] M. Cadik, M. Wimmer, L. Neumann, and A. Artusi, “Evaluation of hdr tone mapping methods using essential perceptual attributes,” *Computers Graphics*, vol. 32, no. 3, pp. 330 – 349, 2008.

[13] M. Gevrekci and B. Gunturk, “Image acquisition modeling for super-resolution reconstruction,” in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2. IEEE, 2005, pp. II–1058.

[14] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[15] P. Vandewalle, S. Süsstrunk, and M. Vetterli, “A frequency domain approach to registration of aliased images with application to super-resolution,” *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–14, March 2006. [Online]. Available: http://rr.epfl.ch/3/

[16] B. Horn and B. Schunck, “Determining optical flow,” *Artificial intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[17] B. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.

[18] A. Bruhn, J. Weickert, and C. Schnörr, “Lucas/kanade meets horn/schunck: Combining local and global optic flow methods,” *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005.

[19] F. Sroubek, G. Cristobal, and J. Flusser, “A unified approach to super-resolution and multichannel blind deconvolution,” *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2322 –2332, Sep. 2007.

[20] S. Farsiu, M. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327 –1344, Oct. 2004.

[21] V. Patanavijit and S. Jitapunkul, "A lorentzian stochastic estimation for a robust iterative multiframe super-resolution reconstruction with lorentzian-tikhonov regularization," *EURASIP J. Adv. Signal Processing*, vol. 2007, no. 2, pp. 21–21, 2007. [Online]. Available: http://dx.doi.org/10.1155/2007/34821

[22] L. Pickup, D. Capel, S. Roberts, and A. Zisserman, "Bayesian methods for image super-resolution," *The Computer Journal*, 2007.

[23] S. Babacan, R. Molina, and A. Katsaggelos, "Variational bayesian super resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984 –999, April 2011.

[24] F. Schubert, K. Schertler, and K. Mikolajczyk, "A hands-on approach to high-dynamic-range and superresolution fusion," in *WACV*, 2009, pp. 1–8.

[25] H. Zimmer, A. Bruhn, and J. Weickert, "Freehand hdr imaging of moving scenes with simultaneous resolution enhancement," *Computer Graphics Forum*, vol. 30, no. 2, pp. 405–414, 2011. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8659.2011.01870.x

[26] M. Gevrekci and B. Gunturk, "Superresolution under photometric diversity of images," *EURASIP J. Appl. Signal Process.*, vol. 2007, pp. 205–205, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1155/2007/36076

[27] A. Tomaszewska and R. Mantiuk, "Image registration for multi-exposure high dynamic range image acquisition," in *Proc. International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2007, pp. 49–56.

[28] H. Zimmer, A. Bruhn, J. Weickert, L. Valgaerts, A. Salgado, B. Rosenhahn, and H.-P. Seidel, "Complementary optic flow," in *Energy minimization methods in computer vision and pattern recognition*. Springer, 2009, pp. 207–220.

[29] J. Hu, O. Gallo, and K. Pulli, "Exposure stacks of live scenes with hand-held cameras," *Computer Vision–ECCV 2012*, pp. 499–512, 2012.

[30] M. Robertson, S. Borman, and R. Stevenson, “Dynamic range improvement through multiple exposures,” in *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, vol. 3. IEEE, 1999, pp. 159–163.

[31] S. Baker and T. Kanade, “Limits on super-resolution and how to break them,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 9, pp. 1167–1183, 2002.

[32] D. Sun, S. Roth, J. Lewis, and M. Black, “Learning optical flow,” *Computer Vision–ECCV 2008*, pp. 83–97, 2008.

[33] H. Nagel and W. Enkelmann, “An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 5, pp. 565–593, 1986.

[34] M. Fairchild and G. Johnson, “icam framework for image appearance, differences, and quality,” *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 126–138, 2004.