Thesis for the Degree of Doctor of Philosophy

Image Reconstruction and Optical Flow Estimation on Image Sequences with Differently Exposed Frames

Tomas Bengtsson

CHALMERS

Department of Signals and Systems
Chalmers University of Technology
Göteborg, Sweden 2015

Image Reconstruction and Optical Flow Estimation on Image Sequences with Differently Exposed Frames

Tomas Bengtsson

ISBN 978-91-7597-307-4

Doktorsavhandlingar vid Chalmers tekniska högskola
Serial no. 3988
ISSN 0346-718X

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

Typeset by the author using LaTeX.

Printed by Chalmers Reproservice
Göteborg, Sweden, December 2015

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. For example, the dynamic range of light intensities that the sensor can capture in a given image is much smaller than the dynamic range of common daylight scenes and that of the human visual system. Thus, the scene content in certain regions is not properly captured due to over- or underexposure of the sensor. The dynamic range limitation is addressed by signal processing methods that produce a high dynamic range representation of an original scene by fusing information from a sequence of images. Digital cameras systems, in addition to producing images of high visual quality, are increasingly being used for automatic image analysis tasks, where a computer algorithm analyzes the captured image data and outputs some extracted information. Image analysis results also rely on the use of image data that represents the relevant content of real-world scenes.

This thesis is concerned with the opportunities and challenges of high dynamic range imaging, in the contexts of high quality image reconstruction and motion analysis by optical flow estimation. A method is proposed that produces a high dynamic range image and jointly enhances the spatial image resolution by exploiting the fact that the input image sequence provides complementary spatial information of the scene. Key characteristics of the human visual system are taken into account in the problem formulation in order to improve the perceived image quality. In addition, a method is proposed for optical flow estimation in high dynamic range scenarios, that benefits from using image sequences with differently exposed frames as input. The produced motion information can be used in motion analysis applications, including active safety systems in vehicles.

Keywords: high dynamic range, super-resolution, image reconstruction, optical flow, motion analysis, inverse problem, human visual system, digital camera system, multiple camera settings

Preface

It gives me immense pleasure to present this doctoral thesis. During my years as a doctoral student, I have learned as much about life and about myself as I have about my research topic, and in my view that is saying quite a lot. 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. To turn things on its head, with regard to the content of this thesis, a quote by photographer and storyteller Sebastião Salgado: "It is more important for a photographer to have very good shoes, than to have a very good camera."

This thesis is in partial fulfillment of the requirements for the degree of Doctor of Philosophy (PhD). It has been organized in two parts. In the first part, the research topics are introduced, taking on a broader view as compared to the second part, in which three papers are appended. The work has been supported in part by VINNOVA (the Swedish Governmental Agency for Innovative Systems) within the projects Visual Quality Measures (2009-00071) and Image Fusion for 3D reconstruction of traffic scenes (2013-04702), and by Volvo Car Corporation. Other project partners have been Fraunhofer Chalmers and Epsilon.

Acknowledgements

I have a great many things to be thankful for. Thanks, first of all, to the taxpayers who have funded a large part of my doctoral study period. Thanks also to Volvo Car Corporation, in particular to Konstantin Lindström, for your generous support throughout these years. To my supervisor, Professor Tomas McKelvey, thank you for all the solid advice, our discussions always leave me with a good feeling. To Professor Irene Yu-Hua Gu, thank you for introducing me to the field of research and for your enthusiastic support. To Professor Mats Viberg, thank you for the wine tastings that have captivated my tastebuds on several occasions. To my dear colleagues, thank you dearly! I really appreciate being part of such a group of ambitious, warm-hearted people with mixed backgrounds and specialities.

To Tilak, thank you for being such an unconventional inspirer. To Lars, thanks for sharing your humorous self-taught bitterness. To Johan, thank you for tips and tricks and solid lunch companionship. To Livia, many thanks for taking the lead with arranging social activities for the group. To Abu, that epic weekend with the food rescue party, the after-party, the brunch and Majorna art walk is definitely one to remember, thank you! To Lennart, thank you for the coaching, for innovative teaching methods, and of course for the sourdough. To Oskar, thank you for your insights into life philosophy. To Nina, thank you for staying cool and for your amazing dinner skills. To Eoin, thank you for the ridiculously transdisciplinary PhD pubs, the much needed social occasions for PhD students in Göteborg. Sláinte! To Mahogny coffee bar, and to my friend Dan with whom I share a passion for coffee and Nebbiolo grapes, grazie mille! To Purre and the wonderful staff at Linsen who serve delicious lunch, daily, thanks a bunch for all the tastiness! To my dance partners over the past few years, thanks for mental revitalization, laughs and a strong sense of belonging! Special mention goes out to Jenny, you just continue to amaze me with your great, gentle presence and fun-loving nature, thank you dearly.

To friends, old and new, whose discovery of the joy and thrill of dancing still lies in the future, thank you for all the precious moments that we have shared together. To my Chalmers mates, thank you for all the great parties and for always having something interesting cooking. To friends from global studies, what a fantastic bunch of adventurous and compassionate people you are, thank you for that! To Markus, thank you for the shared exploration into the world of music and comedy. To Johanna, thank you for helping to broaden my view of the world and question things that I had taken for granted. ¡Olé! To Henrik and Fleur, thank you for taking hospitality to such a high level and sharing the art of how to make one feel welcome. To my mother Boel, my father Tage, my brother Martin, and to my extended family in Halmstad, Kristina, Lasse, Lina, Isak, Albin and Felix, thank you dearly. I love you all! My gratitude also extends to friends yet to be met. To sources of inspiration everywhere.

Carpe diem!

Tomas Bengtsson
Göteborg, Sweden, December 2015

List of publications

This thesis is based on the following three appended papers:

Paper 1

> T. Bengtsson, T. McKelvey and I. Y-H. Gu, "Super-Resolution Reconstruction of High Dynamic Range Images in a Perceptually Uniform Domain," in *SPIE Journal of Optical Engineering, Special Issue on High Dynamic Range Imaging* , vol. 52, no. 10, October 2013.

Paper 2

> T. Bengtsson, T. McKelvey and K. Lindström, "Optical Flow Estimation on Image Sequences with Differently Exposed Frames," in *SPIE Journal of Optical Engineering* , vol. 54, no. 9, September 2015.

Paper 3

> T. Bengtsson, T. McKelvey and K. Lindström, "On Robust Optical Flow Estimation on Image Sequences with Differently Exposed Frames using Primal-Dual Optimization," submitted to *International Journal of Computer Vision, Springer*.

The author is the principal contributor to each of the appended papers and has written the papers himself. All authors jointly determined the general direction of the research. The co-authors have assisted with their expert input, dialogue about the structure of the papers and comments on preliminary manuscripts. The problem formulation in Paper 1 was initially developed together with I. Y-H. Gu. Its solution strategy, including the software implementation, was refined under the supervision of T. McKelvey. The author is the main contributor to the problem formulations and solution strategies in Papers 2 and 3.

Other publications

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," in *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, Kyoto, Japan.

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

T. Bengtsson, T. McKelvey and K. Lindström, "Variational Optical Flow Estimation for Images with Spectral and Photometric Sensor Diversity," in *Proceedings of International Conference on Graphic and Image Processing (ICGIP)*, September 2014, Paris, France.

Contents

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 most people, cameras are strongly associated with photography. Cameras are used to take pictures with, of family and friends, vacation travels, beautiful nature and much more. However, aside from producing visually pleasing images, digital camera technology is increasingly being used for automatic image analysis [1, 2]. Generally speaking, image analysis is about extracting meaningful information from images. It has widespread everyday use for tasks such as reading bar codes on the items in the local grocery store. A current, hot application is motion analysis and tracking of vehicles in traffic situations, which provides information to driver assistance and active safety systems [3, 4]. Computerized image analysis is further included in medical imaging systems [5, 6]. Thus, as in the case of medical imaging, the imaging device is not necessarily a conventional camera. It can be any imaging modality that has an array of sensor elements or in other ways can produce images from measuring physical quantities. The list of scientific and industrial areas where digital image analysis is applied can be made long, and include astronomy, geoscience, identification, machine vision, material science, microscopy, remote sensing and robotics [1].

As the thesis title suggests, this work deals with *image reconstruction*, which has to be defined in this context. Image reconstruction methods attempt to retrieve information of the original real-world scene that has been lost in the imaging process. When we, as human beings, observe a real-world scene, an image is formed in our eyes. If the same scene is imaged

by a camera, useful information is lost due to limitations of the camera sensor. In other words, cameras are more restrictive than the human eye in certain crucial aspects. To exemplify, most of us have probably experienced the difficulty of taking good pictures in circumstances where there is bright sunlight in combination with shadow areas, or of indoor environments with a bright window. Such a scenario contains a wide range of light intensities, or in other words, a high dynamic range. Whereas the human eye is capable of seeing indoor and outdoor environments at the same time, an image taken with a camera results in over- or underexposure of certain image regions, due to the insufficient dynamic range of the camera sensor hardware. Currently, so called high dynamic range (HDR) image capture is emerging as a new functionality of consumer camera devices [7]. The aim is to capture a similar range of light intensities to what the human eye is capable of. In order to produce an HDR image, information from multiple differently exposed images is combined [8, 9]. At least two images are used, one taken with a short exposure duration and the other with a long exposure duration. In overcoming the dynamic range limitation, reliable HDR functionality should actually be seen as quite revolutionary for digital camera technology. However, there are challenges, particularly for non-static scenarios. 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 (SR) image reconstruction [10–12], which is used in order to enhance spatial resolution by utilizing multiple images. With the market dominance of high-definition television HDTV (1920×1080) and other high resolution displays, there is a clear application and potential for SR to convert low resolution, low quality video (image sequence) to be pleasantly viewed on these devices. Both techniques, HDR and SR, attempt to combine information from an image sequence 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 increased 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 SR reconstruction 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 acquire as much meaningful data about the original scene as possible, or as necessary with respect to what a human can perceive. 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 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 less mature 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 [13]. SR techniques may also be subject to future use in image coding. For example, it has been suggested for use in image compression [12]. In addition, SR techniques are of interest for displaying video sequences of a given resolution on a device with a higher resolution, as an alternative to traditional, simpler interpolation. In terms of hardware, having a small pixel size comes at the cost of increasing the exposure duration [14], 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 SR to achieve the same total resolution. Custom sensor equipment has been proposed to accommodate this [15].

This thesis further deals with *optical flow estimation* [2,3,16,17], an automatic image analysis task that produces low-level motion estimates that describe apparent motion of each individual pixel element. Optical flow (OF) estimation is automatic in the sense that no user intervention is required to produce the output. The produced motion information can for instance be used to boost performance of image segmentation [18] or to determine motion of specific higher-level objects, such as vehicles in traffic scenarios for application to driver assistance systems [3, 19]. Another application is to spatially align time-series of image data, for instance in medical imaging [20]. Finally, the essential motion information used for image alignment in SR methods is often obtained by OF estimation [21–24]. Thus, the research topics of this thesis clearly overlap. Furthermore, the mathematical approaches used to solve both these problems share many similarities. Optical flow is defined as the pattern of apparent motion that

can be perceived by a given sensor, such as the eye. OF methods typically, including in this thesis, use two or more images from a standard camera as input to estimate flow. Each image pixel is assigned a vector that describes the projected flow onto the 2D image plane of a corresponding real-world point between the time-instances of two captured images. The quality of the input images naturally impacts the result of the estimated flow. Thus, in HDR scenarios, the limited dynamic range of camera sensors can be an issue for the performance of OF methods, just as it is for the case of high quality image reconstruction.

1.1 Aim of the thesis

Two main topics are discussed in this work. They are both separate and at the same time interlinked. The first topic of the thesis addresses 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? This requires highlighting the impact of the human visual system in the problem formulation. The second topic revolves around using high quality image data as input to motion estimation by optical flow techniques. While this topic enters into the first, it is pursued mainly for its own purposes. Specifically, the thesis aims to

I Present a unified survey of image reconstruction methods based on multiple input images, as well as of optical flow estimation for motion analysis applications, and as a part of image reconstruction methods. This provides a broad view of the research areas, in which the contributions of the included papers are placed.

II Propose a method for joint image reconstruction of high resolution, high dynamic range images that is influenced by important characteristics of human visual perception.

III Propose a method for optical flow estimation in HDR scenarios that is based on using image sequences with differently exposed frames.

1.2 Thesis outline

This thesis is divided into two parts. In Part I, the research areas of image reconstruction and OF estimation based on multiple images are discussed, providing a background for the three papers that are appended in Part II of the thesis. Particularly, a selection of work which is relevant to the proposed

methods of joint SR and HDR image reconstruction (Chapter 4, Paper 1) and OF estimation for HDR scenarios (Chapter 6, Papers 2 and 3) is discussed. The chapters on the proposed methods are relatively short, with the details instead available in the respective papers. In Chapter 2, an introduction to digital camera systems is given, including relations to relevant aspects of human visual perception. The mathematical image acquisition model for the camera that is used throughout the thesis is also presented therein. Chapter 3 treats reconstruction of high dynamic range images from differently exposed LDR input images (Section 3.2) as well as reconstruction of images with enhanced spatial resolution by the use of a super-resolution method (Section 3.3). In Chapter 4, SR of HDR image sequences is discussed, and a method is proposed that takes perceptual characteristics of human vision into account in the mathematical formulation. The method thus improves over previous work on joint SR, HDR reconstruction where the problem is formulated in an unsuitable image domain and no regard is taken to human perception. In Chapter 5, image-based motion estimation is discussed, particularly focusing on OF methods. In Chapter 6, the OF estimation problem is extended to image sequences with differently exposed frames, and a solution method is proposed. A summary of the included papers (in Part II) is given in Chapter 7, and concluding remarks are given in Chapter 8.

Chapter 2

Human vision and digital camera systems

Digital camera systems technology is in many aspects designed to mimic the visual system of its developer and user, the human. The use of cameras is primarily to capture still images or video for digital reproduction of real-world scenes. A more recent, alternative application is (automatic) image analysis, which has developed along with increasingly abundant computer processing power. 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 human observer. An overview of a general digital camera system is depicted in Figure 2.1. To the left of the figure

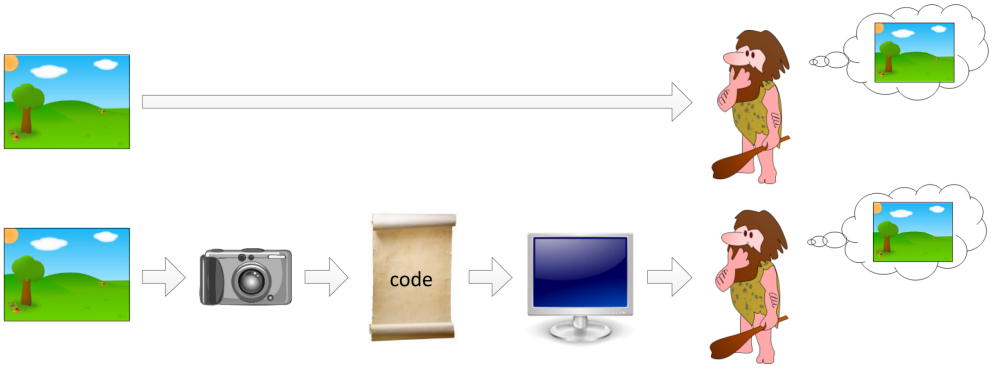

Figure 2.1: A digital camera system. Data of an original scene is captured with a camera, coded with some algorithm and visualized on a display device. The goal is typically that the produced image should be perceptually similar to directly observing the scene.

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 acquisition 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 typical objective of the system in Figure 2.1 is to enable to visualize, on some display device, a high quality image of the original scene. For image analysis applications, however, the objective for the digital camera system differs. The physical data from the camera sensor should then be utilized to perform a certain task, for instance the task of segmenting a specific class of objects. Thus, the code part differs and so does the visualization step, which may consist of highlighting segmented objects. In general, there are numerous automated image analysis applications where the image data should not be visualized at all, but instead be used to trigger some action based on a detected event. For example, motion information analysis of a traffic scenario may be used in a vehicle to issue a warning to the driver or to perform an intervention such as automatic braking.

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. 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, a scalar output value is produced, that may or may not be in the operational range of the sensor [25]. In the case of the camera, these scalar outputs from each pixel element is the raw data, for a given image, that is available for image coding.

An important question that arises related to the digital camera system 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 upcoming sections of this chapter. By using the camera in Figure 2.1 to capture multiple images of the scene, the total information acquired enables to produce and display an image that is free from over- and underexposure, and has a high spatial resolution, properties that are both crucial for a high perceived visual quality. For motion estimation, the more critical of the two discussed camera sensor limitations is the insufficient dynamic range. Thus, using multiple differently exposed images enables to estimate motion in areas that would otherwise be too poorly exposed. Before presenting a mathematical model for the camera, some key concepts in digital image processing and how they relate to the different parts in Figure 2.1 are discussed.

2.1 Key concepts in digital image processing

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 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 quantized, 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,

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

[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.

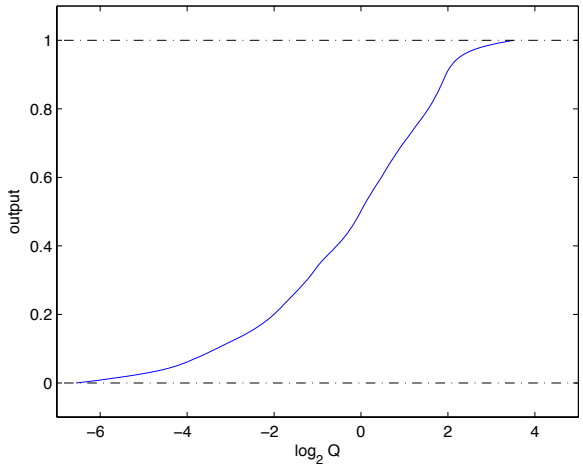

Figure 2.2: The input-output relationship for a signal Q to a sensor.

where $S(\lambda)$ is the incident light spectrum, as a function of the wavelength λ, 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 (SNR) 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 [26]. 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 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 $Stops = \log_2(DR)$, in the unit Exposure Value (EV).

	Dynamic Range	Stops
Original real-world scene	10^5	16.6
Camera (acquisition 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* (see Section 3.2.1) 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 [13]. The contrasts are particularly decreased at distinct image edges, a change that is less noticeable to the HVS than compressing contrasts within textured areas. After tonemapping, the image is coded (and possibly stored) in a general device independent LDR image format, that can be visualized on a display using its LDR intensity interval. The raw HDR image can be retained in a specific HDR format.

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 SNR in the sensor element, the exposure duration Δt needs to be increased [14]. 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 camera sensors

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.

The color filter elements are designed so that they roughly match the average human eye [25]. 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 HVS similarly has three types of cone cells, and like the Bayer filter has a better spatial resolution for brightness than for color perception. The signal at each sensor

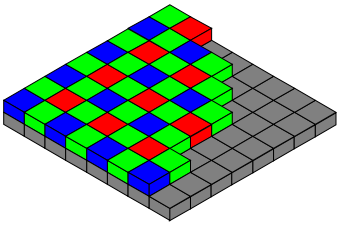

Figure 2.3: The color filter array of the Bayer pattern.

element, that was presented in (2.2), can now be specificed further as

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

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 [27]. The demosaicing could alternatively be formulated within the super-resolution framework, as discussed by Farsiu et al. in [28]. Commonly, however, the SR reconstruction 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 [29]. 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 [29]. 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. [29]. 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, *image formats*, of a relatively low bit depth. This raises the question of how these digital images related to the discussed illuminance images. The answer to that 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 $\lambda \in [380, 700]$ 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 [25]. In combination, the spectral responses of each cone type determine both how colors are perceived as well as perceived brightness. If vision is considered as a greyscale phenomena, 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 [25]. To accommodate this feature, the exposure (2.3) of a camera image (propor-

tional 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 [13].

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 [30, 31], 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.

2.2.1 Perceptual uniformity in HDR imaging

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.

As far as this thesis is concerned, the proposed joint SR and HDR image reconstruction method in Chapter 4 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. 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 [32].

2.3 Camera model

This section presents a mathematical model of a digital camera, which is later used to derive formulations of image reconstruction algorithms, including motion compensation through spatial alignment. The images that the camera delivers are used as input to methods that aim to enhance their dynamic range, spatial resolution, or both. Throughout the thesis we use simplified variants of the camera model, which is formulated to be sufficiently general to encompass all treated problems. For motion estimation between pairs of similarly exposed images, including conventional OF methods, no camera model is typically specified. However, we revisit the camera model and its use for optical flow estimation on image sequences with differently exposed frames in Chapter 6. Consider a sequence of high quality digital images, $\{\mathbf{X}_k\}, k = 1, \ldots, K$, each of size (resolution) $M \times N$, that are in the greyscale illuminance domain (the extension to color images is simply to consider each color channel separately). These images are merely a modeling construction, representing undegraded versions of the actual available images, $\{\mathbf{I}_k\}, k = 1, \ldots, K$, as depicted in Figure 2.4. The $\mathbf{I}_k$ images are observations of the $\mathbf{X}_k$ images, according to the camera model introduced shortly in this section. Both $\mathbf{I}_k$ and $\mathbf{X}_k$ are images, of different quality, of an underlying real-world scene.

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 $\{\mathbf{I}_k\}$. Assuming brightness constancy of scene objects, let the other images be related to the reference according to

$$\mathbf{X}_k(i,j) = \mathbf{X}_r(i + U_{kr}(i,j), j + V_{kr}(i,j)) \tag{2.5}$$

where (i,j) is the pixel location in the image array and $U_{kr}(i,j)$ and $V_{kr}(i,j)$ denote respectively the horizontal and vertical components of the displacement field

$$\mathbf{U}_{kr}(i,j) \triangleq (U_{kr}(i,j), V_{kr}(i,j)), \tag{2.6}$$

that describes the (local) motion of each pixel in image k to its position in the reference image. Notice that (2.5) only holds for pixels (i,j) that are

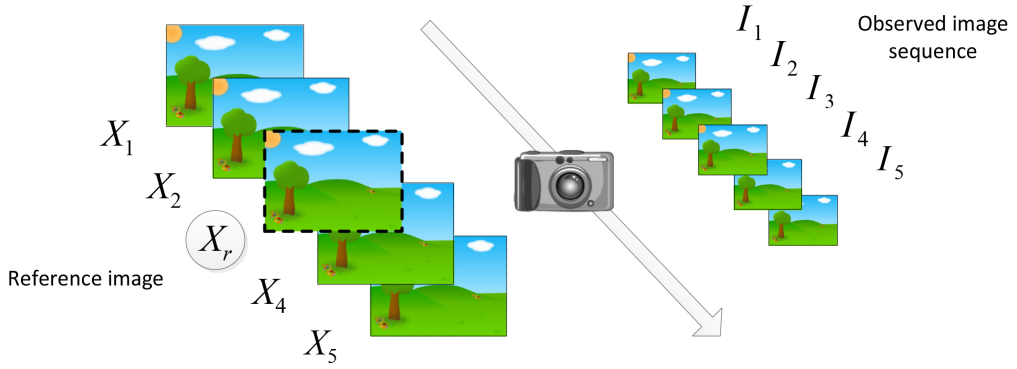

Figure 2.4: An example of $K = 5$ observed images $\mathbf{I}_k$, that could be used to reconstruct a reference image $\mathbf{X}_r$ of, for example, a higher resolution or a higher dynamic range, or to estimate the motion of each pixel in $\mathbf{X}_r$.

non-occluded in $\mathbf{X}_r$, such that a motion vector exists. Since pixel indexes are integer numbers, the displacements, with this formulation, are limited to be integer numbers as well. As an alternative, matrix-vector representation is often used to represent images and image operations. Using $\mathbf{x}_k = vec(\mathbf{X}_k)$, of size $(MN) \times 1 \triangleq n \times 1$, equation (2.5) is re-expressed as

$$\mathbf{x}_k = \mathbf{T}\{\mathbf{U}_{kr}\}\mathbf{x}_r, \tag{2.7}$$

where $\mathbf{T}\{\mathbf{U}_{kr}\}$ is a matrix of size $n \times n$, parameterized by the $M \times N \times 2$ displacements $\mathbf{U}_{kr}$, that relate $\mathbf{x}_k$ and $\mathbf{x}_r$ through a warping operation [33]. The matrix-vector representation is only notation used for analysis, the implementation is realized by image processing operations that for instance allow non-integer pixel displacements in $\mathbf{T}\{\mathbf{U}_{kr}\}$ to be evaluated using interpolation [12, 23].

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

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

For each of the multiple observations, $\mathbf{C}\{\mathbf{H}_k\}$ of size $n \times n$ represents 2-dimensional (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 $\mathbf{D}$, 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 quantization noise is represented by $\mathbf{q}_k$, both are of size $(n/L^2) \times 1$.

The exposure on the camera sensor is $\mathbf{e}_k = \Delta t_k \mathbf{DC}\{\mathbf{H}_k\}\mathbf{x}_k + \mathbf{n}_k$. 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, $I \in \{0, \ldots, 1\}$, in the PU (LDR) image domain of $\mathbf{i}_k$. 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 [25]. 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

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

where the choice of $\gamma_{LDR} = 1/2.2$ is the same exponent as often used for gamma correction applications [34, 35]. 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 4.

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 [25]. 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 $\mathbf{i}_k$, 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 is presently considered as a single channel greyscale image but could contain

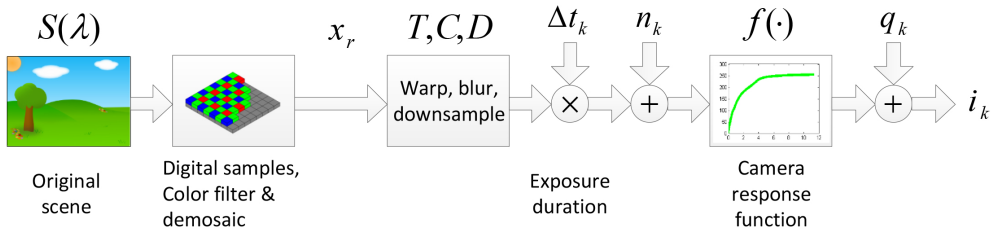

Figure 2.5: The generative camera model.

(demosaiced) r,g,b information, may be warped, blurred and downsampled, as decided by the scenario of interest to model. The exposure image is then mapped by the CRF and finally quantized to produce $\mathbf{i}_k$.

In the following chapters, image sets $\{\mathbf{i}_k\}$ are used to reconstruct images of increased dynamic range (Section 3.2), spatial resolution (Section 3.3) and of both increased dynamic range and spatial resolution jointly (Chapter 4). 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. To conclude this chapter, we comment briefly on the the role of the camera model for automatic image analysis and provide some basics on spatial as well as photometric image alignment that are both re-occurring parts of the presented algorithms throughout the thesis.

2.3.1 Automatic image analysis

As opposed to the case of image visualization, processing or reconstruction, human perception is not necessarily central to image analysis. Thus, how the exposure data is coded by f_{op} in (2.9) is of lesser consequence. Furthermore, downsampling and blurring by $\mathbf{D}$ and $\mathbf{C}\{\mathbf{H}_k\}$ are specifically included in the camera model for image reconstruction purposes and have no use here.

For image analysis purposes, the main point is to give a high weight to physical data with high SNR, excluding human perception. In relation to that, there is a possible, slight shortcoming in the fact that input images to most image analysis methods are taken directly in the pixel value domain without specifying a camera model, when the raw physical data may be more suitable. The issue of saturated image data, naturally, persists in the area of image analysis. If the sensor exposure on a pixel element falls outside of the operational range, $[E_{min}, E_{max}]$, the information associated with it cannot be recovered from that image, which can have a negative impact on the performance of image analysis tasks. In a HDR scenario, to

avoid this from happening, multiple images with varying Δt_k can be taken such that their combined dynamic range exceeds that of the imaged scene.

2.3.2 Spatial image alignment

To describe spatial alignment, consider a pair of two images. Each pixel (i, j) in the first image has a corresponding location in the second image, that differs if the pixel has moved. Spatial alignment is performed by shifting the pixel values of each pixel of the second image back to original location in the first image, an operation called warping. The relation in (2.7) constitutes a *backward* warping $\mathbf{T}\{\mathbf{U}_{kr}\}\mathbf{x}_r$ of the reference image data $\mathbf{x}_r$ to the pixel grid of $\mathbf{x}_k$. The warped image $\mathbf{x}_r^{\text{Warped}} = \mathbf{T}\{\mathbf{U}_{kr}\}\mathbf{x}_r$ is equal to $\mathbf{x}_k$ under the established assumption of brightness constancy. Forward warping, on the contrary, is used for the case where the motion vectors that relate a pair of images are parameterized with respect to the pixel locations of the reference image. In other words, forward warping, $\mathbf{T}\{\mathbf{U}_{rk}\}\mathbf{x}_k$, is based on evaluating $\mathbf{X}_k(i + U_{rk}(i, j), j + V_{rk}(i, j))$ where (i, j) are coordinates of the reference image. For non-integer displacements, warping necessarily includes interpolation to evaluate non-integer pixel locations.

A condition which is important to spatial alignment of image data is *forward-backward consistency* which holds if

$$\mathbf{U}_{rk}(i, j) + \mathbf{U}_{rk}(i + U_{kr}(i, j), V_{kr}(i, j)) = 0. \tag{2.11}$$

In terms of the matrix-vector notation, $\mathbf{T}\{\mathbf{U}_{rk}\}\mathbf{T}\{\mathbf{U}_{kr}\} = \mathbf{Id}$, where $\mathbf{Id}$ is the Identity matrix, holds for consistent points. In practice (for non-static scenarios), there are always points that violate this condition, due to occlusion or moving outside of the imaged area. For *estimation* of displacement fields, a forward-backward consistency check can be useful to detect occluded image areas and discard erroneous estimates at such locations. For image reconstruction purposes, consider the expression $\mathbf{x}_k = \mathbf{T}\{\mathbf{U}_{kr}\}\mathbf{x}_r$, that tells the corresponding location of each point $\mathbf{x}_k(i, j)$ in $\mathbf{x}_r$. A set of points in $\mathbf{x}_k$ are not visible in $\mathbf{x}_r$ due to being occluded there. Observations $\mathbf{x}_k(i, j)$ of such points (i, j) are thus useless in trying to add information to $\mathbf{x}_r$. From the opposite perspective of the reference image, there is a set of points that are visible in $\mathbf{x}_r$ but occluded in $\mathbf{x}_k$. Information from these points would be useful for reconstructing a high quality image $\mathbf{x}_r$ but unfortunately it does not exist in $\mathbf{x}_k$.

Finally, to contrast with spatial image alignment, *image registration* is a widely used concept and a research area in itself for alignment using the best fit of a given global motion model [36, 37].

2.3.3 Photometric image alignment

A set of images are photometrically aligned if the the pixel values of each image $\mathbf{i}_k$ represent intensities on a shared photometric scale. For example, photometric alignment of a set of images taken according to the camera model (2.8) with different exposure durations is achieved by mapping the $\mathbf{i}_k$ 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. If the raw exposure data is available for each image, photometric alignment is achieved directly by dividing with the exposure durations.

Chapter 3

Image reconstruction problems

In this chapter, the separate topics of high dynamic range image reconstruction (Section 3.2) and super-resolution image reconstruction (Section 3.3) are presented. These tasks are then treated jointly in Chapter 4. First, some theoretical concepts that are at the core of image reconstruction methods, as well as of OF methods, are introduced in Section 3.1.

3.1 Robust norms, regularization and learned statistics

Common to all the image reconstruction methods and optical flow methods treated in this thesis is that they solve an *inverse problem*, in other words, a problem where the objective is to estimate a set of parameters that describe the process of producing the observed data. For an inverse problem in linear form, the task is to estimate the variable $\mathbf{x}$, given observed data

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{n}, \tag{3.1}$$

where $\mathbf{A}$ is a system matrix and $\mathbf{n}$ is a noise term. In the general case, $\mathbf{A}$ contains uncertain parameters. In the SR case, the uncertainty in $\mathbf{A}$ is due to incorrectly estimated blur or motion parameters. If $\mathbf{A}$ is deterministic and known, and the elements of $\mathbf{n}$ are independent and identically distributed zero mean Gaussian variables, the estimate $\hat{\mathbf{x}}$ that minimizes the mean squared error $\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_2$, the *maximum likelihood* (ML) estimate in a statistical sense, is

$$\hat{\mathbf{x}} = \mathbf{A}^{\dagger}\mathbf{b}, \tag{3.2}$$

where $\mathbf{A}^{\dagger}$ denotes the pseudo-inverse. Formulated as a minimization problem, the minimizer of $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2$ with respect to $\mathbf{x}$ provides the best estimate under the criteria of minimizing the mean squared error (or equivalently the

PSNR). In the SR literature, alternative norms and norm-like distance functions have been proposed due to the actual noise distribution, and to errors in the system matrix. Farsiu et al. show that, even when the noise term is Gaussian, minimizing the L1 norm of the residual $\mathbf{Ax}-\mathbf{b}$ rather than the L2 norm gives better estimation results due to the uncertainty in the blur and motion parameters of $\mathbf{A}$ [38,39]. The robust Lorentzian norm (not really a norm since it violates the triangle inequality) is adopted in our work on SR reconstruction, as an improvement over using the L1 or L2 norms [40,41].

Super-resolution reconstruction is often imprecisely referred to as an ill-posed problem (in the sense of Hadamard). In more detail, depending on the relation between the downsampling factor and the number of available LR images, estimating the HR image often corresponds to solving an underdetermined system of linear equations, which implies that the problem is ill-posed. If the system matrix of the inverse SR problem is a square or a tall matrix and has full rank, the problem is no longer ill-posed, but it is still often severely ill-conditioned due to the blur and downsampling operators. In the case of an underdetermined problem, regularization of the problem is needed in order for it to have a unique solution. Regularization is achieved by adding additional equations that enforce a certain condition on the solution. Thus, the original objective, to minimize $\|\mathbf{Ax}-\mathbf{b}\|$, is altered to

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{Ax}-\mathbf{b}\|_2^2 + \lambda\rho(\mathbf{x}). \tag{3.3}$$

The new, regularized problem consists of a data term $\|\mathbf{Ax}-\mathbf{b}\|_2^2$ and a regularization term ρ with weight λ. For certain applications, a good choice for the regularization term is $\rho(\mathbf{x}) = \|\mathbf{x}\|$. Then, the resulting estimate

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\| \begin{bmatrix} \mathbf{Ax}-\mathbf{b} \\ \sqrt{\lambda}\mathbf{x} \end{bmatrix} \right\|_2^2 \tag{3.4}$$

is the minimum-norm solution among the set of solutions to the original underdetermined problem. Such a regularization term, however, is not suitable for image reconstruction methods, as the zero solution (or constant solution, if the image data representation is shifted to be symmetric about zero) typically does not represent a reasonable prior for images. On the contrary, regularization terms for image reconstruction are commonly based on the observation that images are typically piecewise smooth, consisting of a set of objects with relatively constant intensities. Due to that, the regularization term should penalize differences in image intensity between nearby pixels on the same imaged object. For SR, being an ill-conditioned problem, a regularization term is typically warranted even if sufficient LR images are available, in order to make the inverse problem more robust to noise (an exception being the case where a very large number of LR images are used).

Regularization is often described as being either deterministic or stochastic [10]. In the Bayesian, stochastic case, the unknown image is distributed according to a prior (representing prior knowledge of $\mathbf{x}$) that roughly models image statistics, also being the result of a trade-off with the need for a practical mathematical expression. Farsiu et al. [38] (a deterministic approach) adopt a regularization term for SR that seeks to minimize the Total Variation (TV) of the image intensities [42–44]. Thus, the TV regularization term penalizes the L1-norm of the image gradient magnitudes. The popular approach of compressed sensing has also been proposed for SR [45]. Several authors formulate their SR methods using a Bayesian framework and discuss reasonable formulations of image priors [24, 46–48]. Statistical justification for using certain image priors is most often based on rather simple observations. More direct attempts to include knowledge of natural image statistics through learning also exist [49, 50]. In the OF literature, notable but rare work to learn statistics for the design of a robust data term norm as well as for regularizing the flow solution is done by Sun et al. [51]. A further discussion on regularizing optical flow is presented in Section 5.2.2.

3.2 HDR image reconstruction

This section discusses how an HDR image can be reconstructed from a set of differently exposed LDR images, $\{\mathbf{i}_k\}$ [7]. The raw sensor exposure of each image is recovered and then merged in the illuminance domain, following spatial alignment of the image set. For HDR image reconstruction, as for the methods presented later, specific assumptions are made with the respect to the operators in the generative camera model (2.8) for $\mathbf{i}_k$. 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), $\mathbf{D} = \mathbf{Id}$. 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 $\mathbf{x}_r$ (the reference image), observed through the differently exposed LDR images

$$\begin{aligned} \mathbf{i}_1 &= f(\Delta t_1 \mathbf{x}_r + \mathbf{n}_1) + \mathbf{q}_1, \\ \widetilde{\mathbf{i}}_2 &= f(\Delta t_2 \mathbf{T}\{\mathbf{U}_{2r}\}\mathbf{x}_r + \widetilde{\mathbf{n}}_2) + \widetilde{\mathbf{q}}_2, \end{aligned} \tag{3.5}$$

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, that should ideally be larger than the dynamic range of the original scene in order to completely avoid over- and underexposure in the reconstructed $\mathbf{x}_r$.

The first step, in order to reconstruct $\mathbf{x}_r$, is to spatially align the observed images to the pixel grid of the reference image. In this case, $\mathbf{i}_1$ shares the same pixel grid locations as $\mathbf{x}_r$, whereas the observations of $\mathbf{x}_r(i,j)$ available through $\widetilde{\mathbf{i}}_2$ need to be aligned to the reference grid by warping to yield

$$\mathbf{i}_2 = \mathbf{T}\{\mathbf{U}_{r2}\}\widetilde{\mathbf{i}}_2. \tag{3.6}$$

If the displacement field between $\mathbf{x}_r$ and $\widetilde{\mathbf{i}}_2$ adheres to a global translational model, that is $\mathbf{U}_{r2}$ is constant for all pixel locations, and the translational shifts are integer numbers of pixels, it follows that, neglecting the image boundaries that are shifted out of the image, $\mathbf{T}\{\mathbf{U}_{r2}\}\mathbf{T}\{\mathbf{U}_{2r}\} = \mathbf{Id}$. Furthermore, because $f(\cdot)$ is a pixelwise function,

$$\begin{aligned} \mathbf{i}_1 &= f(\Delta t_1 \mathbf{x}_r + \mathbf{n}_1) + \mathbf{q}_1, \\ \mathbf{i}_2 &= f(\Delta t_2 \mathbf{x}_r + \mathbf{n}_2) + \mathbf{q}_2. \end{aligned} \tag{3.7}$$

Thus, $\mathbf{i}_1$ and $\mathbf{i}_2$ are two differently exposed, spatially aligned observations of $\mathbf{x}_r$. If, on the other hand, the translational shifts are non-integer numbers, $\mathbf{i}_1$ and $\mathbf{i}_2$ will not be perfectly aligned as suggested by (3.7). This is because, in that case, interpolation is included in $\mathbf{T}$, and thus $\mathbf{T}\{\mathbf{U}_{r2}\}\mathbf{T}\{\mathbf{U}_{2r}\} \neq \mathbf{I}$. Rotation, change of scale or more complex local motion all likewise give rise to interpolation in $\mathbf{T}$. Furthermore, because the warp operator $\mathbf{T}\{\mathbf{U}_{r2}\}$ 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. Such 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. Occluded image regions, however, are not possible to align at all, which may lead to a lack of information in those regions.

In practice, image alignment of differently exposed LDR images is a difficult task. This is due to that motion estimates of high precision are required. For the application to HDR image reconstruction, many approaches to motion compensation exist under the shared name *HDR deghosting*. Tursun et al. propose a taxonomy of HDR motion compensation methods, in which optical flow based methods is one category [9]. New optical flow based methods report increasingly promising results [52, 53]. The earlier method by Zimmer et al. results in severe ghost artifacts for challenging scenarios, according to an evaluation where the patch-based alternative by Sen et al. gives better results [54, 55]. The more recent OF based method by Hafner et al., however, improves over both [53]. In their method, the optical flow and the HDR image are estimated jointly, as opposed to the method by Zimmer et al. where the image alignment is performed as pre-processing.

For image regions with complex motion patterns, the best choice may still be to discard incorrectly aligned data altogether from the reconstruction.

Given a set of K spatially aligned images $\mathbf{i}_k$, for instance $K = 2$ as above, or a larger number, photometric alignment should be performed in order to reconstruct a HDR image $\mathbf{x}_r$. If the CRF f is unknown, it can be estimated from the $\mathbf{i}_k$ images, for example using the non-parametric method of Debevec and Malik [8]. More precisely, the (approximate) inverse CRF g, introduced in Section 2.3, is estimated directly. A set of P pixel positions are selected at random, to provide sample points from each image $\mathbf{i}_k$. If some image areas were not possible to align spatially, these should be avoided in the selection of the sample points. Then, $g(I)$ is estimated for all input values it can take, $I \in \{I_{min}, \ldots, I_{max}\} = \{0, \ldots, 255\}$, jointly with the unknown illuminance values $[\mathbf{x}_r]_i$ of the P sample point pixel positions $i \in \mathbf{p}$, by minimizing

$$\begin{aligned}&\sum_{i\in\mathbf{p}}\sum_{k=1}^{K}\{w([\mathbf{i}_k]_i)[\ln(g([\mathbf{i}_k]_i)) - \ln([\mathbf{x}_r]_i) - \ln(\Delta t_k)]\}^2 + \\ &+ \lambda \sum_{I=I_{min}+1}^{I_{max}-1} w(I)g''(I)^2,\end{aligned} \tag{3.8}$$

where

$$w(I) = \begin{cases} I & , I \leq 127 \\ 255 - I & , I > 127 \end{cases} \tag{3.9}$$

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. More recent research has shown how to improve the weighting function based on more careful modeling of the noise properties of the camera sensor [56]. As seen in (3.8), the minimization is performed in the logarithmic domain, which is much closer to perceptual uniformity than linear illuminance. A smoothness term with weight parameter λ is used to enforce a slowly changing slope of $g(I)$ in the solution. The second derivative can for example be implemented as $g''(I) = g(I-1) - 2g(I) + g(I+1)$. 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 [8] 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 $g(I)$ function is shown. The relation between pixel values $I \in \{0, \ldots, 255\}$ to the exposure $E \in \{g(0), \ldots, g(255\} =$

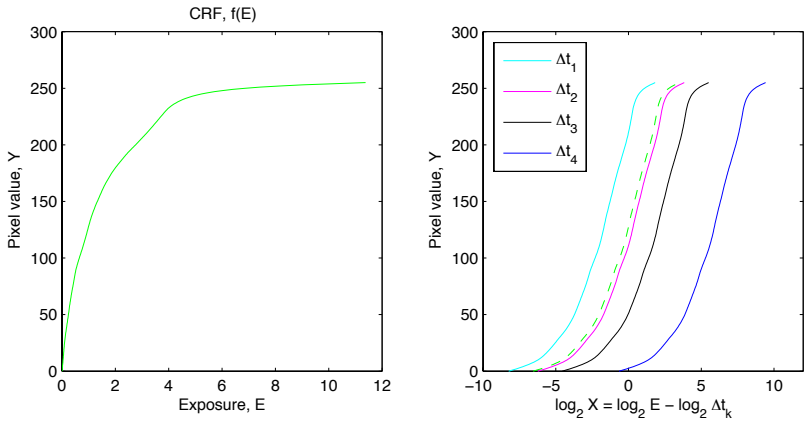

Figure 3.1: From left to right: (a) Results of estimating the inverse CRF from the four LDR images in Figure 3.2, using the method of Debevec and Malik [8]. (b) A plot that shows the combined dynamic range of the LDR observations.

$\{E_{min}, \ldots, E_{max}\} = \{0.0106, \ldots, 11.383\}$ 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 $\mathbf{i}_k$ 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 (2.9), and estimate the polynomial coefficients [57].

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

$$\mathbf{y}_k = g(\mathbf{i}_k)/\Delta t_k, \tag{3.10}$$

such that they become photometrically aligned in a shared domain. The

$\mathbf{y}_k$ images are fused by pixelwise weighted average in the logarithmic (PU) illuminance domain [8, 9]. That is, the pixels values of the reconstructed HDR image $\mathbf{x}_r$ are given as

$$[\mathbf{x}_r]_i = \exp\Big(\frac{\sum_{k=1}^{K} w([\mathbf{i}_k]_i)(\ln g([\mathbf{i}_k]_i) - \ln \Delta t_k)}{\sum_{k=1}^{K} w([\mathbf{i}_k]_i)}\Big). \tag{3.11}$$

Note that a zero weight ($w(I)$ as in (3.9)) 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

$$\mathbf{i}_k = f(\Delta t_k \mathbf{x}_r + \mathbf{n}_k) + \mathbf{q}_k, \tag{3.12}$$

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 (3.11).

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 [13], which is able to better preserve a natural look of colors. The next section gives an overview of existing tonemapping operators.

3.2.1 Tonemapping of HDR images

An image, whether it is generated from an LDR scene or if it contains HDR content, is typically stored 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, captured as a single image with an LDR camera device, 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.

Images of HDR content, such as the $\mathbf{x}_r$ discussed earlier in this chapter, also need to be represented in a standard format that can be interpreted and output on a display device. At this stage, much research is done on how to tonemap HDR images to a PU 8-bit domain, for visualization on LDR devices. Due to the higher dynamic range of the imaged content, however, the 256 quantization levels that 8-bit formats offer is perhaps too restrictive, and higher bit depths may thus be desirable. The simplest tonemapping

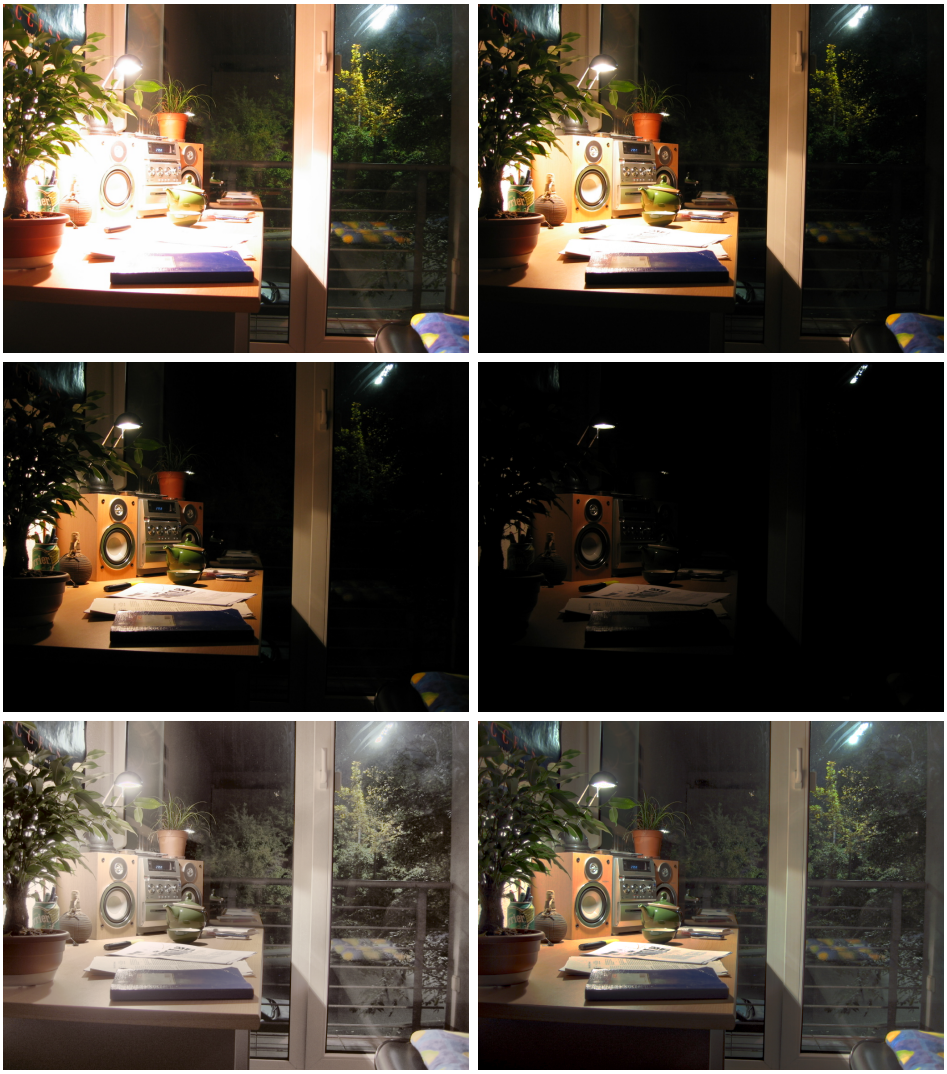

Figure 3.2: Top and middle rows: (a)-(d), Differently exposed LDR input images. Bottom row: Tonemapped HDR result, using the method of Matlab to the left (e), and iCAM06 [13] to the right (f).

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 [13]. 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. [58]. A conclusion that is drawn by the authors from their survey is that, while local processing or multi-resolution 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.

3.3 SR image reconstruction

In this section, a set $\{\mathbf{i}_k\}$ of low resolution images are used to reconstruct, by a super-resolution method, an image $\mathbf{x}_r$ of a higher resolution [10–12]. All LR images are assumed to be taken with the same exposure duration. Thus, the reconstructed image will be a low dynamic range image similar to the input images. What makes SR reconstruction work is that each of the $\mathbf{i}_k$ images provides new information of $\mathbf{x}_r$, as depicted in the example of Figure 3.3. The images, for this to be the case, 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 $\mathbf{i}_r$ in Figure 3.3 (b) provides information about $\mathbf{x}_r$ in Figure 3.3 (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 $\mathbf{x}_r$ (which has a resolution $L = 2$ times higher per dimension than $\mathbf{i}_r$). Taking into consideration more observations, such as those in Figure 3.3 (c) and (d), additional information about $\mathbf{x}_r$ is given.

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

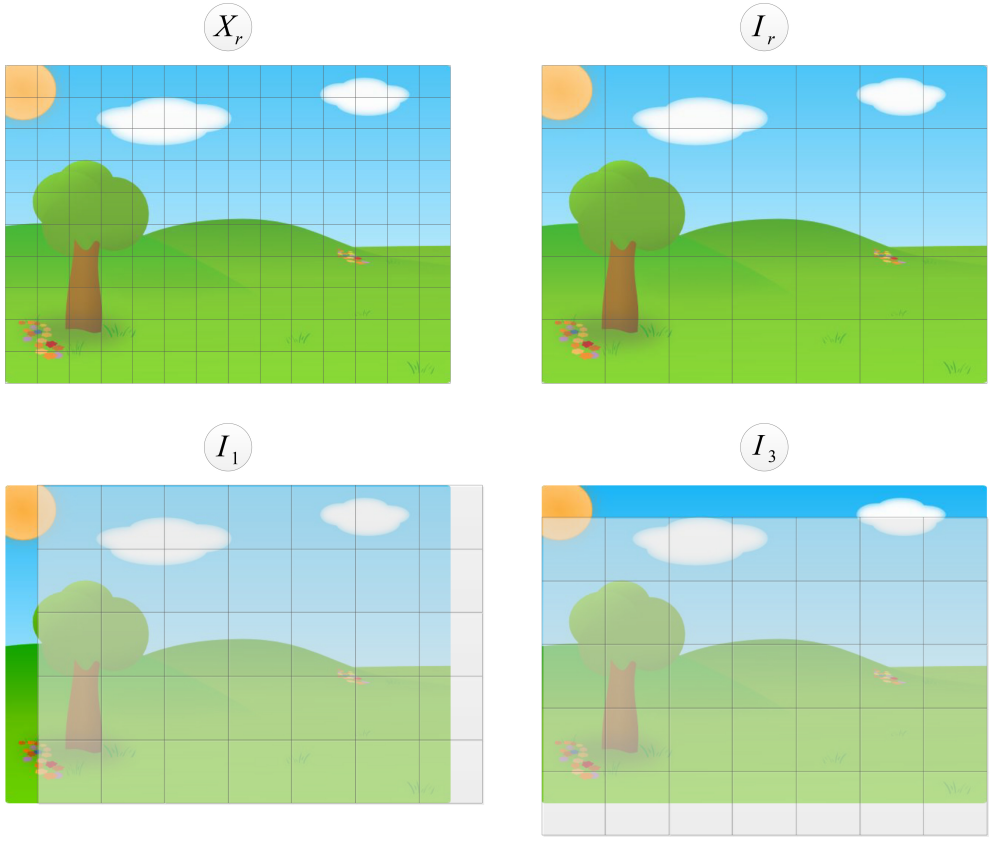

Figure 3.3: Top row, from left to right: (a) The HR reference image. (b) The LR observation of the reference image. Bottom row: (c)-(d) Two more LR observations, that provide additional information by sampling the $\mathbf{x}_r$ pixels using different basis functions. Each square in the grids correspond to the a pixel in the respective image.

and alternatively use the camera model

$$\begin{aligned} \mathbf{i}_k &= \mathbf{DC}\{\mathbf{H}_k\}\mathbf{T}\{\mathbf{U}_{kr}\}f(\Delta t\ \mathbf{x}_r) + \mathbf{n}_k = \\ &= \mathbf{DC}\{\mathbf{H}_k\}\mathbf{T}\{\mathbf{U}_{kr}\}\mathbf{z}_r + \mathbf{n}_k, \qquad k = 1, \ldots, K \end{aligned} \tag{3.13}$$

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 SR reconstruction of LDR images. It was first when differently exposed images were considered that authors on the topic of SR reconstruction 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 [59]. It is possible that the camera model in (3.13) 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 $\mathbf{i} = [\mathbf{i}_1^T, \ldots, \mathbf{i}_K^T]^T$ 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

$$\boldsymbol{\mathcal{H}} \triangleq [(\mathbf{DC}\{\mathbf{H}_1\}\mathbf{T}\{\mathbf{U}_{1r}\})^T, ..., (\mathbf{DC}\{\mathbf{H}_K\}\mathbf{T}\{\mathbf{U}_{Kr}\})^T]^T \tag{3.14}$$

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

$$\mathbf{i} = \boldsymbol{\mathcal{H}}\mathbf{z}_r + \mathbf{n}. \tag{3.15}$$

In order to obtain a unique solution to $\mathbf{z}_r$ given $\mathbf{i}$, 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 SR methods, before proceeding to the discussion of (some of) its challenges, an example with $K = 3$ images is presented that compares SR reconstruction using the inverse problem formulation with two interpolation approaches. The model in (3.15) is used to generate $\{\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3\}$, 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, $\mathbf{D}$ 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 $(D_{1,r}^x = 0.5, D_{1,r}^y = 0)$ and $(D_{3,r}^x = 0, D_{3,r}^y = 0.5)$. Both the PSF and the subpixel shifts in this example match the illustrations in Figure 3.3 (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 3.4. Figure 3.4 (a) displays the original image $\mathbf{z}_r$. Figure 3.4 (b) shows $\mathbf{i}_r$ upsampled by a factor $L = 2$ using bicubic interpolation. The second upsampling approach, shown in Figure 3.4 (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 $\mathbf{i}_k$, as it gave a better MSSIM score compared to using bicubic interpolation on the three $\mathbf{i}_k$. Finally, in Figure 3.4 (d), the result from the SR reconstruction with the regularized inverse problem formulation

$$\hat{\mathbf{z}}_r = \arg\min_{\mathbf{z}_r} \|\boldsymbol{\mathcal{H}}\mathbf{z}_r - \mathbf{i}\|_2^2 + \lambda\|\boldsymbol{\Gamma}\mathbf{z}_r\|_2^2 \tag{3.16}$$

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 $\mathbf{i}$. If not for the linear regularization term $\boldsymbol{\Gamma}\mathbf{z}_r$, of weight $\lambda = 10^{-3}$, the

Figure 3.4: Top row, from left to right: (a) Original image $\mathbf{z}_r$. (b) Bicubic interpolation of $\mathbf{i}_2$. Bottom row: (c) Average of the zero-order hold interpolated $\mathbf{i}_k$ images. (d) Result of solving the SR problem in (3.16).

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 $\boldsymbol{\Gamma}$, 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 order derivative to enforce a smooth solution. Table 3.1 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 SR reconstruction based on the inverse problem formulation are presented. The next section recaps image alignment strategies used for SR reconstruction. Then, the objective function in the SR minimization of (3.16) 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 SR algorithm is outlined.

Method	MSSIM [29]
1. Bicubic Interpolation of $\mathbf{i}_r$	0.7416
2. Upsampled (ZOH interpolation) average	0.8035
3. SR using the inverse formulation (3.16)	0.9396

Table 3.1: MSSIM results that show the superiority of solving the inverse SR problem compared to interpolation methods, for the example in Figure 3.4.

3.3.1 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. These shifts are typically well described by a planar global motion model, for example with affine motion parameters. Furthermore, regardless of 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 $\mathbf{U}_{kr}$ (contained in $\boldsymbol{\mathcal{H}}$) should therefore be estimated, using a suitable model. For a high quality SR 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 $\mathbf{U}_{kr}$, several authors of SR literature assume a global motion model and use low-dimensional parameterizations for the displacements, thus not attempting to model motion within the scene [36, 37]. 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 $\mathbf{i}_k$ image. Additional examples of global alignment strategies include using the Scale Invariant Feature Transformation (SIFT) method [60] along with Random Sample Consensus (RANSAC) [61] to estimate global affine transformation parameters [62], and frequency domain approaches, for instance as proposed by Vandewalle et al. [63], that estimate planar translation and rotation. A class of methods that estimate non-parametric displacement fields, in order to model local motion, are optical flow methods. Seminal papers by Horn-

Schunk [16], for global OF models, and Lucas-Kanade [64], for local OF models, have been the basis for developing OF methods for SR applications. For instance, Baker and Kanade extend the Lucas-Kanade OF method to the specific application of SR reconstruction [21].

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 [12].

An alternative, or rather complementary, approach to perform subpixel scale image alignment is that of blind super-resolution (BSR) [65]. The method is similar to Multichannel Blind Deconvolution (MBD), with the extension of downsampling [66, 67]. Both the unknown image and (non-parametric) kernels of a fixed support, one for each related image $\mathbf{i}_k$, 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.

3.3.2 The inverse SR problem

Earlier in this chapter, the SR problem was posed in an example as solving the minimization problem (3.16), in order to obtain an estimate of the HR image $\mathbf{z}_r$, given observed image data $\mathbf{i}$. The specific objective function contained in (3.16) is a special case of the more general formulation,

$$\hat{\mathbf{z}}_r = \arg\min_{\mathbf{z}_r} \rho_1(\boldsymbol{\mathcal{H}}\mathbf{z}_r - \mathbf{i}) + \lambda\rho_2(\boldsymbol{\psi}(\mathbf{z}_r)), \tag{3.17}$$

where $\rho_1(\boldsymbol{\mathcal{H}}\mathbf{z}_r - \mathbf{i})$ is the data term, $\rho_2(\boldsymbol{\psi}(\mathbf{z}_r))$ is a regularization term of weight λ, $\boldsymbol{\psi}$ is a function of $\mathbf{z}_r$ 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 $\boldsymbol{\mathcal{H}}\mathbf{z}_r - \mathbf{i}$ should be small in $\rho_1(\cdot)$, which should preferably be a function that makes the residual robust, both to noise in the observations $\mathbf{i}$, and to errors in the system matrix $\boldsymbol{\mathcal{H}}$ [39], due to model mismatch or estimation errors in the model parameters. Robust norm functions are discussed in several papers on SR.

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 [38]. 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 [41].

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 $K \geq L^2$ and $\boldsymbol{\mathcal{H}}$ is a full rank matrix, regularization is typically used to improve the otherwise poor condition number of the overall problem, (3.17), 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, because images are known to be relatively smooth (they contain mostly low frequencies), a better alternative is to penalize the first or second derivative of $\mathbf{z}_r$ to enforce a smooth solution. Several authors adopt nonlinear regularization functions that are designed not to over-penalize strong image edges between different image segments, noting that images are somewhat better described as *piecewise* smooth [38,41]. The use of a 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 SR reconstruction [48, 68].

3.3.3 The SR algorithm

Up until now, the two main ingredients of the full SR algorithm, that is, estimating the displacement fields, as well as the HR image, have been discussed separately. A high level SR algorithm, in which displacement field- and HR image estimation may be iterated until some stop condition is met, is presented in Algorithm 3.1.

First, the (HR) image displacements $\mathbf{U}_{kr}$ are estimated for a selected motion model. In the initial estimation, this is done either on $\mathbf{i}_k$ images that are upsampled by interpolation to the higher resolution or on $\mathbf{i}_k$ themselves, followed by upsampling the estimated displacement field. 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 (3.17). If

SR algorithm		
while $\sim$*stopflag*		
1:	$\{\hat{\mathbf{U}}_{kr}\}$	$\leftarrow$ estimate the displacement fields,
2:	$\hat{\mathbf{z}}_r$	$\leftarrow$ solve (3.17) to reconstruct the HR image,
3:	*stopflag*	$\leftarrow$ check if stop condition is met,
end		

Algorithm 3.1: A high level SR algorithm consisting of two main estimation steps.

a nonlinear objective function is adopted, or if the dimension of the problem is very large, such that an iterative minimization method must be used, the estimate may be initialized using an upsampled version of $\mathbf{i}_r$. A gradient update step of (3.17) is

$$\hat{\mathbf{z}}_r^{(n+1)} = \hat{\mathbf{z}}_r^{(n)} - \beta\Big(\nabla\rho_1(\hat{\mathbf{z}}_r^{(n)}) + \lambda\nabla\rho_2(\hat{\mathbf{z}}_r^{(n)})\Big), \tag{3.18}$$

where β is the step length,

$$\nabla\rho_1(\hat{\mathbf{z}}_r^{(n)}) = \mathbf{T}\{\mathbf{U}_{kr}\}^T\mathbf{D}^T\mathbf{C}^T\{\mathbf{H}_k\}\rho_1'\Big(\mathbf{DC}\{\mathbf{H}_k\}\mathbf{T}\{\mathbf{U}_{kr}\}\hat{\mathbf{z}}_r^n - \mathbf{i}_k\Big), \tag{3.19}$$

and the regularization term is left unspecified for the purpose here, which is to analyze the transposed operators in (3.19). The linear operator $\mathbf{C}^T\{\mathbf{H}_k\} = \mathbf{C}^T\{\widetilde{\mathbf{H}}_k\}$ is implemented by 2D convolution with the (real-valued) kernel $\mathbf{H}_k$ flipped in horizontal and vertical directions about its origin, such that $\widetilde{\mathbf{H}}_k(i,j) = \mathbf{H}_k(-i,-j)$. Next, $\mathbf{D}^T$ implements upsampling from the LR pixel grid to corresponding HR locations (without interpolation). Finally, whereas $\mathbf{T}\{\mathbf{U}_{rk}\}$ in the camera model and the objective function (3.17) performs backward warping of the HR image $\mathbf{z}_r$ to the locations of each $\mathbf{i}_k$, $\mathbf{T}\{\mathbf{U}_{kr}\}^T = \mathbf{T}\{\mathbf{U}_{rk}\}$ in (3.19) denotes forward warping [23]. The forward displacement field $\mathbf{U}_{rk}$ cannot directly be obtained from $\mathbf{U}_{rk}$, since multiple pixels in $\mathbf{X}_r$ may map to the same pixel (i,j) in $\mathbf{X}_k$. Thus, it needs to be estimated separately.

If BSR is included in the SR algorithm of Algorithm 3.1, 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 SR algorithm should necessarily be iterated (at least steps 1*b* and 2) 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

not rare that only one iteration is performed, thus estimating displacement fields and the HR image in a sequence. However, recent methods typically perform multiple iterations to refine both estimates [23, 24, 46]. While the presented SR algorithm is fairly general, there are methods that fall outside of it. Notably, an extension of non-local means denoising performs SR reconstruction without explicit motion estimation [69].

Chapter 4

Joint SR and HDR image reconstruction

Similarly to the case of separate HDR or SR image reconstruction, an image set $\{\mathbf{i}_k\}$ is used here to reconstruct a single image, which can benefit from all the information in the multiple observations. In this chapter, the $\mathbf{i}_k$ 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, $\mathbf{y}_k$, are obtained from the $\mathbf{i}_k$ as in (3.10). Using the model (2.8) for $\mathbf{i}_k$, it follows that

$$\begin{aligned} \mathbf{W}_k\mathbf{y}_k &= \mathbf{W}_k(g(\mathbf{i}_k)/\Delta t_k) = \\ &= \mathbf{W}_k(\mathbf{DC}\{\mathbf{H}_k\}\mathbf{T}\{\mathbf{U}_{kr}\}\mathbf{x}_r + \mathbf{n}_k), \quad k = 1, \ldots, K \end{aligned} \tag{4.1}$$

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 $\mathbf{y}_k$ 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 $\mathbf{i}_k$ 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 (4.1) 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, $\mathbf{y} = [\mathbf{y}_1^T, ..., \mathbf{y}_K^T]^T$, $\mathbf{v} = [\mathbf{n}_1^T/\Delta t_1, ..., \mathbf{n}_K^T/\Delta t_K]^T$, both of size $(nK/L^2) \times 1$, and $\mathbf{W} = diag(\mathbf{W}_1, ..., \mathbf{W}_K)$, of size $(nK/L^2) \times$

HDR SR algorithm

while $\sim$*stopflag*

1: $\{\hat{\mathbf{U}}_{kr}\}$ $\leftarrow$ estimate the displacement fields,
2: $\hat{g}(\cdot)$ $\leftarrow$ estimate the mapping from pixel value to exposure,
3: $\hat{\mathbf{x}}_r$ $\leftarrow$ estimate the HR, HDR image,
4: *stopflag* $\leftarrow$ check if stop condition is met,

end

Algorithm 4.1: A high level SR algorithm for differently exposed images.

(nK/L^2), a compact equivalent form of (4.1) is

$$\mathbf{W}\mathbf{y} = \mathbf{W}(\boldsymbol{\mathcal{H}}\mathbf{x}_r + \mathbf{v}), \tag{4.2}$$

where $\boldsymbol{\mathcal{H}}$ is the same system matrix as in Section 3.3. Now, somewhat analogously to the reconstruction of a HR image in Section 3.3, which was achieved by minimizing (3.17), one could solve

$$\hat{\mathbf{x}}_r = \arg\min_{\mathbf{x}_r} \rho_1(\mathbf{W}(\boldsymbol{\mathcal{H}}\mathbf{x}_r - \mathbf{y})) + \lambda\rho_2(\boldsymbol{\psi}(\mathbf{x}_r)) \tag{4.3}$$

in order to obtain a reconstruction of a HR, HDR image, based on the information in $\{\mathbf{i}_k\}$. 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, SR reconstruction is performed on similarly exposed LDR pixel valued images, as was the case in Section 3.3. Whereas the pixel value domain is perceptually uniform, the illuminance domain of $\mathbf{y}$ and $\mathbf{x}_r$ in (4.3) is not. On the contrary, residuals $\rho_1(\mathbf{W}(\boldsymbol{\mathcal{H}}\mathbf{x}_r - \mathbf{y}))$ have a higher perceptual impact for low absolute illuminance levels of $\mathbf{x}_r$.

The published work, so far, on HDR SR 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. [52, 57, 70]. An objective function of the form of (4.3) is minimized in order to obtain the resulting HR, HDR image. Recently, Traonmilin and Aguerrebere presented a method where that specifies a weight matrix $\mathbf{W}$ that depends on the pixel value intensities of the unknown HR, HDR image, but the impact of human perception is not mentioned [71]. In the last section of this chapter, we alter the objective function (4.3) in such a way that the residual vector is expressed in a perceptually uniform domain. First, however, consider the HDR SR algorithm presented in Algorithm 4.1. It is similar to Algorithm 3.1 with the difference that, unlike the case in Chapter 4, the $\mathbf{i}_k$ here are differently exposed, which adds the step of photometrical alignment. The most

common approach, if neither the displacement fields or the inverse CRF are known, is to first estimate the displacement fields. The displacement field estimates are used to warp the $\mathbf{i}_k$ such that they are aligned spatially, at least for certain image areas. Then, for photometric alignment, $g(\cdot)$ is estimated based on spatially aligned image areas, and used to retrieve the illuminance domain information images $\mathbf{y}_k$. Having aligned the LR, LDR observations both spatially and photometrically, the HR, HDR image is finally estimated.

4.1 Spatial and photometric alignment of differently exposed images

This section discusses the case where neither the displacement field or the CRF is known, but mainly the case where the CRF is known or where raw illuminance information of the images is retained. The image displacements are generally not known and thus need to be estimated. To align the $\mathbf{i}_k$ images both spatially and photometrically, including estimating the CRF, is more challenging than performing either spatial or photometric alignment alone. Gevrekci and Gunturk discuss different approaches as to go about with the task [72]. The most common approach, which Gevrekci and Gunturk also adopt, is to first align the differently exposed images spatially, and then to estimate the (inverse) CRF to 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 [73].

If the CRF is known or if the raw exposure information of each image is available, the spatial and photometric alignment for HDR SR reduces to a separate step of spatial alignment of differently exposed images (with different regions being saturated in each image), similar to the case for HDR image reconstruction in Section 3.2. However, as is always the case for SR methods, the motion estimation is extra challenging due to the need to perform it on downsampled images, low resolution images compared to the HR pixel grid. The proposed methods for joint SR and HDR image reconstruction to date use optical flow based methods for spatial alignment. These OF based algorithms should detect troublesome image areas with regard to accurate displacement field estimation, just like other types of HDR deghosting techniques often do [9]. Thus, to increase the robustness of image reconstruction methods that rely on the estimated displacement fields, Hu et al. propose a method that includes a routine for detection of non-rigid motion [74]. These areas then receive special treatment in the

image reconstruction methods, typically by the use of some less ambitious reconstruction method.

An example of a method that uses OF as part of HDR SR reconstruction is that of Zimmer et al. 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 [75]. At the time of publication, it was reported to be the top ranked method at the Middlebury benchmark [76, 77] for evaluations of optical flow methods, but new methods by other authors now show improved results. Still, the precision of the OF method is not sufficient to avoid severe reconstruction artifacts [55]. Hafner et al. provide an improved method that estimates optical flow and a HDR image jointly, however without attempting any resolution enhancement. Once the optical flow (or other motion information) between image frames has been established, image warping can be performed to align the images. Then, a method for photometric alignment, that typically maps pixel valued images (or raw exposure data) to the HDR illuminance domain, is applied [8, 78].

4.2 Proposed objective function for SR reconstruction of HDR images

In this section, which leads up to the appended Paper 1 of this thesis, which is summarized in Chapter 7, an alternative objective function to that of (4.3) is proposed. The illuminance domain formulation of the minimization problem in (4.3) is thus generalized to

$$\hat{\mathbf{x}}_r = \arg\min_{\mathbf{x}_r} \rho_1(\mathbf{r}_{\text{data}}(\mathbf{x}_r)) + \lambda\rho_2(\boldsymbol{\psi}(\mathbf{x}_r)), \tag{4.4}$$

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, (4.4) can be expressed as

$$\hat{\mathbf{x}}_r = \arg\min_{\mathbf{x}_r} \|\mathbf{r}(\mathbf{x}_r)\|_2^2 = \arg\min_{\mathbf{x}_r} \left\| \begin{bmatrix} \mathbf{r}_{\text{data}}(\mathbf{x}_r) \\ \sqrt{\lambda}\boldsymbol{\psi}(\mathbf{x}_r) \end{bmatrix} \right\|_2^2. \tag{4.5}$$

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 $\mathbf{W}(\boldsymbol{\mathcal{H}}\mathbf{x}_r - \mathbf{y})$ for the case where a unique solution exists, that is, $K \geq L^2$

and $rank(\mathbf{W}) > MN$. 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, $\mathbf{T}(\mathbf{U}_{kr})$ 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 [79]. 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 SR reconstruction. 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 (4.4) is taken as the illuminance domain formulation of (4.3), 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, $\mathbf{y}$, is first normalized to [0,1] (the same notation, $\mathbf{y}$, is kept). The data residual is then taken to be $\mathbf{r}_{\text{data}}(\mathbf{x}_r) = \mathbf{W}(\widetilde{f}(\boldsymbol{\mathcal{H}}\mathbf{x}_r) - \widetilde{f}(\mathbf{y}))$, where $\widetilde{f} = (\cdot)^{\gamma_{HDR}}, \gamma_{HDR} < 1$ is a concave, pixelwise function. An interpretation of $\widetilde{f}$ is that it is a global tonemapping operator. It maps illuminance values at each pixel to a PU image domain. Note that $\mathbf{W}(\widetilde{f}(\boldsymbol{\mathcal{H}}\mathbf{x}_r - \mathbf{y}))$ 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, $\boldsymbol{\psi}(\mathbf{x}_r) = \boldsymbol{\Gamma}_{\boldsymbol{\mathcal{L}}}\widetilde{f}(\mathbf{x}_r)$, where $\boldsymbol{\Gamma}$ is a matrix that represents 2D convolution on a vectorized image with the Laplacian kernel

$$\boldsymbol{\mathcal{L}} = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{4.6}$$

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 outside of the main scope of 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 [41]. The Lorentzian-Laplacian norm then effectively fulfills the similar purpose as the often used, nonlinear, edge-preserving Bilateral Total Variation (BTV) regularization function [38]. Better experimental results than the BTV are reported by [41]. Zimmer et al., in their work on optical flow and HDR SR methods, use an amended regularization method, based on the work of Sun et al. [51] that only includes smoothing constraint along image edges, and not across image edges [52, 75]. 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 [80].

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

$$\hat{\mathbf{x}}_r = \arg\min_{\mathbf{x}_r} \left\| \begin{bmatrix} \mathbf{W}(\widetilde{f}(\boldsymbol{\mathcal{H}}\mathbf{x}_r) - \widetilde{f}(\mathbf{y})) \\ \sqrt{\lambda}\boldsymbol{\Gamma}_{\boldsymbol{\mathcal{L}}}\widetilde{f}(\mathbf{x}_r) \end{bmatrix} \right\|_2^2. \tag{4.7}$$

Numerical reconstruction errors are of the same magnitude for any choice of γ_{HDR} in the expression of $\widetilde{f}(\cdot)$, 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* [81]. 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 $[1/5, 1/3.53]$ [13]. The importance of exact perceptual uniformity as well as color fidelity in $\widetilde{f}(\cdot)$ 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 5

Image-based motion estimation

In previous chapters, compensating for motion of pixels has been discussed as a part of image reconstruction methods. In this and the following chapter, the focus is shifted to estimating such motion. Particularly, the focus is on optical flow estimation, a technique that produce a dense motion field estimate, that can describe local motion information of rigid as well as non-rigid (deformable) objects. Some alternative motion estimation techniques are presented to provide a context for Optical Flow (OF) methods. First, however, we discuss the concepts of motion and optical flow in formal terms.

Motion, generally speaking, refers to the 3-dimensional (3D) motion of real-world points. A *motion field* is a dense representation of 3D motion, or of its projection on the 2D image plane. The motion field can be defined as the time-derivatives of the image location of each point. To estimate it from an image sequence, a pair of images are typically used to produce a time-integrated motion field estimate between the two images, using discrete derivative approximations. Optical flow is a concept that comes from ecological psychology, particularly the study of the perception of the visual world by animals (and humans) [82]. It is the pattern of apparent motion that can be observed by, for instance, the human eye, or in our context, the camera sensor. The rotation of a perfect sphere with constant reflectance across its surface is an example of motion that is not visible. It moves, but it is not a case of apparent motion to the human observer or to the camera sensor. Apparent motion, or flow, for a camera connected to an image analysis algorithm, is loosely speaking the displacements of brightness patterns, which may be due to illumination changes, including shadows, and not due to actual motion. A human, on the contrary, is able to infer from the structure of the scene that such effects are in fact not due to motion.

Global motion estimation is the task of estimating motion parameters that are shared for the whole image region [37]. Such a parametric model can for instance describe translational, rotational or affine motion. The

dimension of the parameter representation is typically very low (relative to the number of pixels in the image). Global motion of an image relative to a reference image can be estimated by maximizing a similarity function, for instance based on a correlation or mutual information measure. Another motion estimation approach is to detect a sparse set of points per image that have distinctive features, and to match each feature descriptor in the first image with a similar feature descriptor from the second image [60, 83, 84]. The information of matched points can then either be interpreted as separate motion corresponding to those particular points (or to the image segments that they belong to), or all the feature matches can be used jointly to reach a consensus of global motion parameters [61, 62].

5.1 Dense motion estimation

The objective for dense motion estimation methods is to estimate the motion corresponding to each pixel location in an image. The term dense is introduced to contrast versus sparse techniques, where only a subset of the image region, typically a set of scattered points, are considered. OF methods, by convention, provide dense flow field estimates of the apparent motion seen in the 2D image plane. In Figure 5.1, a pair of overlaid images from the Middlebury [76] *Grove2* sequence are shown together with matches of sparsely extracted SURF image features [83], as well as a flow field estimated by an OF method (the one in Section 5.2.4). The maximum ground truth 2D displacement has a magnitude of slightly above 5 pixels, thus the overlaid images coincide to the degree that it looks like one image. The color encoded flow field describes estimated pixel displacements between the two input images. The ground truth flow is essentially the projection of global 3D motion that results from panning of the camera. No local object motion is present. Upon careful inspection of Figure 5.1 (a), two clear mismatches of the respective SURF features are seen to exhibit far too large displacements to fit the global motion of the scene. One of these mismatches is shown in the zoomed in area at the bottom left corner.

Certain 3D flow cannot be estimated from a set of 2D images taken from the same view (static camera position) due to geometrical reasons, including occlusion. The 3D counterpart of optical flow methods, scene flow methods, typically use a stereo camera setup to circumvent this issue and provides 3D flow vectors as well as their corresponding 3D real-world coordinates based on pairs of stereo images [3]. Alternative methods that use a single, moving camera to capture different views of the scene also exist for 3D flow estimation [85]. Such methods go by the name structure from

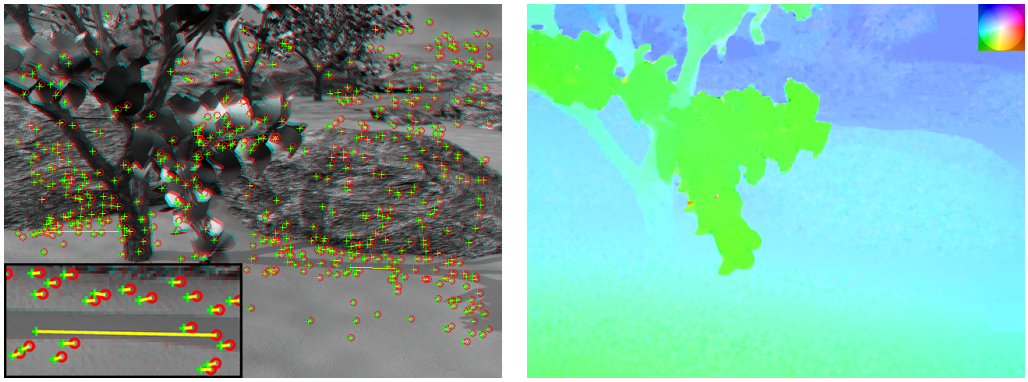

Figure 5.1: From left to right: (a) Two overlaid images. The respective matched subsets of their extracted SURF features are denoted by circles and plus-signs. Matches are connected by a line. A zoomed-in area is shown in the bottom left corner. (b) Visualization of an estimated flow field. Each pixel has corresponding 2D flow vector with direction and magnitude encoded according to the color chart shown in the upper right corner.

motion. Scene flow methods are far less common than OF methods, but are gaining in interest, particularly for research related to autonomous driving. The intended application of the corresponding motion analysis system determines whether a stereo camera hardware setup should be prioritized over a single camera. If depth information is required, an alternative is to use optical flow data in combination with a separate depth (range) sensor. The mathematical formulation of scene flow estimation is very similar to using an OF formulation but involving a larger number of images. Typically, separate OF estimation cost terms and stereo disparity estimation (which is essentially OF estimation for 1-dimensional flow) cost terms are coupled in a joint optimization, from which 3D flow is obtained.

5.1.1 Performance assessment

There are a number of popular performance benchmarks for evaluating dense motion estimation techniques, particularly optical flow methods (2D motion estimation). The benchmark websites provide rankings of OF methods on different error metrics for estimated flow fields relative to the ground truth flow $\mathbf{u}_{rk}^{gt}$, averaged over a set of test image sequences. The most widely used error metric is the average endpoint error (AEPE) of the estimated flow vectors $\hat{\mathbf{u}}_{rk}$. The pointwise endpoint error is EPE $= \|\hat{\mathbf{u}}_{rk} - \mathbf{u}_{rk}^{gt}\|$. Aside from the test sequences, the respective benchmark datasets also contain equally challenging training sequences with public ground truth data avail-

able. These can be used to tune or train OF algorithms. The introduction of new datasets has historically contributed significantly to showing where previous methods fail. Access to publicly available image sequences with ground truth flow information is clearly valuable, since it allows researchers to compare the performance of their methods to existing algorithms.

A classical benchmark whose dataset is still widely used to date is Middlebury [76, 77]. Only a few years after its introduction in 2007, OF methods performed very well on its image sequences. Thus, to help distinguish between the performances of state-of-the-art methods, a new benchmark, MPI Sintel [86, 87], introduced image sequences of animated but realistic-looking data that contain several challenging aspects that are missing in the Middlebury sequences. These challenges are natural illumination changes, motion blur as well as large displacements (flow magnitudes) that in addition to being a challenge in their own right lead to larger occluded regions. Particularly, large displacements of small objects are hard to estimate and often lead convergence of the numerical solution to poor local minima when solving the OF estimation problem. This is because small objects vanish at coarse resolution levels where their large displacement need to be estimated. Furthermore, motion that includes severe object deformations is also hard to estimate and tends to be oversmoothed in estimated flow fields. Another recent benchmark is KITTI [88, 89], which specializes on road sequences captured from inside a moving vehicle. It specifically targets and facilitates research towards autonomous driving. Because the sequences are taken outdoor, they contain many natural illumination changes. In its original release, ground truth flow was provided only for static scenes, that include large displacements but not of small or deformable objects. A new release was made in 2015 where accurate OF (as well as scene flow) ground truth has been established for sequences that contain dynamic scenes with moving traffic. Moving objects (vehicles) are described by a rigid body motion model [90].

Comparing the results of OF methods on MPI Sintel and KITTI, it is clear that each dataset rewards certain characteristics of the OF method. In other words, how well an OF method works is highly data dependent. This points to the inherent tradeoff between generalization versus specialization of an OF algorithm. Should it work sufficiently well for all possible scenarios, or specialize on the demands of a certain end application. Prominent researchers and practitioners within the OF field of research aim to raise a more active discussion on performance analysis in a broader sense than merely evaluating flow error metrics [91]. They observe that new OF methods are published at a rate so high that the achievements of existing OF methods are not consolidated to a satisfactory degree. A more systematic

way to analyze performance, dealing with whether motion can be estimated all, confidence measures associated with the flow estimates (at each point) and robustness with respect to model violations, would help an engineer to choose among the numerous published methods for a specific application. Furthermore, publication of reference implementations of OF methods is encouraged, to allow the research community to dissect and compare all details of an OF implementation. Due to their data dependent nature, it is hard to give conclusive answers as to what elements of OF methods are best. An example of a systematic performance evaluation is performed by Vogel et al. [92], but they only use image sequences from Middlebury and KITTI, which again makes it questionable how well their systematic results generalize. Sun et al. investigate the performance of numerical implementation choices, regarding for example the interpolation method and discrete derivative approximations, and unveil some "secrets" of optical flow [93].

5.2 Variational optical flow estimation

Optical flow estimation methods are often formulated using variational mathematics [94–96]. Observed images are thus seen as time-discrete samples of an underlying image function $I(\mathbf{x}, t)$, where $\mathbf{x} = (x, y)$ are the spatial coordinates and t denotes time. A flow field estimate is obtained by minimizing a cost functional expression with respect to the flow that relate the points of the input images. Whereas the word *domain* is also used to describe the representation of image intensities (for instance linear or logarithmic domain), its proper usage in the variational context is to describe the extent of the continuous-valued spatial locations $\mathbf{x}$, that take values on the domain $\Omega \subset \mathbb{R}^2$. The variational formulation fits nicely with the subpixel nature of flow data and the fact that optical flow is defined in continuous time. This, together with the solid mathematical foundations of the calculus of variations are the reasons for its widespread use in the OF research community. In numerical implementations, continuous derivatives are replaced by discrete approximations and non-integer points $\mathbf{x}$ are evaluated by interpolation.

Connected to the variational formulation, some notation needs to be reintroduced. Images are denoted by $I_k(\mathbf{x}) = I(\mathbf{x}, t_k)$. They are considered to be greyscale images $I_k : \Omega \to \mathbb{R}$ unless otherwise stated. A displacement field parameterized with respect to the points of a reference image I_r is denoted by $\mathbf{u}_{rk} = (u_{rk}(\mathbf{x}), v_{rk}(\mathbf{x}))$. It describes the time-integrated flow between the sampling time instances. The corresponding locations of a point $\mathbf{x}$ in $\mathbf{I}_k$ is thus $\mathbf{x} + \mathbf{u}_{rk}$. If $t_k = t_r + 1$, as assumed from here on, a flow field estimate is given directly as the estimated displacement field, otherwise it is

given by scaling with $1/(t_k - t_r)$. The ground truth flow $\mathbf{u}_{rk}$ is the actual 2D projected motion, discussed in Section 5.1. However, to estimate it based on image data, we have to rely on some assumption to relate points in different images. The most common and traditional approach is to assume that any given point has a constant brightness along its motion trajectory, and use this as a basis to formulate OF estimation mathematically. This assumption is called the Brightness Constancy Assumption (BCA), or the optical flow constraint (the word constraint is often used in a sloppy manner and should not be confused with constraints in optimization). Under the BCA, the best estimate $\hat{\mathbf{u}}_{rk}$ satisfies $I_k(\mathbf{x} + \hat{\mathbf{u}}_{rk}) - I_r(\mathbf{x}) = 0$. In practice, the OF data cost term is based on relaxing the left hand side and minimizing

$$I_k(\mathbf{x} + \mathbf{u}_{rk}) - I_r(\mathbf{x}), \tag{5.1}$$

with respect to $\mathbf{u}_{rk}$ in a suitable norm or norm-like distance function as part of a total cost functional. Because the flow field contains two unknowns, horizontal and vertical flow, for each point, an additional criteria is needed for the problem to have a unique solution for arbitrary images. There are two primary approaches used to handle this, both based on the assumption that nearby points move in a similar manner. In other words, the flow field is assumed to be smooth. One approach is to assume that every pixel in a local neighbourhood share the same flow and include that into the OF formulation [64]. The second approach, which is the more common for optical flow methods nowadays, is to add a regularization term that enforces a spatial condition on the flow solution globally. Its formulation is based on statistics of flow patterns or, rather, at least on the observations that flow is typically piecewise smooth, due to that images consist of a collection of objects where the points of each object exhibit similar flow [16]. Naturally, mixtures of these so called local and global approaches also exist [97].

Focusing on the case of a pixelwise data cost and a global regularization term, the overall OF cost functional to be minimized is

$$E_{\text{total}}(\mathbf{u}_{rk}) = E_D + \alpha_S E_S = \int_\Omega F_D + \alpha_S F_S \; d\mathbf{x}, \tag{5.2}$$

where α_S is a regularization weight and F_D, F_S are corresponding pointwise terms to E_D, E_S. There are many design choices for the data term as well as for the spatial regularization term, as discussed in turn in the next two sections. These two are the fundamental terms in variational OF estimation. An additional term is sometimes included to enforce temporal coherency of flow estimates. Furthermore, modeling of natural illumination changes between input images can be attempted. These additions are discussed in Chapter 6. Recent methods often include a term that penalizes the

flow solution against sparse flow estimates derived from pre-matched image features. The purpose of such a term is to aid the iterative minimization method in avoiding poor local minima, as discusses in Section 5.2.3.

5.2.1 OF data cost term

The OF data cost term typically has the form

$$F_D = \Psi_D(T(I_k(\mathbf{x} + \mathbf{u}_{rk})) - T(I_r(\mathbf{x}))), \tag{5.3}$$

where Ψ_D is a distance function, for instance a norm, and T denotes an arbitrary operation of the pixel values. The standard BCA based data term is obtained when T is the Identity. More general expressions $\Psi_D(I_r, I_k, \mathbf{u}_{rk})$ exist, sometimes with an extra variable $l(\mathbf{x})$ to model illumination changes, and including additive mixtures of different data costs, for instance combining the BCA with a second data term based on the gradient constancy assumption (GCA) [33]. Many proposed formulations were recently evaluated by Vogel et al. on the Middlebury and KITTI datasets, including the conventional data term based on the BCA, a version of the BCA that uses structure-enhanced input images obtained via a structure-texture decomposition of the original images, a normalized cross-correlation measure, a mutual information data term and finally one based on census transformed images [92]. On the Middlebury sequences, the simple approaches, such as the BCA based data term, perform equally well or better than the other. However, on the KITTI sequences, the patch-based census transform and a proposed convex approximation of it significantly outperform the simple pixelwise data terms [98–100]. Nevertheless, many highly ranked OF methods use variants of the BCA and GCA in combination with a normalisation scheme that prevents undesirable overweighting of the data term at large image gradient locations [101, 102]. A recent publication reports improved results on the KITTI benchmark when adding an illumination offset variable and learning characteristic brightness transfer functions that vary across different imaged objects [103]. The authors argue that discarding information about absolute contrast magnitudes, which the census transform does, degrades performance. Among the proposed data terms in the literature, only a few methods explicitly deal with occlusion in its formulation, notable exceptions include [104–107].

5.2.2 Spatial regularization for optical flow

The OF spatial regularization term is designed based on the statistical properties of flow fields, and aim to penalize deviations from those properties.

The expression

$$F_S = F_S(\nabla I_r, \nabla u_{rk}, \nabla v_{rk}), \tag{5.4}$$

where the arguments are the first order derivatives of the reference image and the flow covers many of the proposed formulations. Weickert and Schnörr discuss regularizers on that form using the categorization of flow-driven versus image-driven, and isotropic versus anisotropic regularization terms [108]. Image-driven formulations are based on the observations that the flow edges in an image tend to be a subset of the image edges, and thus less penalty is given to flow edges if their location coincide with image edges. Flow-driven formulations, on the other hand, does not use any information of image data. Particularly image-driven regularizers are either isotropic or anisotropic. In the isotropic case, a function of the gradient magnitudes of the reference image simply acts as a spatially dependent weight term c in $F_S = c(\|\nabla I_r\|)\Psi_S(\nabla u_{rk}, \nabla v_{rk})$, where $\Psi_S(\nabla u_{rk}, \nabla v_{rk})$ is a purely flow-driven expression. Anisotropic image-driven regularizers in addition use edge orientation information to smooth the flow solution along but not across image edges [51, 101].

Other variants of spatial regularization terms include higher order derivatives of the flow. The most popular such method is the total generalized variation (TGV) penalty term, particularly its second order variant (the first order is the traditional TV penalty) that enforces the solution to be piecewise affine [109]. In general, a TGV^k regularizer of order k assigns zero cost to polynomials of order $k-1$. TGV^2 outperforms first order methods on the KITTI benchmark [88, 89], due to its dataset consisting primarily of flat surfaces such as roads and houses whose flow solution is affine. There is also a non-local TGV regularizer that achieves improved boundary localization and robustness to scale changes between images by incorporating larger neighborhoods into the regularizer [110]. Finally, it should be mentioned that a number of recent top performing OF methods use a rather different technique to enforce spatial coherency of flow estimates. They start from a sparse set of matched points and obtain dense flow field estimates by performing edge-preserving interpolation [102, 111]. These sparse-to-dense methods nevertheless include a (post-processing) refinement step where a conventional OF estimation method is used to obtain subpixel precision.

5.2.3 Coarse-to-fine iterative minimization

Variational OF that consists of minimizing a cost functional of the form (5.2) is typically solved using a coarse-to-fine, iterative warping strategy [33, 112, 113]. The warping strategy includes performing successive linearizations of the non-convex data cost term (5.3) about the current estimate $\mathbf{u}_{rk}^{(0)}$,

according to

$$F_D = \Psi_D(I_k(\mathbf{x} + \mathbf{u}_{rk}^{(0)}) + (\nabla I_k)^T \mathbf{du}_{rk} - I_r(\mathbf{x})), \tag{5.5}$$

where $\nabla I_k = \nabla I_k(\mathbf{x} + \mathbf{u}_{rk}^{(0)})$ and $\mathbf{du}_{rk} = \mathbf{u}_{rk} - \mathbf{u}_{rk}^{(0)}$. The coarse-to-fine approach is started by initializing the flow to zero and minimizing (5.2) for downsampled and smoothed versions of the input images. After a number of iterations, the OF estimation proceeds by re-scaling the current flow estimate to a finer resolution level, and continues up until the original resolution of the input images. This is essentially a heuristic method that works well in many cases in its aim to avoid convergence to poor local minima. A pseudo-algorithm for the coarse-to-fine iterative minimization scheme is given in Algorithm 5.1. First, re-sampled images I_r^s, I_k^s are produced at each of the S pixel resolution levels of the coarse-to-fine image pyramid. At each resolution level, flow update terms are obtained by solving the Euler-Lagrange equations associated with the linearized cost functional. An outer iteration loop over n is included that updates the warping point in each iteration, thus linearizing the argument of the data term about the current estimate $\mathbf{u}_{rk}^{(0)}$ as in (5.5). An additional inner iteration loop is required for the general case of non-linear, robust distance functions Ψ_D, Ψ_S, that consists of fixing the occurrences of the flow field variable inside the expressions of the derivatives Ψ'_D, Ψ'_S to the flow estimate at the previous iteration in order to reach a linear expression for the original cost functional. Additional details are available for a similar case in the appendix of Paper 2 and for example in the work of Brox et al. [33].

Despite the almost universal use of coarse-to-fine minimization in OF methods, its limitations have become clear in recent years, when the OF cost functional formulations in modern state-of-the-art methods are considered to be relatively good. The main difficulty, it is argued, is to effectively find the global minimum of the OF cost functional, whose numerical implementation on the image pixel grid is of a very large dimension [114]. The iterative minimization methods that are employed face the challenge of avoiding massive amounts of local minima. To guide the solution towards the global minima, several recent methods perform pre-matching of a sparse set of image locations that contain highly discriminative image features and introduce the motion information from the feature matches into the OF estimation procedure. The introduction of the feature matching information is achieved using one of two different approaches. A first set of methods include a third term E_M to the overall cost functional expression (5.2) and minimize it within the conventional coarse-to-fine estimation strategy while reducing the influence of the feature matches as the resolution levels become finer [107, 114–118]. Another set of methods use edge-aware interpolation

Coarse-to-fine warping algorithm

generate image pyramids $\{I_r^s\}, \{I_k^s\}$, initialize $\mathbf{u}_{rk}^{(0)} = 0, \ \forall \mathbf{x}$
for $s = 1, \ldots, S-1$
  re-sample $\mathbf{u}_{rk}^{(0)}$ to the current resolution level
  for $n = 0, \ldots, N-1$
    compute ∇I_k^s for the current warping points $\mathbf{x} + \mathbf{u}_{rk}^{(0)}$
    set $\mathbf{du}_{rk}^{(n,0)} = 0, \ \forall \mathbf{x}$
    for $l = 0, \ldots, L-1$
      compute $\Psi'_D\{\cdot\}^{(n,l)}$ and $\Psi'_S\{\cdot\}^{(n,l)}$
      $\mathbf{du}_{rk}^{(n,l+1)} \leftarrow$ solve the associated Euler-Lagrange equations
    end
    $\mathbf{u}_{rk}^{(n+1)} = \mathbf{u}_{rk}^{(n)} + \mathbf{du}_{rk}^{(n,L)}$
  end
  $\mathbf{u}_{rk}^{(N)} \rightarrow \mathbf{u}_{rk}^{(0)}$, new warping point
end
output $\mathbf{u}_{rk}^{(N)} \rightarrow \hat{\mathbf{u}}_{rk}$

Algorithm 5.1: Optical flow estimation by minimizing a variational cost functional with the coarse-to-fine warping strategy.

of the flow from the sparse feature matches to obtain a dense, segmented representation of the flow, that is used as initialization to a one-level refinement step [102, 111, 119, 120]. The pre-processing step to extract and match features is typically performed using approximate nearest neighbor (NN) methods [121, 122]. A method that is well suited to perform matching for repetitive image structures is the multi-layered technique of DeepFlow [117]. A downside with using pre-matched features is the significant risk of introducing false matches that can lead to poor initialization of the flow in certain image regions, that are likely to persist in the end result. Due to this, it is important with careful selection of what features to match. For example, including feature matches for image regions where the OF method fares well regardless only introduces an unnecessary risk.

5.2.4 Real-time implementation

From an optical flow application perspective, the ability to compute quality flow field estimates in real-time is often necessary. In 2007, Zach et al.

proposed a solution strategy to (5.2) that achieves real-time performance at 30 frames per second for video inputs at a resolution of 320×240 pixels for a GPU-accelerated (graphics processing unit) implementation [123]. The minimization of $E_{\text{total}}(\mathbf{u}_{rk})$ in (5.2) is performed by forming an equivalent equality constrained problem

$$\begin{aligned} \min_{\widetilde{\mathbf{u}}_{rk}, \mathbf{u}_{rk}} \quad & \int_\Omega F_D(\mathbf{u}_{rk}) + \frac{1}{2\theta}\|\mathbf{u}_{rk} - \widetilde{\mathbf{u}}_{rk}\|_2^2 + \alpha_S F_S(\widetilde{\mathbf{u}}_{rk}) \; d\mathbf{x}, \\ \text{s.t.} \quad & \widetilde{\mathbf{u}}_{rk} = \mathbf{u}_{rk}, \end{aligned} \tag{5.6}$$

where $\widetilde{\mathbf{u}}_{rk}$ is an auxiliary variable. Then, the equality constraint is relaxed and the unconstrained version of (5.6) is solved using alternating minimization of the two coupled subproblems

$$\min_{\widetilde{\mathbf{u}}_{rk}} \int_\Omega \alpha_S F_S(\widetilde{\mathbf{u}}_{rk}) + \frac{1}{2\theta}\|\mathbf{u}_{rk} - \widetilde{\mathbf{u}}_{rk}\|_2^2 \; d\mathbf{x}, \tag{5.7a}$$

$$\min_{\mathbf{u}_{rk}} \int_\Omega F_D(\mathbf{u}_{rk}) + \frac{1}{2\theta}\|\mathbf{u}_{rk} - \widetilde{\mathbf{u}}_{rk}\|_2^2 \; d\mathbf{x} \tag{5.7b}$$

with a coupling weight decided by θ, embedded in a coarse-to-fine multiresolution framework. The cost functionals in (5.7) can be minimized efficiently, with closed-form expressions for their minimizers, if for example

$$F_S = \sqrt{\|\nabla \widetilde{u}_{rk}\|_2^2 + \|\nabla \widetilde{v}_{rk}\|_2^2}, \tag{5.8a}$$

$$F_D = |I_k(\mathbf{x} + \mathbf{u}_{rk}^0) + \nabla I_k^T \mathbf{u}_{rk} - I_r(\mathbf{x})| \; d\mathbf{x}, \tag{5.8b}$$

where F_S integrated over the domain Ω corresponds to the (isotropic) total variation semi-norm $\|\widetilde{\mathbf{u}}_{rk}\|_{TV}$ [124]. With these expressions, the first subproblem (5.7a) corresponds to the classic Rudin-Osher-Fatemi denoising problem that can be solved efficiently using a dual formulation of the TV expression [42, 125–127]. The second subproblem (5.7b) is solved by a soft thresholding operation [3, 123, 127]. Steinbrücker et al. propose to solve (5.7b) without warping or linearizing the data term, and instead use an exhaustive search, such that each subproblem in (5.7) is solved globally (which still does not guarantee convergence to the global minima of the original problem) [128]. While their proposed method is slower, it shows some improvements over using the coarse-to-fine solution strategy. In Paper 3, we minimize an OF cost functional formulation using a primal-dual algorithm which can be derived similarly as in (5.6), but instead setting the auxiliary variable to be equal to the gradient of the flow [127, 129]. The dual subproblem then corresponds to optimizing the convex conjugate of

F_S with respect to the dual variables associated with the introduced equality constraint [130]. The solutions to both subproblems are obtained by evaluating their respective proximal operators [131].

To conclude the section, it seems that systematically investigating different solution strategies to minimize an OF cost functional is of major interest, with regard to computational speed, quality of the convergence point, and the trade-off between these two aspects. Many options can be considered, for instance comparing traditional coarse-to-fine minimization of the original (primal) cost functional E_{total} to primal-dual alternatives. If the cost functional is formulated using non-linear expressions of Ψ_D, Ψ_S, which is the case for both the total variation expression and the absolute value in (5.8), as well as for the commonly used TGV^2 term, these expressions need to be regularized. For instance, the expressions in (5.8) need to be re-formulated according to $\sqrt{z^2} \rightarrow \sqrt{z^2 + \epsilon^2}$, $|z| \rightarrow \sqrt{z^2 + \epsilon^2}$, where ϵ is a small constant. The primal-dual solver readily handles these non-differentiable expressions. What are the consequences of this difference? OF methods that incorporate pre-matched features to initialize or guide the minimization process towards the global minima may perform its pre-processing step to match features using exact nearest neighbor. In other words, they include an exhaustive search similar to what Steinbrücker et al. use, but only for a subset (although sometimes large) of the pixels as opposed to for every pixel. Is there a good balance point with regard to how many features locations to include in an exhaustive NN search to guide the OF estimation? What is the consequence of switching the exact search to an approximate NN search method?

Chapter 6

Optical flow estimation for HDR scenarios

Image-based optical flow estimation is traditionally performed using two or more similarly exposed images. That setup is clearly limiting in HDR scenarios, where an image taken with any given exposure setting contains regions that are either over- or underexposed, due to the insufficient dynamic range of the camera sensor. The dynamic range limitation can be overcome by adding more input images, taken with a different exposure setting, to the OF estimation. To use only two images with different exposure settings is worse than the original setup of two similarly exposed images, as the amount of points that are visible in both images is reduced. The default setup in this chapter is to use image sequences $\{I_k(\mathbf{x})\}$ with 4 frames. The images I_1, I_3 are taken using exposure setting I, which is adjusted to the dim image regions and thus contain other regions that are overexposed. The images I_2, I_4 are taken using exposure setting II, which is adjusted to the bright image regions and therefore leads to other regions being underexposed. The aim is that the combined dynamic range obtained by using two exposure settings is high enough such that all image regions are properly exposed (unsaturated) for at least one of the exposure settings. An example image sequence is shown in Figure 6.1. Four frames from the Sintel [86] *Alley2* sequence have been altered by clipping high intensity regions in I_1, I_3 and low intensity regions in I_2, I_4 to simulate a HDR scenario. Such animated sequences are useful for performance assessment, as ground truth flow data is available. Before proceeding to discuss a revised camera model and the proposed OF method for image sequences with differently exposed frames in the next two sections, we review some related work on motion estimation in HDR scenarios, particularly with a focus on OF methods.

Optical flow estimation, in our work, is pursued mainly for its own purposes as an enabler to motion analysis applications. Other OF methods

Figure 6.1: An image sequences $\{I_k(\mathbf{x})\}$ with 4 frames taken with alternating exposure settings every other frame.

exist in the HDR context as part of HDR image reconstruction methods [9]. Most methods for HDR image reconstruction, however, do not estimate dense motion. They typically opt for simpler motion compensation methods, using a global motion model, in combination with HDR deghosting techniques to avoid reconstruction artifacts [9, 70, 72, 132, 133]. Zimmer et al., however, use optical flow based image alignment as pre-processing in their SR, HDR image reconstruction method [52]. As input they use sets of 5 to 9 images, each taken with a different exposure duration relative to the others. Hafner et al. estimate optical flow and a HDR image jointly using alternating minimization for a common cost functional [53]. Their results indicate that the joint approach benefits both the flow- and the HDR image estimates. The flow estimate benefits from using estimates of the HDR image in an image-driven, anisotropic spatial flow regularization term.

6.1 Image sequences with differently exposed frames

For the proposed method in this chapter, an image is assumed to be generated according to the camera model

$$\widetilde{I}_k(\mathbf{x}) = f(\Phi_k(X(\mathbf{x}) + N_k(\mathbf{x}))), \tag{6.1}$$

where $X(\mathbf{x})$ is the (filtered) illuminance incident on the sensor for the specific lightning condition of the imaged scene at the time instance of the image $\widetilde{I}_k(\mathbf{x})$ and $N_k(\mathbf{x})$ is a noise term. The camera response function, f, as in (2.9), clips the sensor exposure $E_k(\mathbf{x}) = \Phi_k(X(\mathbf{x}) + N_k(\mathbf{x}))$ outside of its operational range $[E_{min}, E_{max}]$. The function Φ_k models the specific exposure

setting used for image k. Under the assumption of brightness constancy, another image $\widetilde{I}_{k+1}$ of the same scene can be related to the non-occluded regions of $\widetilde{I}_k$ through

$$\widetilde{I}_{k+1}(\mathbf{x} + \mathbf{u}_k(\mathbf{x})) = f(\Phi_{k+1}(X(\mathbf{x}) + N_{k+1}(\mathbf{x}))), \tag{6.2}$$

where $\mathbf{u}_k$ denotes the displacement (i.e. the integrated optical flow between the time instances of $\widetilde{I}_k$ and $\widetilde{I}_{k+1}$) of point $\mathbf{x}$ in $\widetilde{I}_k$. The camera model specified by (6.1) and (6.2) is more general than in the related work on optical flow methods for HDR image reconstruction, where only the exposure duration is altered. This model, in addition, allows other exposure settings to be changed, for example to use flash illumination every other frame. If the images are generated using the same camera settings, such that $\Phi_{k+1} = \Phi_k$, OF estimation between $\widetilde{I}_k$ and $\widetilde{I}_{k+1}$ reduces to the traditional case. For the case where $\widetilde{I}_k, \widetilde{I}_{k+1}$ are taken with different exposure durations $\Delta t_1, \Delta t_2$, and all other exposure settings are equal, they are given by

$$\begin{aligned} \widetilde{I}_k(\mathbf{x}) &= f(\Delta t_1(X(\mathbf{x}) + N_k(\mathbf{x}))), \\ \widetilde{I}_{k+1}(\mathbf{x} + \mathbf{u}_k(\mathbf{x})) &= f(\Delta t_2(X(\mathbf{x}) + N_{k+1}(\mathbf{x}))). \end{aligned} \tag{6.3}$$

The images can be aligned photometrically by inverting the effect of the CRF in its non-saturated regions and scaling with the inverse of the respective exposure durations. In the remainder of the chapter, we denote image sequences whose frames are photometrically aligned version the corresponding $\widetilde{I}_k$, if there exists a mathematical expression for the Φ_k to relate them, without the tilde as $\{I_k(\mathbf{x})\}$. For the case when $\widetilde{I}_{k+1}$ is taken with flash but $\widetilde{I}_k$ is not, the flash illuminates the scene in a spatially varying manner, causing changes to $X(\mathbf{x})$ that are difficult to model. It may still be possible to implicitly align the images photometrically by using transformed image functions such as the census transform [98]. Nevertheless, the proposed method in Section 6.3 is designed to handle even such cases where the differently exposed images cannot be aligned photometrically. An advantage of using flash instead of a long exposure duration in order to capture low intensity image regions is that the issue of motion blur can be mitigated. A prototype camera system, mounted on a vehicle, that uses near-infrared flash illumination every other frame is discussed in Paper 2. Sellent et al., however, design an OF method based on adding a long exposed image in between two short exposed images, and specifically aims to model the motion blur in order to aid the OF estimation [105]. Thus, motion blur in the input image sequence is not strictly undesired. In any case, flash illumination is useful if a high frame rate is required.

6.2 Proposed method for OF estimation of HDR image sequences

In this section, which leads up to the appended Paper 2 and Paper 3, summarized in Chapter 7, we outline our proposed method for optical flow estimation on sequences with differently exposed frames. In particular, we consider image sequences $\{I_k(\mathbf{x})\} = I(\mathbf{x}, \{t_k\}),\ k = 1, \ldots, 4$, assuming $t_{k+1} = t_k + 1,\ \forall k$, where every second frame is taken with exposure setting I and II respectively. The objective is to estimate the flow field $\mathbf{u}_2$, that describes the optical flow at the time instance and with respect to the spatial locations of the reference frame, fixed here as I_2. Three flow fields $\mathbf{u}_k = (u_k(\mathbf{x}), v_k(\mathbf{x})),\ k = 1, 2, 3$ are used to form data cost terms

$$\begin{aligned} F_{D13} &= \Psi((I_3(\mathbf{x} + \mathbf{u}_2) - I_1(\mathbf{x} - \mathbf{u}_1))^2), \\ F_{D24} &= \Psi((I_4(\mathbf{x} + \mathbf{u}_2 + \mathbf{u}_3) - I_2(\mathbf{x}))^2), \\ F_{D23} &= \Psi((I_3(\mathbf{x} + \mathbf{u}_2) - I_2(\mathbf{x}))^2), \end{aligned} \tag{6.4}$$

where Ψ is a robust distance function. These terms are summed to form a total data cost

$$F_D = \theta_{13} F_{D13} + \theta_{24} F_{D24} + \theta_{23} F_{D23}, \tag{6.5}$$

where $\theta_{13}(\mathbf{x}), \theta_{24}(\mathbf{x}), \theta_{23}(\mathbf{x})$ are weights that are set to be non-zero for image regions for points where the respective image pair is mutually non-saturated, and in the case of θ_{23}, if the differently exposed I_2 and I_3 are photometrically aligned. Importantly, θ_{23} is set to zero for regions where one of I_2, I_3 is saturated but not the other, as such regions provide false correspondences. Note that the flow fields are not parameterized as $I_1(\mathbf{x} + \widetilde{\mathbf{u}}_1)$, $I_3(\mathbf{x} + \widetilde{\mathbf{u}}_3)$, $I_4(\mathbf{x} + \widetilde{\mathbf{u}}_4)$ to directly relate each of the non-reference images to the reference coordinates $\mathbf{x}$. Instead, the flow field terms in (6.4) describe flow increments between a pair of adjacent frames with respect to the locations of each point in the reference frame [134]. For example, $\mathbf{u}_3(\mathbf{x})$ describes the flow between frames I_3, I_4 of the point that was in location $\mathbf{x}$ in the reference image I_2.

A flow estimate $\hat{\mathbf{u}}_2$ is obtained by minimizing the cost functional

$$E_{\text{total}}(\{\mathbf{u}_k\}) = \int_\Omega F_D + \alpha_S F_S + \alpha_T F_T + \alpha_M F_M \; d\mathbf{x} \tag{6.6}$$

with respect to its arguments $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$, where F_S is a spatial regularization term, F_T is a temporal regularization term, F_M is a feature matching term and $\alpha_S, \alpha_T, \alpha_M$ are their respective weights. The spatial regularization term F_S can either consist of a sum of separate terms for each flow field, such as the expression in (5.8a), or it can be formulated jointly [134]. The temporal regularization term F_T penalizes differences between the flow

terms to enforce the flow to be temporally (piecewise) smooth. Finally, the feature matching term F_M integrates a sparse set of pre-processed feature matches into the minimization, in accordance with the motivation in Section 5.2.3. Additional variables, $l_k(\mathbf{x})$, can be added to the data term of (6.6) to model natural illumination changes that occur between similarly exposed (or photometrically aligned) image pairs, along with a regularization term that enforces piecewise smooth illumination changes. Additive illumination offsets have successfully been integrated into conventional OF formulations [103, 107]. Statistical deviations from the BCA can thus be learned from the image data.

Chapter 7

Summary of included papers

This chapter provides a brief summary of the three papers that are included in Part II of the thesis and how they relate to Part I. The papers have been reformatted to comply with the layout of the thesis. The contents have not been changed, aside from some adaptation of notation, stated at the end of the chapter, to match that of Part I. The first paper addresses SR reconstruction of HDR images, and incorporates the perceptual characteristics of the HVS into the problem formulation. Papers 2 and 3 address OF estimation for HDR scenarios.

Paper 1

In Paper 1, a method is proposed that performs reconstruction of HR, HDR images by solving an inverse problem in an image domain that is designed to be perceptually uniform with respect to the HVS, as opposed to previous work where the problem is formulated directly in the illuminance domain. To a certain extent, Paper 1 is a continuation of Chapter 5, that elaborates on the mathematics of the PU formulation of the HDR SR problem. A nonlinear objective function of the form in (4.4) is proposed. Other choices for $\rho_1(\cdot)$ and $\rho_2(\cdot)$ than the L2 norm in (4.7) are discussed at greater lengths and specifically the use of the Lorentzian norm is evaluated. Whether the image reconstruction is performed in the proposed PU domain or in the illuminance domain, the reconstructed image contains numerical errors across image edges that are of similar magnitudes. Illuminance domain image data, however, needs to be tonemapped for visualization, contrary to the PU domain image that is already in a tonemapped domain. Then, the numerical errors in the darker (low illuminance) regions exhibit themselves as severe artifacts in the tonemapped result. Experimental reconstruction results are presented alongside the objective quality measures PSNR and MSSIM and

demonstrate the benefit of using a PU domain formulation, such as (4.7), as compared to the illuminance domain formulation (4.3). Results on color image sequences, as well as MSSIM quality maps, are provided in our earlier conference papers [135, 136].

Paper 2

In Paper 2, a method is proposed for OF estimation on sequences with differently exposed frames. This setup is useful in HDR scenarios, to avoid poor flow estimates in image regions that contain saturated data. The method is formulated such that any number of input frames can be used, as well as any number of exposure settings. The default setting, however, is to use 4 frames that are taken with two different exposure settings every other frame, as described in Chapter 6. The method works well even when images taken with different exposure settings cannot be related mathematically by photometric alignment, thanks to the flow parametrization in (6.4). We show qualitatively that the performance of OF estimates is degraded due to saturation in the input images. A set of different OF data terms are evaluated quantitatively among themselves and compared to the conventional setup of image sequences that are captured using a single exposure setting. Experimental results are given for two cases, one where the image sequences is generated using different exposure durations and the other where the image sequence is generated using flash illumination every other frame. In the latter case, photometric alignment is not attempted. The best performing data term in a HDR scenario is adopted in Paper 3.

Paper 3

Paper 3 builds on the results of Paper 2, and extends the method for OF estimation on sequences with differently exposed frames to handle challenging scenarios such as natural illumination changes and large displacements of small objects. Illumination changes are included in the modeling of the OF method and flow information from nearest neighbor feature matches are included to aid the estimation of points that exhibit large displacements. The improved performance of flow estimates due to these two additions is shown in qualitative experimental results. The OF estimation is performed by minimizing an OF cost functional using an efficient primal-dual method.

Notational differences

There are a few noteworthy differences in notation between the introductory chapters in Part I and the appended papers. For Paper 1, the notation is completely consistent with that of Part I of the thesis. However, compared to the published version of the paper, $\mathbf{i}_k$ and $\mathbf{y}_k$ are interchanged. For Paper 2 and Paper 3, the significant notational differences are listed in Table 7.1.

	Thesis Part I	Papers 2 and 3
Camera response function	f	CRF
Illuminance	X	R
Exposure	E	X

Table 7.1: Different notation used in Papers 2 and 3 compared to the introductory chapters.

Chapter 8

Concluding remarks

Part I of the thesis is concluded with these final words. Throughout the introductory chapters, two main topics have been discussed; image reconstruction of HR, HDR images as well as optical flow methods. In Chapter 2, the relevance of human visual perception and its connection to digital camera systems was emphasized. The concept of high dynamic range imaging, also introduced therein, has been a theme throughout most of the chapters. In Chapter 3, high dynamic range and super resolution image reconstruction were treated as separate topics. Joint reconstruction of HR, HDR images was discussed in Chapter 4, including a proposed method that takes human visual perception into account in its inverse problem formulation. Interesting future work includes to consider other formulations of the objective function for the inverse SR problem, perhaps replacing the pointwise PSNR-like measure with a structure-aware measure, inspired by MSSIM, that correlates better with perceived image quality. Conventional optical flow methods that estimate motion based on two or more images of the same scene were introduced in Chapter 5, as a background to the proposed method for OF estimation in HDR scenarios in Chapter 6. It is of major interest to perform a measurement campaign to capture real-world HDR scenes with differently exposed frames, as a basis for further investigation of the proposed OF method. The three papers that are appended in the thesis were summarized in Chapter 7, to bridge between the introductory chapters and Part II.

References

[1] P. Soille, *Morphological image analysis: principles and applications*. Springer, Berlin-Heidelberg, 2013.

[2] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[3] A. Wedel and D. Cremers, *Stereo scene flow for 3D motion analysis*. Springer, London, 2011.

[4] D. Geronimo, A. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1239–1258, 2010.

[5] M. Abràmoff, P. Magalhães, and S. Ram, "Image processing with ImageJ," *Biophotonics international*, vol. 11, no. 7, pp. 36–42, 2004.

[6] J. Prince and J. Links, *Medical imaging signals and systems*. Pearson Prentice Hall Upper Saddle River, NJ, 2006.

[7] E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann, San Francisco, 2010.

[8] P. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Conference on Computer Graphics and Interactive Techniques*, 1997, pp. 369–378.

[9] O. Tursun, A. Akyüz, A. Erdem, and E. Erdem, "The state of the art in HDR deghosting: A survey and evaluation," in *Computer Graphics Forum*, vol. 32, 2015, pp. 348–362.

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

[11] P. Milanfar, *Super-resolution imaging.* CRC Press, 2010.

[12] A. Katsaggelos, R. Molina, and J. Mateos, *Super resolution of images and video.* Morgan & Claypool, 2007.

[13] J. Kuang, G. Johnson, and M. Fairchild, "iCAM06: A refined image appearance model for HDR image rendering," *Journal of Visual Communication and Image Representation*, vol. 18, pp. 406–414, 2007.

[14] 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.

[15] M. Angelopoulou, C.-S. Bouganis, P. Cheung, and G. Constantinides, "Robust real-time super-resolution on FPGA and an application to video enhancement," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 2, no. 4, 2009.

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

[17] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[18] D. Cremers and S. Soatto, "Motion competition: A variational approach to piecewise parametric motion segmentation," *International Journal of Computer Vision*, vol. 62, no. 3, pp. 249–265, 2005.

[19] A. Giachetti, M. Campani, and V. Torre, "The use of optical flow for road navigation," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 1, pp. 34–48, 1998.

[20] W. Crum, T. Hartkens, and D. Hill, "Non-rigid image registration: theory and practice," *The British Journal of Radiology*, vol. 77, pp. 140–153, 2004.

[21] S. Baker and T. Kanade, "Super-resolution optical flow," *Technical Report CMU-RI-TR-99-36, Carnegie Mellon University*, 1999.

[22] R. Fransens, C. Strecha, and L. Van Gool, "Optical flow based super-resolution: A probabilistic approach," *Computer vision and image understanding*, vol. 106, no. 1, pp. 106–115, 2007.

[23] D. Mitzel, T. Pock, T. Schoenemann, and D. Cremers, "Video super resolution using duality based TV-L1 optical flow," in *Pattern Recognition*, 2009, pp. 432–441.

[24] C. Liu and D. Sun, "On bayesian adaptive video super resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 346–360, 2014.

[25] E. Allen and S. Triantaphillidou, *The Manual of Photography and Digital Imaging.* Focal Press, 2011.

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

[27] X. Li, B. Gunturk, and L. Zhang, "Image demosaicing: A systematic survey," in *Proceedings of SPIE*, vol. 6822, 2008.

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

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

[30] G. Wyszecki and W. Stiles, *Color science,* vol. 8. Wiley, New York, 1982.

[31] R. W. Hunt and M. Pointer, *Measuring colour.* John Wiley & Sons, 2011.

[32] T. Aydin, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," *Proceedings of SPIE*, vol. 6806, 2008.

[33] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conference on Computer Vision (ECCV)*, 2004, pp. 25–36.

[34] International Color Consortium, "Specification of sRGB," 2015. [Online]. Available: http://www.color.org/sRGB.pdf

[35] M. Anderson, R. Motta, S. Chandrasekar, and M. Stokes, "Proposal for a standard default color space for the internet - sRGB," in *Color and imaging conference*, vol. 1. Society for Imaging Science and Technology, 1996, pp. 238–245.

[36] M. Irani and S. Peleg, "Improving resolution by image registration," *Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991.

[37] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.

[38] 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, 2004.

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

[40] M. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer vision and image understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[41] V. Patanavijit and S. Jitapunkul, "A Lorentzian stochastic estimation for a robust iterative multiframe super-resolution reconstruction with Lorentzian-Tikhonov regularization," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 2, pp. 21–21, 2007.

[42] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[43] A. Chambolle and P. Lions, "Image recovery via total variation minimization and related problems," *Numerische Mathematik*, vol. 76, no. 2, pp. 167–188, 1997.

[44] P. Blomgren and T. Chan, "Color TV: total variation methods for restoration of vector-valued images," *IEEE Transactions on Image Processing*, vol. 7, no. 3, pp. 304–309, 1998.

[45] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[46] R. Hardie, K. Barnard, and E. Armstrong, "Joint map registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997.

[47] L. Pickup, "Machine learning in multi-frame image super-resolution," Ph.D. dissertation, Oxford University, 2007.

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

[49] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 977–984.

[50] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[51] D. Sun, S. Roth, J. Lewis, and M. Black, "Learning optical flow," in *European Conference on Computer Vision (ECCV)*, 2008, pp. 83–97.

[52] 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.

[53] D. Hafner, O. Demetz, and J. Weickert, "Simultaneous HDR and optic flow computation," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2014, pp. 2065–2070.

[54] P. Sen, N. Kalantari, M. Yaesoubi, S. Darabi, D. Goldman, and E. Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes." *ACM Transactions on Graphics*, vol. 31, no. 6, p. 203, 2012.

[55] K. Hadziabdic, J. Telalovic, and R. Mantiuk, "Comparison of deghosting algorithms for multi-exposure high dynamic range imaging," in *Proceedings of the 29th Spring Conference on Computer Graphics*, 2013, pp. 21–28.

[56] M. Granados, B. Ajdin, M. Wand, C. Theobalt, H.-P. Seidel, and H. Lensch, "Optimal HDR reconstruction with linear digital cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 215–222.

[57] J. Choi, M. Park, and M. Kang, "High dynamic range image reconstruction with spatial resolution enhancement," *The Computer Journal*, vol. 52, pp. 114–125, 2009.

[58] 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.

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

[60] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[61] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

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

[63] 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, 2006.

[64] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 81, 1981, pp. 674–679.

[65] 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, 2007.

[66] S. Harmeling, S. Sra, M. Hirsch, and B. Schölkopf, "Multiframe blind deconvolution, super-resolution, and saturation correction via incremental EM," in *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 3313–3316.

[67] G. Harikumar and Y. Bresler, "Perfect blind restoration of images blurred by multiple filters: Theory and efficient algorithms," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 202–219, 1999.

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

[69] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Transactions on Image Processing*, vol. 18, no. 1, pp. 36–51, 2009.

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

[71] Y. Traonmilin and C. Aguerrebere, "Simultaneous high dynamic range and superresolution imaging without regularization," *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1624–1644, 2014.

[72] M. Gevrekci and B. Gunturk, "Superresolution under photometric diversity of images," *EURASIP Journal of Applied Signal Processing*, vol. 2007, 2007.

[73] M. Grossberg and S. Nayar, "Determining the camera response from images: What is knowable?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1455–1467, 2003.

[74] J. Hu, O. Gallo, and K. Pulli, "Exposure stacks of live scenes with hand-held cameras," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 499–512.

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

[76] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011.

[77] Baker, S. and Scharstein, D. and Lewis, J. and Roth, S. and Black, M. and Szeliski, R., "The middlebury computer vision pages, optical flow page," 2015. [Online]. Available: http://vision.middlebury.edu/flow/eval/

[78] M. Robertson, S. Borman, and R. Stevenson, "Dynamic range improvement through multiple exposures," in *IEEE International Conference on Image Processing (ICIP)*, vol. 3, 1999, pp. 159–163.

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

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

[81] 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.

[82] J. Gibson, *The perception of the visual world*. Houghton Mifflin, Boston, 1950.

[83] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 404–417.

[84] O. Miksik and K. Mikolajczyk, "Evaluation of local detectors and descriptors for fast feature matching," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2012, pp. 2681–2684.

[85] R. Newcombe and A. Davison, "Live dense reconstruction with a single moving camera," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1498–1505.

[86] D. Butler, J. Wulff, G. Stanley, and M. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 611–625.

[87] ——, "MPI Sintel flow dataset," 2015. [Online]. Available: http://sintel.is.tue.mpg.de/

[88] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[89] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "KITTI vision benchmark suite," 2015. [Online]. Available: http://www.cvlibs.net/datasets/kitti/

[90] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[91] D. Kondermann, S. Abraham, G. Brostow, W. Förstner, S. Gehrig, A. Imiya, B. Jähne, F. Klose, M. Magnor, H. Mayer *et al.*, "On performance analysis of optical flow algorithms," in *Outdoor and Large-Scale Real-World Scene Analysis*, 2012, pp. 329–355.

[92] C. Vogel, S. Roth, and K. Schindler, "An evaluation of data costs for optical flow," in *German Conference on Pattern Recognition (GPCR)*, 2013, pp. 343–353.

[93] D. Sun, S. Roth, and M. Black, "Secrets of optical flow estimation and their principles," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2432–2439.

[94] I. Ekeland and R. Témam, *Convex analysis and variational problems*. SIAM, Philadelphia, PA, 1999.

[95] I. Gelfand and S. Fomin, *Calculus of variations*. Courier Corporation, 2000.

[96] G. Strang, *Computational science and engineering*. Wellesley-Cambridge Press, 2007.

[97] 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.

[98] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European Conference on Computer Vision (ECCV)*, 1994, pp. 151–158.

[99] T. Müller, C. Rabe, J. Rannacher, U. Franke, and R. Mester, "Illumination-robust dense optical flow using census signatures," in *Pattern Recognition*. Springer, 2011, pp. 236–245.

[100] D. Hafner, O. Demetz, and J. Weickert, "Why is the census transform good for robust optic flow computation?" in *Scale Space and Variational Methods in Computer Vision*, ser. Lecture Notes in Computer Science. Springer, Berlin-Heidelberg, 2013, vol. 7893, pp. 210–221.

[101] H. Zimmer, A. Bruhn, and J. Weickert, "Optic flow in harmony," *International Journal of Computer Vision*, vol. 93, no. 3, pp. 368–388, 2011.

[102] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Epicflow: Edge-preserving interpolation of correspondences for optical flow," *arXiv preprint arXiv:1501.02565*, 2015.

[103] O. Demetz, M. Stoll, S. Volz, J. Weickert, and A. Bruhn, "Learning brightness transfer functions for the joint recovery of illumination changes and optical flow," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 455–471.

[104] L. Alvarez, R. Deriche, T. Papadopoulo, and J. Sánchez, "Symmetrical dense optical flow estimation with occlusions detection," *International Journal of Computer Vision*, vol. 75, no. 3, pp. 371–385, 2007.

[105] A. Sellent, M. Eisemann, B. Goldlücke, D. Cremers, and M. Magnor, "Motion field estimation from alternate exposure images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1577–1589, 2011.

[106] D. Sun, J. Wulff, E. Sudderth, H. Pfister, and M. Black, "A fully-connected layered model of foreground and background flow," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2451–2458.

[107] R. Kennedy and C. Taylor, "Optical flow with geometric occlusion estimation and fusion of multiple frames," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2015, pp. 364–377.

[108] J. Weickert and C. Schnörr, "A theoretical framework for convex regularizers in pde-based computation of image motion," *International Journal of Computer Vision*, vol. 45, no. 3, pp. 245–264, 2001.

[109] K. Bredies, "Recovering piecewise smooth multichannel images by minimization of convex functionals with total generalized variation penalty," in *Efficient Algorithms for Global Optimization Methods in Computer Vision*, 2014, pp. 44–77.

[110] R. Ranftl, K. Bredies, and T. Pock, "Non-local total generalized variation for optical flow estimation," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 439–454.

[111] M. Leordeanu, A. Zanfir, and C. Sminchisescu, "Locally affine sparse-to-dense matching for motion and occlusion estimation," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[112] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *International Journal of Computer Vision*, vol. 2, no. 3, pp. 283–310, 1989.

[113] E. Memin and P. Perez, "A multigrid approach for hierarchical motion estimation," in *IEEE International Conference on Computer Vision (ICCV)*, 1998, pp. 933–938.

[114] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, 2011.

[115] M. Stoll, S. Volz, and A. Bruhn, "Adaptive integration of feature matches into variational optical flow methods," in *Asian Conference on Computer Vision (ACCV)*, 2013, pp. 1–14.

[116] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1744–1757, 2012.

[117] P. Weinzaepfel, J. Revaud, Z. Harchaoui, C. Schmid *et al.*, "Deepflow: Large displacement optical flow with deep matching," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[118] R. Timofte and L. Van Gool, "Sparseflow: Sparse matching for small to large displacement optical flow," in *Conference on Applications of Computer Vision (WACV)*, 2015, pp. 1100–1106.

[119] Z. Chen, H. Jin, Z. Lin, S. Cohen, and Y. Wu, "Large displacement optical flow from nearest neighbor fields," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2443–2450.

[120] L. Bao, Q. Yang, and H. Jin, "Fast edge-preserving patchmatch for large displacement optical flow," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 4996–5006, 2014.

[121] M. Muja and D. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," *International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 331–340, 2009.

[122] K. He and J. Sun, "Computing nearest-neighbor fields via propagation-assisted KD-trees," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 111–118.

[123] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," in *German Conference on Pattern Recognition (GPCR)*, 2007, pp. 214–223.

[124] Y. Lou, T. Zeng, S. Osher, and J. Xin, "A weighted difference of anisotropic and isotropic total variation model for image processing," *SIAM Journal on Imaging Sciences*, 2015.

[125] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical imaging and vision*, vol. 20, no. 1-2, pp. 89–97, 2004.

[126] M. Zhu and T. Chan, "An efficient primal-dual hybrid gradient algorithm for total variation image restoration," *UCLA CAM Report*, pp. 08–34, 2008.

[127] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[128] F. Steinbrucker, T. Pock, and D. Cremers, "Large displacement optical flow computation without warping," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 1609–1614.

[129] V. Estellers and S. Soatto, "Detecting occlusions as an inverse problem," *Journal of Mathematical Imaging and Vision*, pp. 1–18, 2015.

[130] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2009.

[131] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[132] P.-Y. Lu, T.-H. Huang, M.-S. Wu, Y.-T. Cheng, and Y.-Y. Chuang, "High dynamic range image reconstruction from hand-held cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 509–516.

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

[134] S. Volz, A. Bruhn, L. Valgaerts, and H. Zimmer, "Modeling temporal coherence for optical flow," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1116–1123.

[135] T. Bengtsson, I.-H. Gu, M. Viberg, and K. Lindström, "Regularized optimization for joint super-resolution and high dynamic range image reconstruction in a percetually uniform domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 1097–1100.

[136] T. Bengtsson, T. McKelvey, and I.-H. Gu, "Super-resolution reconstruction of high dynamic range images with perceptual weighting of errors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 2212–2216.