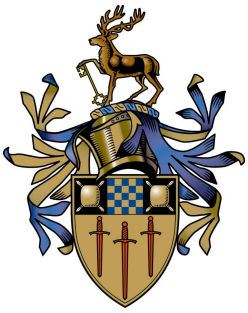
Image Enhancement and Fusion Methods For Mobile Camera Platforms

Falk Schubert

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey



Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

July 2013

Summary

Mobile camera platforms are becoming omni-present, spawning many new applications. This trend is accelerated by cheap, small and powerful cameras. Although the cameras in these mobile devices are already of good quality, they still have physical limitations in terms of spatial resolution, dynamic range and temporal information. In this thesis we investigate algorithmic solutions using image fusion to overcome those limitations. We consider three main techniques: super-resolution, high-dynamic-range imaging and motion detection. The first one aims at generating high-resolution images from low-resolution input videos. The second one combines multiple images taken with different camera settings to generate an image with a greater dynamic range than in any of the input images. The third technique analyzes multiple consecutive frames in a video to extract pixels belonging to moving objects. As mobile cameras are not static, they impose the challenge of compensating the ego-motion. Therefore, we investigate for each of the image fusion approaches the required image registration steps and propose the best suited algorithms.

For the application of increasing resolution, we present a solution to enhance low-resolution videos using multi-frame reconstruction-based superresolution. We propose to use high-resolution images from the Internet as priors within a maximum-a-posteriori formulation. We demonstrate that this superresolution framework increases the resolution of low-quality input videos taken with mobile cameras.

Similar to superresolution, high-dynamic-range imaging is also an algorithmic solution to overcome the physical limitation of a sensor. Many approaches have been proposed for both enhancement applications individually, but little research has been carried out to provide solutions which address both problems simultaneously. We present an approach that combines multi-frame reconstruction-based superresolution and a new minimal high-dynamic-range imaging method into a unified framework.

Relating multiple consecutive frames from a video allows to detect moving objects. This task requires many registration operations, hence demanding very efficient registration methods. We present a fast algorithm to compute homographies based on phase correlation, which benefits from implementations using GPUs. Despite accurate registration, errors in the motion extraction process due to parallax and sensor noise are inevitable and generate false-alarms. We present an efficient filtering scheme, which significantly reduces the false detections, thus improves the performance of moving object detection.

Many image enhancement algorithms that increase the level of image details are motivated by the hope that they improve the performance of subsequent computer vision tasks. We investigate how much the performance of two common applications, i.e. image retrieval and scene recognition, can be increased when applying image filters as a pre-processing step. We consider standard and advanced gradient-based filtering techniques using state-of-the-art benchmark datasets for evaluation and show that reducing the level of image details, e.g. in terms of image abstraction, improves retrieval and recognition.

Keywords: Registration, Superresolution, High Dynamic Range Imaging, Moving Object Detection, Gradient Domain Rendering, Image Fusion, Image Enhancement, Phase Correlation, Computational Photography

Email: falk.schubert@gmail.com

WWW: http://www.imagefusion.eu

Contents

List of Figures

List of Tables

1 Introduction

Many Images Are Better Than One.

Mobile camera platforms are more and more present in various industrial applications but also in the consumer market. This trend is fostered by the availability of cheap, small and powerful cameras. However, the sensors in those digital cameras exhibit several types of limitations and discretization in terms of spatial resolution, dynamic range and temporal information. For instance the number of pixel elements of a sensor is fixed and defines the resolution of the captured images. In this thesis we investigate algorithmic solutions using image fusion and image enhancement methods to overcome those hardware limitations. For static setups image fusion techniques are already well studied. However, for moving platforms the ego-motion of the camera needs to be compensated, requiring adequate image registration. We investigate how image fusion techniques can be extended and applied to produce useful results on images captured with moving camera platforms. We focus on three methods of image fusion: superresolution, high-dynamic-range (HDR) imaging and motion detection. These methods are of great interest to mobile cameras and in closely related surveillance tasks.

The old saying "Two eyes see more than one" is very relevant for the focus of this thesis. In statistics the support of multiple measurements allows to make more precise observations than with one measurement alone. In the field of computer vision image fusion techniques draw from this phenomenon. In the following we give a brief introduction to the most common image fusion methods, state the objectives and explain the context of this thesis. Finally we summarize the main contributions.

1.1 Image Fusion Methods

Limitation: Spatial Resolution **Solution: Superresolution**

In certain applications such as forensic investigations the spatial resolution of a video is often not sufficient to recognize the content, for instance for identifying a suspect in a video by its face or a car by its license plate. However, the imaging process in cameras introduces several artifacts including blur, downscale and compression, which reduce the image content, i.e. the level of detail. This results from limited spatial discretization of the camera sensor, the digital processing inside the camera and blurring caused by lenses and atmosphere. To overcome those, image reconstruction methods such as superresolution [130] have been proposed. These combine multiple low-resolution input images to reconstruct a high-resolution image, which has finer details and a greater number of pixels. Fig. 1.1 depicts typical improvements achievable with superresolution.

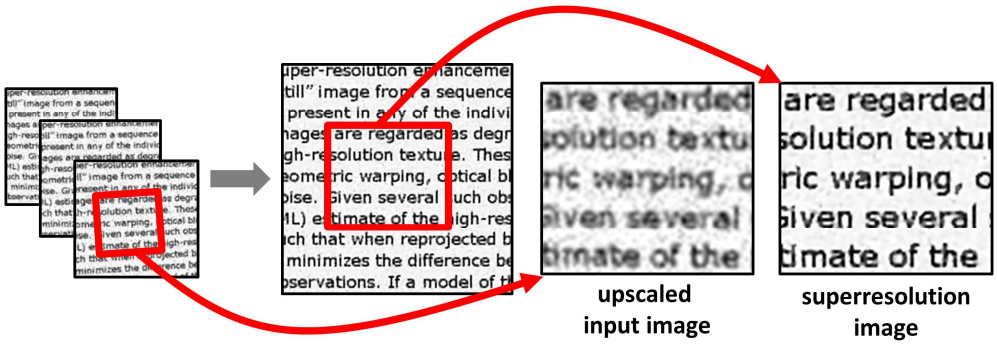


Fig. 1.1: Comparison of up-scaled low-resolution input with superresolution result.

Limitation: Dynamic Range **Solution: HDR Imaging**

Camera sensors also have limitations in the range of light intensities, i.e. radiance values, that can be measured and distinguished. The range of light hitting the sensor is usually dramatically compressed into a much smaller range of intensity values, e.g. 256 for 8-bit images as used in the common JPEG format. This causes very bright/dark areas to appear over-/under-exposed in the captured image, especially if the 8-bit range is not efficiently used. However, the human eye can still perceive all the details in the real

scene in those areas. In applications such as surveillance, where cameras have to cover differently illuminated areas within a single view, image fusion can be used to address this issue. Multiple input images with low-dynamic range are recorded with different exposure settings, e.g. varying the exposure time, and fused to recover an image with a much higher dynamic range [38]. In Fig. 1.2 an example of such fusion is illustrated.

Fig. 1.2: Two low-dynamic-range images are combined to produce HDR result.

Limitation: Temporal Dimension **Solution: Motion Detection**

In some surveillance applications from aerial camera platforms, detecting moving objects is an essential part of analyzing the scene and recognizing anomalies of ground-based objects. Comparing two or more images, which have been recorded at different times, allows to identify areas that have changed over time (see Fig. 1.3). Typically a background image is constructed by fusing multiple consecutive video frames and all other frames are compared to this background, where differences between them can indicate changes due to motion, but also due to registration artifacts caused by motion blur [3].

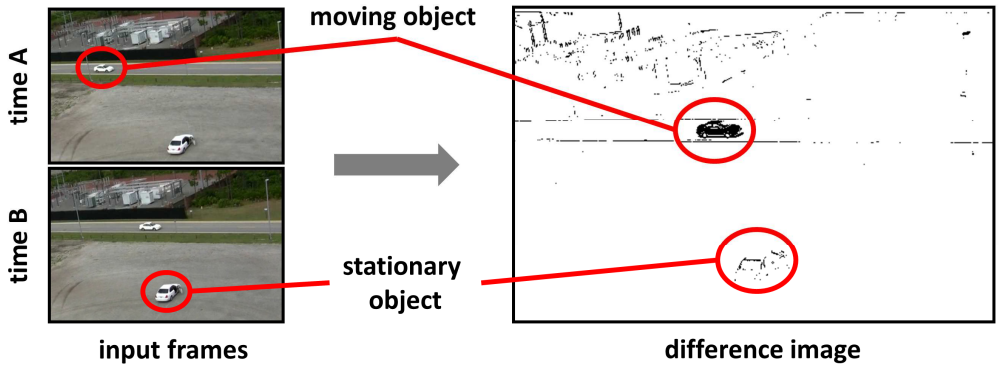


Fig. 1.3: Illustration of image fusion for moving object detection [5].

Limitation: Field-Of-View **Solution: Mosaicing**

Typical cameras have a limited field-of-view to capture a scene, which is smaller than the one of the human eye. Using special lenses and apertures it is possible to significantly increase this viewing angle. In cases where the field-of-view is limited, e.g. a fixed lens is used, and yet a larger scene needs to be explored, multiple images can be stitched together to form a large mosaic image. Such a panorama image is constructed out of multiple overlapping small input images [157]. Fig. 1.4 illustrates this fusion process.

Fig. 1.4: Illustration of image fusion for mosaicing.

Limitation: 2D Images **Solution: Structure From Motion**

During the imaging process the 3D nature of the real world scene is lost as the light rays are projected onto the 2D plane of the camera sensor. However, it is possible to recover the 3D scenery using multiple images recorded from different viewpoints. This can be achieved by moving a single camera to generate those different views, which is called *structure from motion* [40] and is illustrated in Fig. 1.5. Another possible method is based on calculating the *visual hulls* of 3D objects from their 2D recordings [151].

Fig. 1.5: Illustration of image fusion for 3D reconstruction.

Limitation: Band-Of-Focus **Solution: Multi-Focus Image Fusion**

Many cameras have a lens system that can only focus on either very close or very distant objects. After the image has been taken this point of focus can not be changed. In such case multiple input images of the same scene, each with a different focus in depth can be merged to compute an image with focused areas at all depth levels or to refocus at a different point [52]. In Fig. 1.6 an example result generated by fusing two images is depicted.

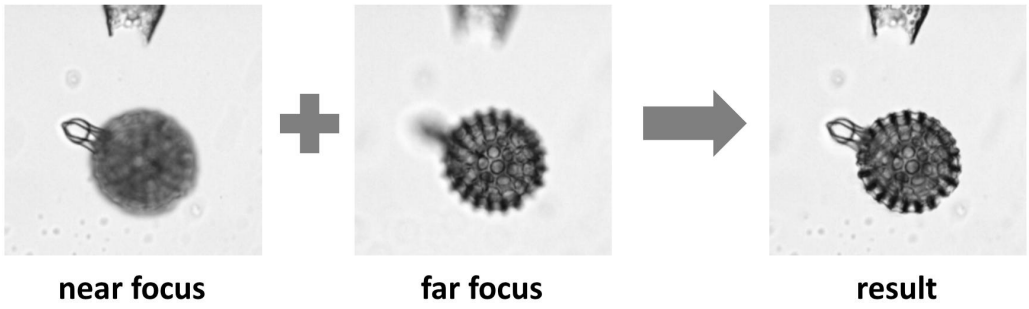


Fig. 1.6: Combining differently focused images to produce result with greater focus range [52].

Limitation: Single Sensor **Solution: Multi-Modal Image Fusion**

In most cameras the image is generate from the read-out of a single sensor. However, often multiple sensors, which are sensitive to different wavelength, capture complementary information from the scene. For instance, to detect humans robustly and independent of illumination and temperature conditions, it is helpful to combine two images captured by two cameras, one sensitive to visible light and the other to infra-red light [37]. An example result of such fusion is depicted in Fig. 1.7.

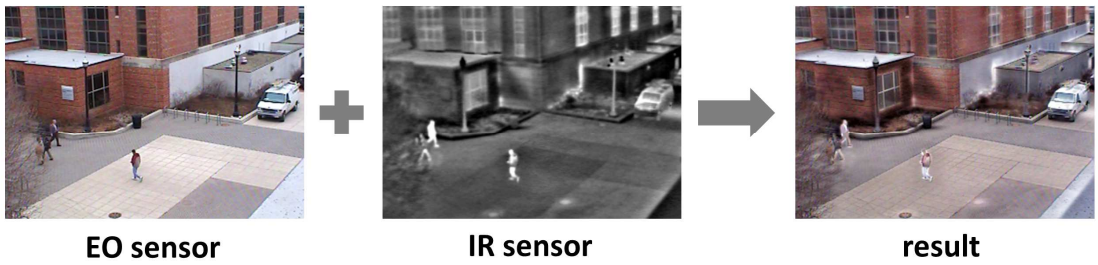


Fig. 1.7: Illustration of image fusion for multi-modal sensor fusion [37].

1.2 Objectives

Although image fusion methods have been extensively studied in the past, most of the work did not explicitly address mobile cameras. Depending on the type of camera and application different kind of camera motion is to be expected. In consumer photography the imagery is typically acquired with hand-held cameras, which have very cheap sensors and are subject to arbitrary motion. In aerial imagery the motion is much more defined and the sensors are of better quality. The objective for this thesis is to investigate the impact of camera motion in general for the three fusion methods: superresolution, high-dynamic-range imaging and motion detection. Depending on the fusion methods and the corresponding application, this requires the investigation of different registration methods. It is the goal to find out what registration methods are suitable for the individual fusion methods or to develop new ones in case the existing state-of-the-art is not sufficient. The types of cameras considered depend on the application and range from hand-held cameras (e.g. mobile phones, digital cameras) with cheap sensors up to high-quality airborne cameras.

For superresolution, camera motion is a fundamental requirement as many images with subtle motion changes are capturing the scene from slightly different viewpoints, which in the end allows the reconstruction of an image with a higher resolution. In this thesis we investigate how to apply this fusion method in applications with wide baseline camera motion. The input images and videos are captured with a digital consumer camera.

High-Dynamic-Range imaging usually assumes static cameras and static scenery. In this thesis we investigate how to handle subtle dynamic scene changes or registration errors resulting from moving cameras. The input images captured with a digital consumer camera. In addition, current state-of-the-art methods require the acquisition of more than just two input image in order to produce useful results. It is the goal of this thesis to explore ways on minimizing these requirements to enlarge the range of possible applications.

Moving object detection is a common application for aerial platforms (i.e. unmanned aerial vehicles). A core component of this fusion method is a fast registration. Due to the hardware restrictions for onboard processing and the employed platforms (e.g. only FPGAs are available to speed-up computations rather than multi-core consumer computers) we find that existing methods are not fast enough when ported to the available, certifiable onboard platforms.

In this thesis, we therefore aim to develop a new algorithm which meets those requirements. As aerial camera platforms are subject to turbulences in the air as well as motion blur, the errors in image registration have to be handled in the moving object detection step. In addition, we investigate filtering methods that efficiently suppress such errors and significantly reduce the high false-positive rate in current state-of-the-art methods.

1.3 Contributions

Registration The key prerequisite to any multi-frame fusion is registration. Depending on the application and fusion method various requirements have to be met by the registration step. Some methods such as superresolution, require accurate alignment. Other techniques such as multi-modal sensor fusion can handle loose alignment, e.g. two images from two different sensors are only merged in form of an overlay. For real-time applications, e.g. on-board processing of aerial videos, the registration has to be efficient as already the fusion process itself is computationally demanding. For instance high-quality superresolution algorithms take up to minutes per frame to produce a result. We therefore consider two important criteria for the registration: speed and accuracy.

Although many different approaches have already been proposed for fast and accurate registration methods, the research area of geometric and photometric image alignment is still very active as the registration problem is far from being solved. In this thesis we investigate different registration methods and evaluate their performance for the three image fusion techniques considered. Furthermore, we present a very efficient, new

registration method which is suitable for many real-time applications such as moving object detection. The algorithm is based on phase correlation, which benefits from parallel implementations using graphic cards. This work is presented in chapter 2.

Superresolution has a long history in the research community and has picked up new momentum in the last years. Although basic reconstruction-based superresolution produces notable enhancement, fundamental limits prohibit further improvement of the input images. Recent developments in the research community encourage to use as much a-priori information as possible to lift the enhancement effect one step up. We present a method that combines such a-priori information with a reconstruction-based formulation. We apply this superresolution scheme to combine high-quality still images with low-resolution videos. This work is presented in chapter 3.

High-Dynamic-Range Imaging Superresolution can be interpreted as a fusion algorithm that uses redundant low-resolution image information to reconstruct a hidden, underlying high-quality image. Along similar lines, high-dynamic-range imaging can be interpreted as a fusion algorithm, that uses complementary low-dynamic-range image information to reconstruct a hidden, underlying high-dynamic-range image. We present a method that combines both of these enhancement methods into a single framework, which produces high-dynamic-range and high-resolution images from low-resolution, low-dynamic-range videos. Furthermore, we present an optimized high-dynamic-range fusion process called *Minimal-HDR*, which uses only two input images to generate a result image instead of many, which is typical for state-of-the-art methods. This work is presented in chapter 4.

Advanced Filter Methods Much research effort in the literature is focused on improving recognition and retrieval tasks. Besides various modifications of the individual algorithmic steps, pre-processing the input data with image enhancement methods is one way to achieve this. It is the hope, that increasing the level of detail will im-

prove the overall performance of computer vision tasks such as recognition and retrieval. However, as most state-of-the-art approaches use gradient-based features, it is not clear whether realistically looking images produce best results or ones with emphasized or suppressed gradients. We therefore present a performance evaluation of numerous existing advanced image filtering techniques that explicitly alter the gradients of an image. We show that for some applications, image abstraction, i.e. reducing the level of detail, actually improve the performance of state-of-the-art feature extraction methods. This work is presented in chapter 5.

Moving Object Detection Multi-frame motion detection approaches are far better suited for aerial camera platforms than methods based on two-frame-differencing. Because this requires many registration operations to reconstruct the background, we employ the high frame-rate tiled phase correlation which is presented in chapter 2. However, despite accurate registration, errors in the motion extraction process due to parallax and sensor noise are inevitable and generate false-alarms. To handle these errors, we present an efficient filtering step to reduce incorrect motion hypotheses, that arise from background subtraction. We show that the proposed filtering significantly improves the precision of the motion detection, while maintaining high recall. This work is presented in chapter 6.

In chapter 7 the research presented in this thesis is summarized and directions for future work are outlined.

1.4 Publications

Thesis Related Publications The following papers have been published during the PhD studies which are summarized in this thesis:

1. F. Schubert and K. Mikolajczyk, “Combining High-Resolution Images With Low-Quality Videos”, BMVC, 2008

2. F. Schubert, K. Schertler and K. Mikolajczyk, "A Hands-On Approach To High-Dynamic-Range And Superresolution Fusion", WACV, 2009

3. F. Schubert and K. Mikolajczyk, "Performance Evaluation of Image Filtering for Classification and Retrieval", ICPRAM, 2013

4. F. Schubert and K. Mikolajczyk, "Benchmarking GPU-Based Phase Correlation For Homography-Based Registration Of Aerial Imagery", CAIP, 2013

5. F. Schubert and K. Mikolajczyk, "Robust Registration and Filtering For Moving Object Detection In Aerial Videos", BMVC, 2013 (**submitted**)

Other Publications As a collaborative student, additional work was carried out as an employee of the EADS Cooperation "European Aeronautic Defense and Space Company". During the PhD studies the following papers have been published as part of the professional work at the company:

1. T. Waanders, Q. Qian, R. Scheiblhofer, B. van Noort, R. Körber, A. Giere, F. Schubert and V. Ziegler, "Helicopter Miniaturized And Low Cost Obstacle Warning System", European Rotorcraft Forum, 2012

2. V. Ziegler, F. Schubert, B. Schulte, A. Giere, R. Körber and T. Waanders, "Low Cost And Miniaturized Helicopter Near Field Obstacle Warning Radar", International Microwave Symposium, 2012

3. V. Ziegler, F. Schubert, B. Schulte, A. Giere, R. Körber and T. Waanders, "Helicopter Near-Field Obstacle Warning System Based on Low-Cost Millimeter-Wave Radar Technology", Transactions on Microwave Theory and Techniques, 2013

4. V. Belagiannis, F. Schubert, N. Navab and S. Ilic, "Segmentation Based Particle Filtering for Real-Time 2D Object Tracking", ECCV, 2012

5. F. Schubert, T. Fath and H. Haas, "Coloured Video Code For In-Flight Data Transmission", ICVS, 2013

6. R. Fernandez, F. Schubert, V. Potyka, "Flexible Trajectory Planning Framework For Rotary And Fixed-Wing Aircrafts Target-Pursuit Using Optimal Control", SAE AeroTech Congress, 2013

Patents Furthermore, the following national (Germany) patents have been granted or filed, which are based on the PhD work at the company:

1. K. Schertler, J. Liebelt, F. Schubert, "Bildverarbeitungsvorrichtung", DE102008014381 (B4), 2010 (**granted**)
2. F. Schubert, K. Schertler, "Verfahren und Einrichtung zur Erzeugung von fehlerreduzierten hochauflösenden und kontrastverbesserten Bildern", DE102008034979 (B4), 2011 (**granted**)
3. K. Schertler, F. Schubert, "Bilderfassungsvorrichtung und Verfahren zum Reduzieren von Bewegungsunschärfe", DE102009057724 (A1), 2011 (**filed**)

Reviewer Work During the PhD studies the student has actively taken part in the research community by reviewing papers for the following conferences and journals:

- IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)
- British Machine Vision Conference (BMVC)
- European Conference on Computer Vision (ECCV)
- IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- International Conference on Pattern Recognition (ICPR)
- IEEE International Conference on Computer Vision (ICCV)
- Journal of Electronic Imaging (JEI)
- ACM Workshop Multimedia in Forensics, Security and Intelligence (MiFor)

2 Image Registration

No Image Fusion Without Good Registration.

Image registration is a fundamental prerequisite to image fusion. Each fusion method imposes individual requirements for the registration. For instance superresolution requires very accurate alignment. On the opposite moving object detection requires very fast registration rather than the most accurate one. Therefore, different registration methods that need to be considered in the thesis we discuss in this chapter. In section 2.1 we introduce the characteristics of image registration methods that are important to our fusion methods. In section 2.2 we summarize the main challenges in image registration and how they are addressed in the literature. In section 2.3 we discuss details of the existing algorithms which we employ in the following chapters. Finally, in section 2.4 we present our new registration method which is faster than the state-of-the-art approaches.

2.1 Introduction

Fast registration algorithms are the key component in real-time image fusion applications such as mosaicing, high-dynamic-range imaging, motion detection and superresolution. For registering image pairs using homographies with 8 degrees of freedom (DOF), local-feature-based methods have proven flexible, robust and fast enough for aligning image pairs accurately at real-time (see section 9.2.4 in [157]). However, in some applications even faster methods are required to increase the performance, for instance, background-subtraction [188] typically merges sets of images containing approximately 10-20 video

frames. In order to avoid accumulation of pairwise homography errors, all images in the stack are aligned to the reference image for every analysed frame [86] requiring many registration operations. Therefore, image registration methods need to be much faster than the current state-of-the-art to allow such applications to run at real-time. In section 2.4 we present a new global registration method based on phase correlation, which can process images at up to 200 fps when using a hardware accelerated implementation. In chapter 6 we successfully apply this registration method to align video frames for moving object detection from aerial camera platforms.

The superresolution fusion presented in chapters 3 and 4 requires sub-pixel accurate geometric and sufficient photometric registration. Instead of the typical two-stage approach with a geometric registration followed by a photometric alignment, as described in sections 2.2.1 and 2.2.2, we employ a recently published approach [12], which computes both types of registration simultaneously. In applications in which images taken from very different view-points (i.e. images with a wide baseline) need to be registered, global dense methods produce unsatisfactory results. In section 2.3.1 we discuss feature-based registration, which is much better suited in these cases. In chapter 3 we show that this registration method can successfully register high-resolution images retrieved from the Internet with low-resolution video frames recorded with a mobile camera.

2.2 Existing Approaches

Most image fusion methods merge two or more images based on pixel information. Typically, the extracted pixel information from all images should correspond to the same scene location. If the camera does not move during the capture of multiple images or the scene has not changed, then a pixel at location (x, y) in one image corresponds to a pixel at location (x, y) in another one. However, in practical applications this assumption does not always hold. Often the camera is moving, even when using a tripod due to e.g. camera chassis vibrations, wind or parts of the scene are moving. In addition to the geometric misalignment, e.g. image shifted to the lower right corner as illustrated

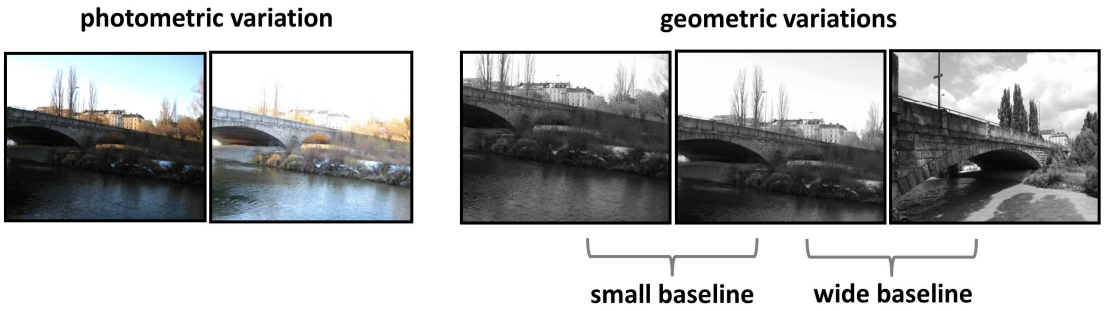


Fig. 2.1: Left: different camera parameters lead to photometric misregistration. Right: small camera motion (i.e. small baseline) or large camera motion (i.e. large baseline) lead to geometric misalignment.

in Fig. 2.1 (geometric variations), images may also exhibit a photometric misalignment, which are typically caused by varying camera parameters, e.g. auto exposure of the camera, as illustrated in Fig. 2.1 (photometric variation). The photometric and geometric variations discussed above can be eliminated with image registration methods.

All fusion methods, i.e. superresolution (chapter 3), high-dynamic-range imaging (chapter 4) and motion detection (see chapter 6), presented in this thesis require geometric registration as the input images are recorded from mobile camera platforms. Photometric registration is necessary for the reconstruction-based superresolution presented in chapter 3 as the underlying camera model assumes a photometric aligned input.

2.2.1 Photometric Registration

In this section we discuss methods for photometric registration. Variations in global image intensity and coloring originate from two main sources:

1. the global illumination of the scene changed
2. the camera parameters changed (e.g. the exposure time) altering the depiction of the scene

In most fusion applications images are both, geometrically and photometrically misaligned which leads to a "chicken-and-egg" problem. On the one hand, most approaches to photometric registration assume geometrically registered input images. On the other hand, many geometric registration problems are much easier to solve if the input images are photometrically aligned. To escape this dilemma there are two options. First, the geometric and photometric registration are performed independently from each other, where the geometric part is either robust against photometric misalignment or the photometric registration is invariant to geometric misalignment. Second, the geometric and photometric registration are performed simultaneously. In this section we briefly summarize existing methods that address the former option. In section 2.3.2 we address the latter option and motivate the use of this approach for the fusion methods presented in this thesis.

Photometric Registration Assuming Robust Geometric Alignment

Similar to the geometric registration, for which a motion model of the camera is assumed describing the geometric transformation between images, photometric registration requires a photometric model. The model must be capable of describing either the effect of changes in global illumination conditions or of altering camera parameters (e.g. exposure time). Given two images I_1 and I_2 that are photometrically misaligned, the goal is to find a mapping function τ that relates corresponding intensity values to each other

$$I_1(p(x, y)) = \tau(I_2(x, y))$$

Where p is a geometric warping function that maps a coordinate pair (x, y) from one image onto another and $I(x, y)$ denotes the intensity value at coordinates (x, y). Depending on the origin of the photometric misalignment the photometric warping function (i.e. photometric model) can take different forms for its parameterization. A very simple and generic form for parameterization, which does not relate to any physical meaning

in terms of the source of the photometric misalignment is an affine relation

$$I_1(p(x,y)) = \tau(I_2(x,y)) = \alpha \cdot I_2(x,y) + \beta \tag{2.1}$$

where α is often considered as a "contrast" value and β as a change in "brightness". For applications in which no additional information about the scene (e.g. how much light was emitted/reflected from the scene) or camera (e.g. which internal camera parameters were used) is available, photometric registration using the affine model is sufficient and possibly the only option. For instance Capel et al. [28] successfully applied this photometric model in recovering latent marks in forensic images. In order to estimate α and β multiple corresponding intensity values $I_1(p(x,y))$ and $I_2(x,y)$ were extracted after a geometric registration of the images. This assumed that the feature-based geometric registration was robust to the photometric difference between the images in question. With these corresponding intensity values a least-squares problem can be formulated, which computes α and β [28].

Robust Photometric Registration Invariant to Geometric Misalignment

If the source for the photometric misalignment is purely due to changes in camera parameters, then the mapping function τ can be parametrized more accurately using the camera response function f. The relation between the radiance E, the exposure time t and the intensity values I (Eq. 4.2) is defined as

$$I = f(E \cdot t)$$

Given two corresponding intensity values I_1, I_2 which were generate by two different exposure times t_1 and t_2

$$I_1 = f(E \cdot t_1) \text{ and } I_2 = f(E \cdot t_2)$$

these can be related via the scene radiance E, which is constant for both images

$$I_1 = f\left(\frac{t_1}{t_2} \cdot f^{-1}(I_2)\right)$$

Hence the intensity mapping function τ can be fully parametrized by the camera response curve and the exposure times t_1, t_2

$$\tau = f\left(\frac{t_1}{t_2} \cdot f^{-1}\right)$$

Since the camera response curve is estimated once for a camera, all subsequent photometric registration operations do not require any geometrically aligned images. The process of estimating the function f is described in section 4.2.1.

A different approach was taken by Grossberg et al. [64], who defined the intensity mapping τ solely using cumulative histograms H. This also assumes that the only source of photometric misalignment is due to camera parameter changes (e.g. exposure times). Consider two images I, their cumulative histograms H and the intensity mapping function τ between them

$$H_1(I_1) = H_1(\tau(I_2)) = H_2(I_2)$$

Substituting $I_2 = u$, the intensity mapping function can now be formulated in terms of these cumulative histograms

$$\tau(u) = H_1^{-1}(H_2(u))$$

where the inverse of the cumulative histograms are well defined as these are monotonic. The advantage of using histograms is the robustness to geometric misregistration as long as the histogram of scene radiance is constant for different exposures. The disadvantage is the sensitivity if a continuous range in radiance is not present in the scene. This leads to empty bins for which the inverse of the cumulative histograms is not defined. The interpolation is used to generate values for these empty bins. Intensity mapping functions based on the estimation of the camera response curve are usually more robust in these scenarios as they either employ an adequate model or smoothness priors.

2.2.2 Geometric Registration

The problem of geometric image registration has been studied for many decades and many methods are available. However, because of the large number of different ap-

proaches very often it is not clear which method should be used. Furthermore, geometric registration is needed in a wide range of applications each imposing different requirements. In mosaicing a scene is captured from different viewpoints and images overlap only by a small percentage of their actual size. In medical image analysis registration is often required to merge images taken with different sensors (e.g. X-Ray, CT, PET, MRI). Images taken at different points in time might be fused to visualize changes in the scene, e.g. satellite images capturing changes of the polar caps. In the latter class of applications the displacements are rather small, but the intensity values may vary for corresponding pixels. Some applications such as superresolution require rather accurate alignment, whereas others such as multi-modal sensor fusion can handle loose alignment. For real-time applications, e.g. on-board processing of aerial videos, the registration step must be efficient, whereas offline applications such as superresolution do not impose time restrictions. To cope with these different challenges a number of methods have been proposed [23, 193, 157].

The first distinction between registration algorithms can be made based on how pixel variations are introduced in two different images. In many cases the movement of the camera or zoom change results in misalignment between two images. This can be described in mathematical terms and it applies to all pixels. Any registration algorithm that builds on such an assumption, where a camera movement globally effects all pixels, is called a *global registration* method. Parts between two captured images may also change, even if the camera was not moved at all. This might be due to dynamic scene motion, e.g. a walking person. Such a complex movement can not be described mathematically for all pixels in a global form. Hence, each pixel has its own motion model from one image to another. Methods that estimate the motion of each pixel individually are called *local registration* methods. In Fig. 2.2 the different categories of local and global registration methods are illustrated. These coarse categories can be further subdivided based on how information about the camera motion is extracted from the images. Either all pixels are used, i.e. dense, or only a sparse subset of pixels are used. For global registration, the first type of methods are called *global dense registration*. These work either

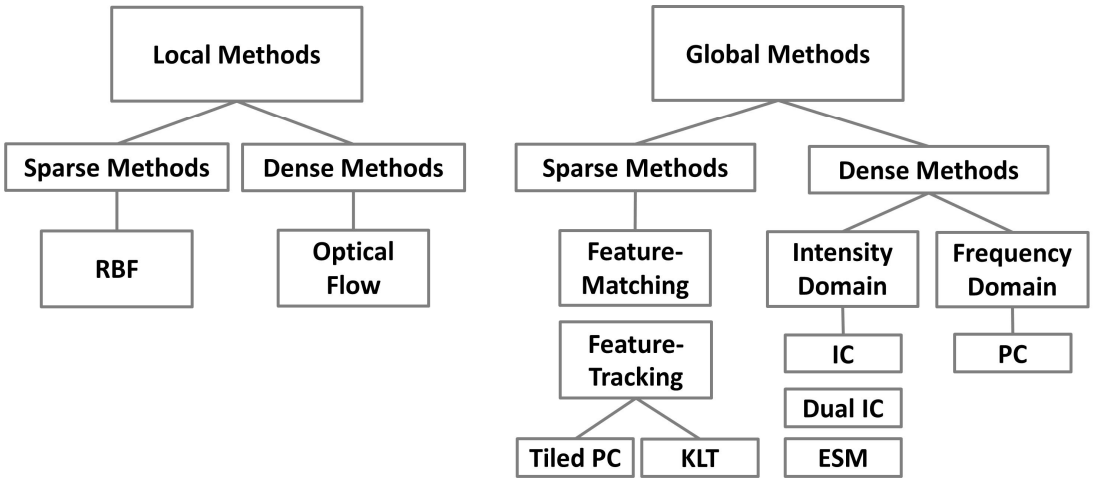


Fig. 2.2: Overview of different registration methods.

directly on intensity values (i.e. intensity-based methods often called *direct methods*) or they transform the images into the frequency domain. Frequency-based methods are very robust to noise artifacts, but are limited in the types of global image transformation that can be computed. In section 2.4 we introduce a new frequency-domain-based registration method called *tiled phase correlation* (Tiled PC), which allows to compute 8-DOF homographies. The most well-known global, dense algorithms are the Inverse-Compositional (IC) [10], the Dual-IC [12] and the Efficient Second-Order Minimization (ESM) [16], which are computed in the intensity domain. The Phase-Correlation (PC) [93] is a well-known algorithm operating in the frequency domain and certainly one of the oldest registration method. The second type of methods are called *global sparse registration* methods, where sparse locations are either defined by feature detectors or extracted from a defined grid. The Kanade-Lucas-Tomasi (KLT) feature tracker [148] and feature matching, e.g. using SIFT [105], are very popular methods from this category.

Well-known local dense registration approaches are based on optical flow, which computes translational offsets between two images on a dense grid. A good general overview of these methods can be found in [140]. In [7, 192, 57] these methods were applied to image fusion. For reducing the computational costs, the offsets can be measured on a

sparse grid and interpolated over a fine grid, e.g. using radial basis functions (RBF) [11].

Although generally in real applications the assumption of a global transform for registering images is too restrictive, many image fusion algorithms rely on such a simplification to ensure accurate registration. This is because the global transform is computed with the support of many measurements, densely or sparsely. Local registration does not have the global support leading to many outliers. Therefore, we also focus in this thesis on registering images using a global transform, i.e. a 8-DOF homography. In the following sections 2.3.1 and 2.3.2 we discuss the two existing homography-based registration algorithms used in this thesis.

2.3 Efficient Registration

In this section we discuss in detail the registration approaches related to the work in this thesis and used in the considered fusion applications. We categorise them into feature-based and intensity-based methods reviewed in section 2.3.1 and section 2.3.2 respectively. We then review their efficient implementations in section 2.3.3.

2.3.1 Feature-Based Methods

Feature-based registration can be divided into three steps: *detection*, *matching* and *estimation of transformation*. This type of registration is very well suited for scenarios with large displacements between images. This is often referred to as *wide-baseline registration*. This capability is due to robust matching algorithms. These rely either on descriptors computed for feature points [105], result from stable tracking of feature points [148] or build upon robust feature point recognition [98, 25, 120]. Furthermore, these methods are robust to illumination changes, since the detection is designed to be photometrically invariant. However, reliable and accurate localization of the feature points is both the strength and weakness of these approaches. If precise features can

be found these methods show great robustness to large displacements. However, false matching or errors in the location of correspondences amplify errors in the transformation parameters. Fortunately, tools like outlier detection (e.g. RANSAC [55]) exist to greatly improve the performance of feature-based methods. In the following a basic description of each of the three steps is given.

Detection

Features are characteristic points in the image, which are invariant to transformations, i.e. often affine ones such as translation, scale, rotation and to illumination changes. Different feature detectors exist, which select distinctive regions, that are repeatable across different images showing the same scene [168]. They differ in terms of robustness, repeatability and computational complexity.

Matching

Finding correspondences is one of the most computational time consuming tasks of feature-based matching, especially with large numbers of features. In essence, every feature $\vec{f_1}$ needs to be identified via a so called *feature descriptor* $\vec{d_1}$ and compared to every other descriptor $\vec{d_2}$ using a distance measure $\langle \vec{d_1}, \vec{d_2} \rangle$. False matches, so called *outliers*, of not corresponding features need to be eliminated to improve the estimation of transformation parameters. A possible approach to outlier rejection is a thresholding based on a ratio of descriptor distances to the first and the second nearest neighbor. Other matching strategies than this nearest neighbour distance ratio (NNDR) have shown lower performance [113]. Various distance measures can be used, e.g. sum-of-squared differences or angle between the feature vectors. An exhausting evaluation of all possible matches using the nearest neighbour distance ratio (NNDR) has a run-time, which is quadratic with the number of features, limiting the range of applications for this approach. Therefore, various indexing structures (e.g. kd-trees) and search strategies (e.g. best-bin-first) have been proposed [116], some of which do not find the closest neighbours needed for

computing the NNDR criteria, but a very close approximation of it. To further speed up the matching a geometric neighbourhood constraint can be introduced for each feature $\vec{f}$. Clearly, for optimal settings, the expected displacements between the images have to be known.

Simple Estimation of Transformation

Once matching features have been found, the transformation can be estimated. Depending on the application, different types of transformations are of interest. For many computer vision problems (e.g. mosaicing, superresolution) a global homography, which has 8 degrees of freedom and assumes that matching features are located on a planar surface, is often sufficient [28, 157]. A homography H is a 3×3 matrix which transforms a feature point $\vec{x}_i$ onto another $\vec{x}'_i$

$$\vec{x}'_i = \alpha \cdot H \cdot \vec{x}_i \tag{2.2}$$

The feature points $\vec{x}_i$ and $\vec{x}'_i$ are homogeneous vectors, hence the mapping through the matrix H is only defined up to an unknown scalar α. Given a set of feature matches the parameters of a homography can be either directly estimated, i.e. via the direct linear transform algorithm, or in an iterative manner using different kind of cost-functions and optimization schemes.

The direct linear transform (DLT) is derived by formulating the homography constraint [70] as a cross-product

$$\vec{x}'_i \times H \cdot \vec{x}_i = 0 \tag{2.3}$$

This eliminates the unknown scaling factor α and simply requires that the homography transformation of the feature point is parallel to the corresponding match. The cross-product can be written using a skew-matrix

$$\vec{x}'_i \times H \cdot \vec{x}_i = A_i \cdot \vec{h} = \begin{bmatrix} \vec{0}^T & -x'_{i,3}\vec{x}_i^T & x'_{i,2}\vec{x}_i^T \\ x'_{i,3}\vec{x}_i^T & \vec{0}^T & -x'_{i,1}\vec{x}_i^T \\ -x'_{i,2}\vec{x}_i^T & x'_{i,1}\vec{x}_i^T & \vec{0}^T \end{bmatrix} \cdot \vec{h} = 0 \tag{2.4}$$

where $\vec{v}_i = (v_{i,1}, v_{i,2}, v_{i,3})^T$ and $\vec{h}$ represents the row-major linearized matrix H. The third row of matrix A_i depends linearly on the first two and can be neglected. The remaining two rows represent two linear equations for the unknown variables in $\vec{h}$. Hence using four feature matches $(\vec{x}_i, \vec{x}'_i)$ with $i \in [1, 4]$ all 8 parameters of the homography are defined and can be estimated in a closed form. Combining all sub-matrices A_i resulting from each feature match into one large matrix A leads to the final linear equation set

$$A \cdot \vec{h} = 0 \qquad \text{subject to} \quad \|\vec{h}\| = 1 \tag{2.5}$$

If more than four matches are used, the equation system is over-determined and does not have an exact solution most of the time. In this case a matrix $\vec{h}$ is sought, which minimizes a residual error

$$A_{over} \cdot \vec{h} = \vec{e} \tag{2.6}$$

In other words a minimization procedure finds the matrix $\vec{h}$, which minimizes the following algebraic cost function

$$O_{algebraic} = \|\vec{e}\|^2 = \|A_{over} \cdot \vec{h}\|^2 = \sum_i \left(\begin{pmatrix} 0 \\ -x'_{i,3} \\ x'_{i,2} \end{pmatrix}^T \vec{\hat{x}}_i \right)^2 + \left(\begin{pmatrix} x'_{i,3} \\ 0 \\ -x'_{i,1} \end{pmatrix}^T \vec{\hat{x}}_i \right)^2 \tag{2.7}$$

where $\vec{\hat{x}}_i = H \cdot \vec{x}_i$. This direct approach (DLT) can be computed very quickly and with only 4 matches an exact solution can be obtained. Unfortunately, the resulting homographies $\vec{h}$, which minimize this cost function $O_{algebraic}$, are often not accurate enough to register the two involved images, especially when many but noisy feature matches are used. These are the optimal solution in respect to the specified cost function, but since the cost function $O_{algebraic}$ does not explicitly reflect the property of aligning the two images in a geometrical sense, the "algebraically optimal" solution is not the "geometrically optimal" solution for the registration problem. Another reason for possible failure cases is the absence of an error model for the feature matches. Mismatching features are treated the same as correctly matching features with a localization error.

In other words, the DLT assumes perfectly located feature matches, which do not exist in most applications. To overcome these problems an alternative cost function needs to be specified, which does reflect the geometrical alignment of the involved images.

Robust Estimation of Transformation

In order to allow for noisy feature locations, an error model is needed. Assume an isotropic, normal distribution with zero mean and standard deviation σ of the localization error of a detected feature point. Then the probability for a detected feature $\vec{x}_i$, given its true location $\vec{\bar{x}}_i$, is specified by

$$P(\vec{x}_i, \vec{\bar{x}}_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|\vec{x}_i - \vec{\bar{x}}_i\|^2}{2\sigma^2}\right) \tag{2.8}$$

Using this error model, the homography estimation can now be formulated as a probability maximization problem. The homography H and the true unknown feature locations $\vec{\bar{x}}_i$ are sought which maximize the following probability

$$P(M|H, \vec{\bar{x}}_i) = \prod_i P(\vec{x}_i, \vec{\bar{x}}_i) \cdot P(\vec{x}'_i, H \cdot \vec{\bar{x}}_i) \tag{2.9}$$

where M is the set of all matched features $\{\vec{x}_i, \vec{x}'_i\}$. This maximization problem is converted into a minimization problem by considering the negative log-likelihood leading to the following geometric cost function O_{mle}

$$O_{mle} = \sum_i \|\vec{x}_i - \vec{\bar{x}}_i\|^2 + \|\vec{x}'_i - H \cdot \vec{\bar{x}}_i\|^2 \tag{2.10}$$

which is minimized by the optimal homography H_{mle} (maximum likelihood estimate) and the true unknown feature locations $\vec{\bar{x}}_i$. To solve this non-linear minimization problem the Levenberg-Marquardt algorithm [112] is often used.

Because this cost-function incorporates an explicit noise model for incorrect feature locations and it reflects the geometrical alignment of the two involved images, the solution H_{mle} is considered as the geometrically optimal transformation to align the two images.

However, since the optimization needs to be solved in an iterative manner, the computation usually takes much longer and requires an acceptable initial estimate. That is why this method is usually only used to refine a rough estimate, which is computed using the direct approach (DLT).

The quality of the results obtained via DLT depends on the data (i.e. the feature coordinates) being used [70]. Therefore the features points need to be normalized prior to the computation. For each set of features points (X containing the features from the first image and X' containing the features from the second image) normalization transformations (T for all features X and T' for all features in X') are computed, which translate and scale the features in such a way, that their mean is centered at the zero-origin and the average distance to the mean is $\sqrt{2}$. The DLT is then applied to the normalized feature matches $(\tilde{x}_i, \tilde{x}'_i)$ where $\tilde{x}_i = T \cdot \vec{x}_i$ and $\tilde{x}'_i = T' \cdot \vec{x}'_i$. The resulting homography $\tilde{H}$ is then denormalized to obtain the final result

$$H_{DLT} = T'^{-1} \cdot \tilde{H} \cdot T \tag{2.11}$$

Outlier Detection

The geometrically optimal solution H_{mle} accounts for Gaussian noise in the feature locations. However, very often the matches are either incorrect or the assumption that all feature matches lay on the same plane is violated, e.g. by parallax effects. These matching errors cannot be easily modeled in terms of a cost function. Therefore, robust estimation techniques have been proposed which try to find the correct homography H_{robust} using a set of matches which include false or inconsistent matches, so called *outliers*. The most popular methods are based on a sampling strategy. A certain number N of submatches $m = \{\vec{x}_i, \vec{x}'_i\}$ are randomly selected, where each subset contains the minimal number of matches needed to compute a homography (i.e. 4 matches). The intuition behind this strategy is that the probability of selecting a subset, which contains only correct matches (i.e. *inliers*), is high if the fraction of inliers is larger than the number of outliers in the whole set. The homography computed from such a perfect subset would be

also valid for all other correct matches. In other words, many homographies H_{trials} are computed from randomly selected 4 matches. The homography, which receives the best support from all other matches, is selected as the true underlying homography H_{inlier}. Different methods have been proposed on how to compute the support of a homography. The most often applied ones are the Random Sampling Consensus (RANSAC) [55] and the Least Median of Squares (LMedS) [70].

RANSAC counts the number of inliers for each H_{trials} by computing a reprojection error

$$err = \|\vec{x}_i' - H_{trials} \cdot \vec{x}_i\| \tag{2.12}$$

for each match $(\vec{x}_i, \vec{x}_i')$. If the error is below some threshold, called *inlier threshold*, the match is considered as an inlier that supports the current homography H_{trials}. The homography H_{trials}, which produces the greatest amount of inliers, is chosen as the final robust homography H_{inlier}.

LMedS computes the k^{th} median reprojection error among all other matches than the ones from the current subset. The homography H_{trials} that generates the smallest median reprojection error is the final robust homography H_{inlier}. All matches that have an error smaller than the minimum reprojection error are considered as inliers. The advantage of this approach is that no inlier threshold has to be selected as for RANSAC. However, the approximate fraction of expected outliers has to been known, which is only the case for well studied image material.

Because of the iterative random sampling of minimal subsets, these robust estimation schemes can be quite time consuming depending on the fraction of inliers. The complete automatic robust homography estimation [70] is summarized in Algorithm 1.

2.3.2 Intensity-Based Methods

It is also possible to use raw intensity values for geometric registration, which are usually much denser than sparse local feature points. The underlying core of these methods is the optimization of a cost measure such as sum-of-squared-differences (SSD) [76], mutual

Algorithm 1 Gold Standard Homography Estimation.

1. compute features at distinctive locations
2. match features using descriptor and efficient matching strategy
3. robust estimation of inliers and their best homography H_{inlier}:
   (a) select random sample of minimum number of correspondences (4 matches)
   (b) compute homography H_{trials} using normalized DLT
   (c) compute number of inliers (i.e. estimate support) using RANSAC or LMedS
4. compute H_{mle} using only inlier matches and with H_{inlier} as initialization using the non-linear homography estimation (MLE)
5. use H_{mle} to refine matching and list of inliers and re-estimate H_{mle} (repeat until number of inliers converge)

information (MI) [172] or cross-correlation [17], which are based on intensity values. These methods show weaknesses when encountering large displacements between the images to be aligned. To eliminate this disadvantage and to speed up the computation often a coarse-to-fine strategy is applied [162]. However, in cases of less structured scenes an optimal transformation might still be found, where the problem of finding enough features impedes the use of feature-based approaches as discussed in section 2.3.1. By far the most popular intensity-based methods are based on minimizing the sum-of-squared differences (SSD) between two images. Since the well-known *Lucas-Kanade tracker* [106] many variants based on this idea have been proposed [9]. A recent method showing superior convergence rates, robustness and accuracy is called *efficient second-order minimization* (ESM) [108, 16].

Dual Inverse Compositional

When computing registration parameters for geometric and photometric alignment separately, one has to either rely on the robustness of the geometric registration towards photometric variations (e.g. [28, 177]) or on a photometric registration method, which does not require geometrically aligned images (e.g. [64]) as discussed in section 2.2.1. However, for many applications it is hard to predict whether the photometric variations between two images is small enough that it can be handled by a photometrically robust geometric registration method and vice-versa. Instead of treating the two registration problems independently of each other, a more elegant approach is to solve these simultaneously.

Recently such a direct registration method was proposed by A. Bartoli [12], which is called *dual inverse compositional*. The work extends the *inverse compositional* approach of Baker et al. [10] to photometric warps. Both, the parameters for the geometric and photometric warp are then solved simultaneously. The photometric warp is formulated only in terms of affine parameters. The author shows in experiments that the simultaneous approach is more robust and efficient than performing geometric registration alone.

Given two geometrically transformed images T and I, the direct intensity approach aims to solve the following cost-function

$$err = \sum_{\vec{x}} \left[T(\vec{x}) - I\left(W(\vec{x}, \vec{p})\right) \right]^2 \tag{2.13}$$

where $W(\vec{x}, \vec{p})$ is a warping function, which projects the coordinates $\vec{x}$ into the frame of the template image T using the parameters $\vec{p}$ of the underlying motion model (e.g. a homography). The goal is to find the unknown parameter set $\vec{p}$, which perfectly aligns the input image I with the template image T in terms of the intensity error. Because of the non-linearity nature of this cost-function, this is usually solved in a two-step iterative manner. The first step solves for a small update of the current parameter vector $\Delta\vec{p}$, which is initialized with $\vec{p} = 0$, and then updates $\vec{p}_{new} \leftarrow (\Delta\vec{p}, \vec{p}_{old})$. Different strategies

have been proposed on how to include the parameter update $\Delta\vec{p}$ in the cost-function Eq. 2.13 and how to update $\vec{p}_{new}$ [10]. The inverse compositional is one of these method that has proven to be very efficient as most time-consuming steps can be precomputed.

The standard inverse compositional method aims to minimize the following cost-function [10] at each iteration

$$err = \sum_{\vec{x}} \left[T\left(W(\vec{x}, \Delta\vec{p})\right) - I\left(W(\vec{x}, \vec{p})\right)\right]^2 \tag{2.14}$$

where the parameter vector $\vec{p}$ for the next iteration is updated as

$$W(\vec{x}, \vec{p}) \leftarrow W(\vec{x}, \vec{p}) \circ W(\vec{x}, \Delta\vec{p})^{-1}$$

The dual inverse compositional [12] extends this efficient scheme to also include a photometric warping function, which leads to the following cost-function at each iteration

$$err = \sum_{\vec{x}} \left[V\left(T\left(W(\vec{x}, \Delta\vec{p})\right), \Delta\vec{q}\right) - V\left(I\left(W(\vec{x}, \vec{p})\right), \vec{q}\right)\right]^2 \tag{2.15}$$

where $V(a, \vec{q})$ is a photometric warping function, for instance an affine model as presented in Eq. 2.1.

Many alternative approaches combining photometric and geometric registration have been proposed. In the following a few notable ones are highlighted. In [91, 190] the KLT cost-function based on color constancy was extended to include a simple multiplicative factor (i.e. gain) on the intensity values to make the KLT feature tracking more robust to the auto-exposure control of the camera. This proved very useful for outdoor applications. A similar approach was also proposed in [83], where an affine model (i.e. gain and bias) was used for photometric registration. A more complex photometric model, i.e. a spatially varying affine model approximated by low-order polynomials, was combined with a geometric registration in [95]. Very recently in [107] a similar approach to [12] was presented, which further relaxes assumptions on noise constraints.

2.3.3 Implementations Using Parallelization

Current local-feature-based methods are by far the most popular ones for image registration [157]. All of them share a similar processing chain as discussed in section 2.3.1 and summarized in the following. First, sparse local features are detected in the two images to be registered. Second, corresponding features are matched using descriptors computed at the feature locations. Third, a transformation (i.e. 8-DOF homography) is computed. The most time consuming parts are the detection of feature locations and the computation of the descriptors. Although very efficient algorithms exist [13, 138] which already allow real-time processing for small image sizes, some applications require even faster algorithms. With the increasing ease of programming field programmable gate arrays (FPGA) and graphic cards (GPU) a common way to speed up such registration algorithms is to implement them on such parallel platforms. However, not all algorithms including efficient ones are easily parallelizable and hence benefit less from an implementation on a FPGA or GPU. In [149] SIFT and KLT features achieved speed-up factors of 10 and 20 respectively compared to CPU implementations. FPGA implementations of SIFT features show similar speed-up factors around 10 as presented in [145]. In [33] the efficient SURF features have shown to speed up by a factor of around 26 compared to the original CPU implementation by [13]. Because many local features (e.g. around 1000) are typically computed for robust homography estimation and each one can be computed independently, feature detection is an inherent parallel problem and benefits from GPU and FPGA implementations as shown by the work referenced above. However, depending on the complexity of the feature, different speed-up factors can be achieved, e.g. simple KLT features (speed-up by 20) vs. complex SIFT features (speed-up by 10). The feature matching is also parallel in nature and speed-up factors between 10-30 have been achieved, when implementing it using GPUs [29]. However, even with those speed-up factors, the overall frame rates achieved for typical videos, i.e. 640×480, converge to 100 fps [33]. Although direct image registration methods, which use all pixels of the image for alignment, are much slower on CPUs than feature-based

methods, their GPU-implementations such as the one presented in [79] achieve similar frame rates. Registering 10 images per processed frame is therefore bound to around 10 fps. To achieve faster processing or to allow the registration of more images (e.g. 20 images to compute a background image), a faster registration method is needed.

2.4 Tiled Phase Correlation

In this section we present an efficient registration approach, which is based on phase correlation. The core components of phase correlation are Fourier transforms, which are highly parallelizable [60]. This results in very fast processing (i.e. up to 200 fps), when implementing this registration method on a modern GPU. Other advantages of phase correlation are its simplicity and robustness. It can handle narrow and larger baselines, given sufficient image overlap, at the same computational cost, it is robust to noise, has the ability to register less textured images and requires considerably less parameter tuning than feature-based and intensity-based methods.

In section 2.4.1 we briefly introduce the phase correlation and explain how it benefits from parallel implementations. In section 2.4.2 we present our extension of the phase correlation, which allows the computation of 8-DOF homographies. In section 2.4.3 we discuss the homography estimation. In addition to the previously discussed standard MLE-based approach, we include a non-uniform location uncertainty directly into the MLE-based optimization scheme. Finally, in section 2.4.4 we present the evaluation results, which show high registration accuracy and demonstrate large speed-up factors between CPU vs. GPU implementations.

2.4.1 Phase Correlation

Registration methods based on the Fourier transform have been studied since [93] and many extensions and applications have been proposed to date [134, 4]. In Fig. 2.3 the computation scheme of standard phase correlation is illustrated. Consider two images

I_1, I_2 that are related by simple translation $\Delta\vec{x}$

$$I_1(\vec{x}) = I_2(\vec{x} - \Delta\vec{x}) \tag{2.16}$$

According to the Fourier shift theorem, their Fourier transforms $\hat{I}_1, \hat{I}_2$ are related such that

$$\hat{I}_2(\vec{u}) = \exp\left[-2\pi i(u_1\Delta x_1 + u_2\Delta x_2)\right]\hat{I}_1(\vec{u}) \tag{2.17}$$

To find the shift in phase, the correlation of them is computed

$$\frac{\hat{I}_1(\vec{u})\hat{I}_2^*(\vec{u})}{\|\hat{I}_1(\vec{u})\hat{I}_2^*(\vec{u})\|} = \exp\left[2\pi i(u_1\Delta x_1 + u_2\Delta x_2)\right] \tag{2.18}$$

where $\hat{I}_2^*$ is the complex conjugate. The term is also called *cross-power spectrum* of I_1 and I_2. Its phase is equivalent to the phase shift between both images. Hence, transforming the cross-power spectrum back into spatial domain, results in a Dirac impulse at the position of the displacement. The sub-pixel location is estimated by fitting a quadratic to a data triplet around the maximal peak at location x and y respectively. The refinement offsets Δx and Δy are given by

$$\Delta x = \frac{\log(v(x+1,y)) - \log(v(x-1,y))}{4\log(v(x,y)) - 2\log(v(x-1,y)) - 2\log(v(x+1,y))} \tag{2.19}$$

and analogous for Δy where $v(x,y)$ denotes the value of the cross correlation peak in the spatial domain at location (x,y). Strong signal changes at image boundaries can

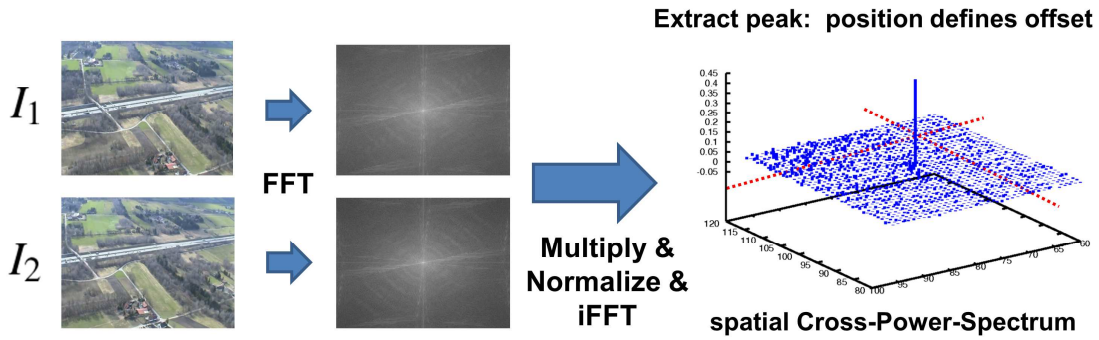


Fig. 2.3: Phase correlation scheme.

significantly affect the correlation surface. Therefore, peaks are well enhanced if the

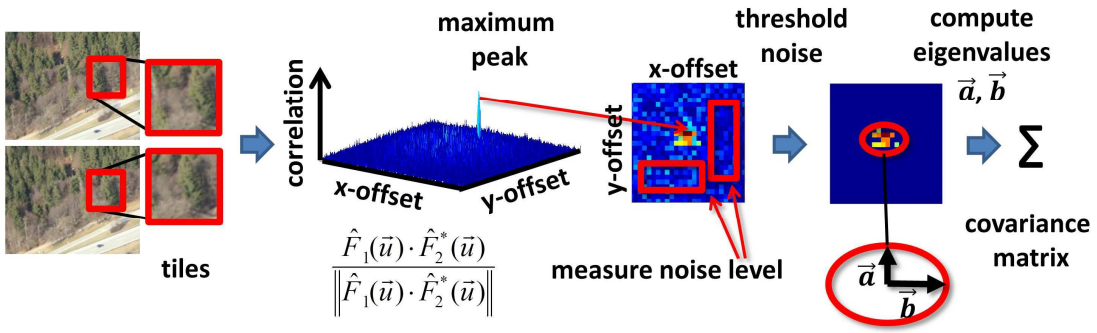


Fig. 2.4: Extraction of covariance matrix from phase correlation peak calculated from two tiles.

image is filtered prior to applying phase correlation. This is typically handled by using windowing methods, i.e. by multiplying the image with a Blackman window function. In theory different types of windowing functions such as Cosine, Gaussian and Tukey can be applied. In section 2.4.4 we evaluate their impact on the registration accuracy.

Similar to the Fourier shift theorem, rotation and scale can also be analyzed in the frequency domain [134]. In practice these extensions are very sensitive to registration errors as these accumulate and amplify during the processing chain, i.e. from the transformation to the estimation of the various parameters. Furthermore, in most applications mobile cameras introduce also perspective transformations in addition to translation, rotation and scale. Therefore, a full 8-DOF homography needs to be estimated in order to provide accurate registration. In the following section we present such an approach which extends the standard phase correlation to these transformation parameters.

2.4.2 Tiling

To overcome the drawbacks associated with global image registration such as sensitivity to local occlusions and its limitation to 2-DOF homographies, we propose a tiled phase correlation. In our approach multiple regions, which we refer to as *tiles*, are extracted from a uniform grid in the image as illustrated in Fig. 2.4. A translational offset is then computed using the peak of the phase correlation [4] between each pair of tiles at

corresponding locations in consecutive video frames. We employ Blackman windowing prior to the phase correlation. The peak itself is extracted at subpixel accuracy by fitting a quadratic function to the global maximum of the spatial cross-power-spectrum. Moreover, we propose to derive an uncertainty measure, i.e. a covariance, by computing the eigenvalues of the ellipsoid footprint of the thresholded correlation peak. The threshold is estimated from the noise level in other parts of the correlation surface. The tile centers from the first frame are then placed into the second frame using the offsets, which form the matching point pairs for homography estimation as described in section 2.4.3. We show in section 2.4.4 that the choice of the tile layout is crucial for accurate and robust estimation. Intuitively, using large but few tiles, e.g. 4, would ensure a fast and robust matching but increases the registration error when the assumption that each tile is only translated from frame to frame is violated. Using small but many tiles with overlaps may result in slower but more accurate registration as the assumption is more likely to be valid. However, it increases the probability of a mismatch as only little image information is contained in each tile. We propose to distribute the tiles in 13 different layouts. For most layouts, i.e. with 4, 8, 9, 16, 25, 36, 64, 100, 225, 400 and 625 tiles, a uniform grid is used. The 5-tile layout is a variation of the uniform 4-tile layout with an additional center tile. The 25+-tile layout is a variation of the uniform 16-tile layout with an additional 9-tile layout as a second layer. Tiles overlap by roughly 25%, except for the simple 4-tile layout. In section 2.4.4 we provide a quantitative evaluation of the registration accuracy for different tile layouts.

2.4.3 Homography Estimation

Homography matrix P is estimated from the matching point pairs and their associated uncertainty covariances. As the true locations of the matched points are not known due to motion blur, noise, etc. the geometrically optimal way to compute P is via an iterative optimization scheme which minimizes the following maximum-likelihood (MLE)

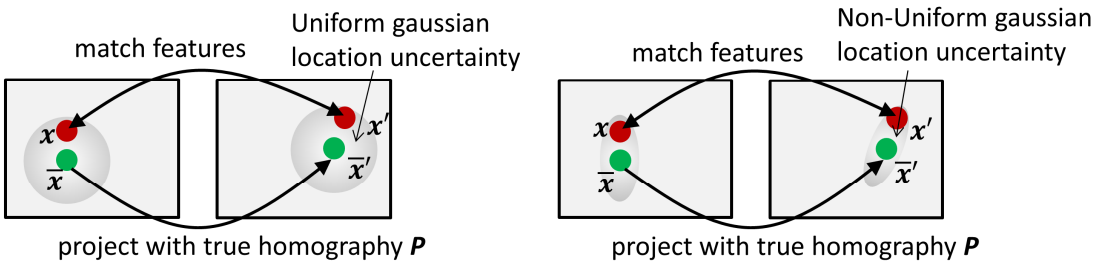


Fig. 2.5: Illustration of matching two points at estimated (red) and true (green) location using homography with MLE cost function and uniform (left) as well as non-uniform (right) Gaussian location error.

cost-function [70] as discussed in section 2.3.1. Recall the cost-function from Eq. 2.10

$$O_{mle} = \sum_i \left\|\vec{x}_i - \vec{\bar{x}}_i\right\|^2 + \left\|\vec{x}'_i - P \cdot \vec{\bar{x}}_i\right\|^2$$

This cost function accommodates the errors in the location of the matched points $\vec{x}_i$ and $\vec{x}'_i$, which lay close to their true locations $\vec{\bar{x}}_i$ and $\vec{\bar{x}}'_i$, as illustrated in Fig. 2.5 (left). The iterative MLE simultaneously computes the true homography as well as the true locations of matched points.

Typically uniform Gaussian is assumed as the point location error. If more information is available it can be integrated allowing for more accurate estimation, as illustrated in Fig. 2.5 (right). As discussed above we model the shape of the correlation peak by a covariance matrix Σ, which describes the direction and magnitude of the two main axes of the elliptic shape of the location error (see right of Fig. 2.4). The error has then the non-uniform distribution

$$P(\vec{x}_i, \vec{\bar{x}}_i) = \frac{1}{\sqrt{2\pi}\,|\Sigma|^{\frac{1}{2}}} \cdot e^{\left(-(\vec{x}_i - \vec{\bar{x}}_i)^T \Sigma^{-1} (\vec{x}_i - \vec{\bar{x}}_i)\right)} \tag{2.20}$$

Substituting the standard univariate Gaussian error function in the MLE cost-function with the multivariate one leads to the following cost-function

$$O_{mle-multivariate} = \sum_i (\vec{x}_i - \vec{\bar{x}}_i)^T \Sigma_i^{-1} (\vec{x}_i - \vec{\bar{x}}_i) + (\vec{x}'_i - P \cdot \vec{\bar{x}}_i)^T \Sigma_i^{-1} (\vec{x}'_i - P \cdot \vec{\bar{x}}_i) \tag{2.21}$$

which can be written as a standard weighted least-squares cost-function

$$O_{mle-multivariate} = \sum_i \vec{a}_i^T \cdot C \cdot \vec{a}_i \tag{2.22}$$

where

$$\vec{a}_i = \begin{pmatrix} \vec{x}_i - \vec{\bar{x}}_i \\ \vec{x}'_i - P \cdot \vec{\bar{x}}_i \end{pmatrix} \tag{2.23}$$

and

$$C = \begin{pmatrix} \Sigma_i^{-1} & 0 \\ 0 & \Sigma_i'^{-1} \end{pmatrix} \tag{2.24}$$

The covariance matrix C can be decomposed using upper-triangle cholesky decomposition

$$C = R^T R \tag{2.25}$$

leading to

$$O_{mle-multivariate} = \sum_i \vec{a}_i^T \cdot R^T R \cdot \vec{a}_i \tag{2.26}$$

which can be rewritten as

$$O_{mle-multivariate} = \sum_i (R\vec{a}_i)^T \cdot R\vec{a}_i \tag{2.27}$$

which can be treated and solved as a standard least-squares problem

$$O_{mle-multivariate} = \sum_i \vec{b}_i^T \cdot \vec{b}_i \tag{2.28}$$

Formally, the uncertainty in the *i-th* point pair match is thus captured by the covariance matrix Σ_i. The numerical effect of the covariances in the least-squares formulation is that the point matches with high uncertainty get a lower weight than the other ones. Hence, only the relative values of the covariances matter rather than the absolute ones.

It has been shown that this improves the performance [70, 191] for feature-based registration methods and we make a similar observation for our phase correlation based approach. An exemplary situation for such non-uniform Gaussian location error is motion blur, which blurs the image in a particular direction and confuses feature detectors.

Another example are feature detectors that also respond to edges [168]. Hence, the likelihood of misplacing them is along the edge, but not along a direction perpendicular to it.

To further reduce the influence of mismatches in the overall homography estimation, we employ *RANSAC* for robust outlier detection. As already a small camera motion can cause large translation of consecutive frames, we also apply a global 2-DOF pre-registration using phase correlation computed on the whole image.

Fig. 2.6: Dataset A used for evaluation of computational speed. Left two: image size 640×480px. Right two: image size 720×576px.

2.4.4 Results

Computational Speed

In the following we present a benchmark demonstrating the speed-up of the phase correlation by using a GPU-implementation instead of a CPU one. In order to allow a fair comparison with the results for local feature-based methods reported in the literature, we use the registration scheme (see section 2.4.2) which outputs 8-DOF homographies. We benchmark the implementations using 4 different systems with increasingly powerful CPUs and GPUs (see Tab. 2.1). The CPU implementation is based on the FFTW library [59] whereas the GPU implementation employs the CUFFT library [34] from the CUDA framework. In the latter all steps (i.e. extraction of regions along grid, forward/backward FFTs, computing the pixel-wise correlation and fitting of the quadratic

to detect the offset at sub-pixel level) except the DLT are carried out on the GPU minimizing expensive data transfers between GPU and CPU memory.

System 1 –	**slow GPU and medium CPU**
GPU:	NVIDIA GeForce 8600M GT, 128MB RAM, 32 Cores, 475 MHz
CPU:	Intel Core2 Duo 2.2GHz, 2GB RAM, 2 Cores
System 2 –	**medium GPU and slow CPU**
GPU:	NVIDIA Quadro FX 4600, 768MB RAM, 128 Cores, 500 MHz
CPU:	Intel Core2 Duo 1.86GHz, 2GB RAM, 2 Cores
System 3 –	**fast GPU and fast CPU**
GPU:	NVIDIA Quadro FX 5800, 2GB RAM, 240 Cores, 650 MHz
CPU:	Intel Xeon 3.2GHz, 12GB RAM, 4 Cores
System 4 –	**fast GPU and fast CPU**
GPU:	NVIDIA Tesla C1060, 4GB RAM, 240 Cores, 1.3 Ghz
CPU:	Intel Xeon 3.2GHz, 12GB RAM, 4 Cores

Table 2.1: CPU and GPU configurations used for benchmarking.

Although the registration algorithm is independent of the image content it does depend on the image size. We therefore use two typical image sizes that are produced by aerial cameras: 640×480 and 720×576. To demonstrate the generalization of the results we use two different image sets per size configuration, resulting in 4 sequences with aerial scenery. This *dataset A* is depicted in Fig. 2.6. The results of the benchmark are presented for each system in Fig. 2.7. The run-times were computed by averaging the processing times over 10 runs per sequence. These average times were further summarized by averaging across all 4 sequences of the dataset. The final average run-times

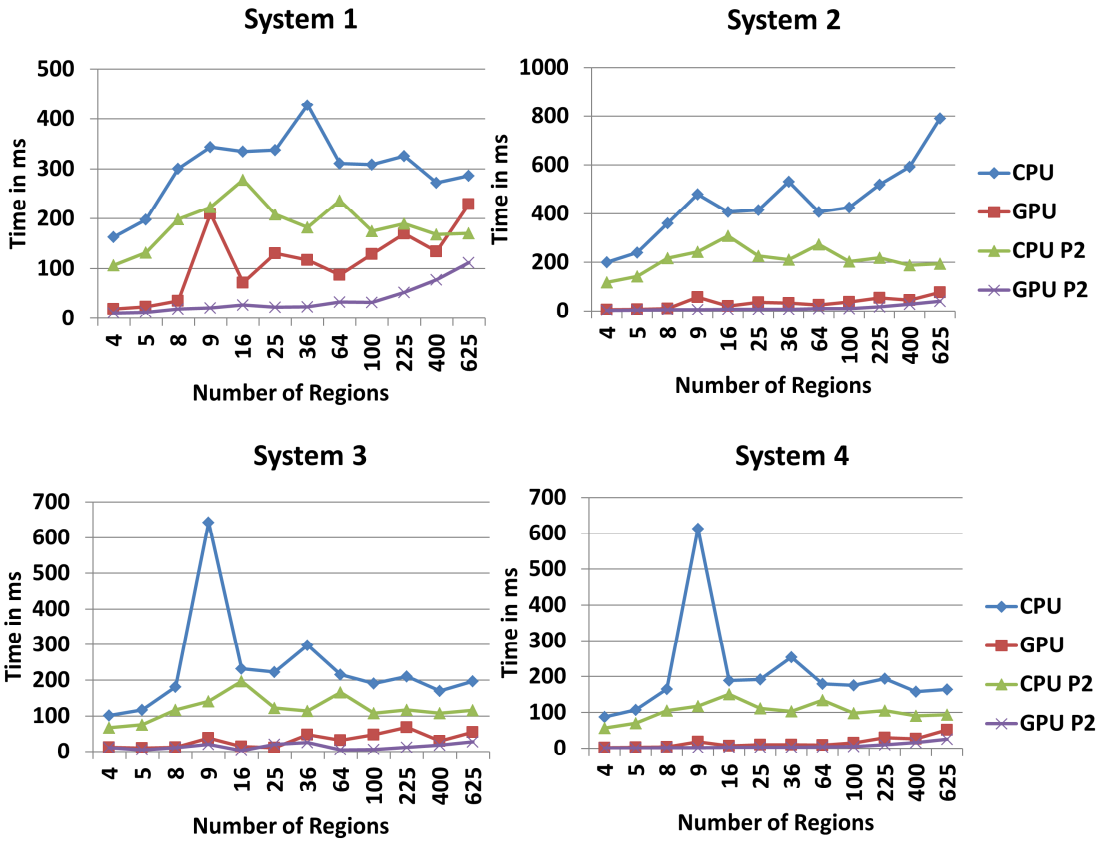


Fig. 2.7: Timing results for system 1-4.

in milli-seconds are then plotted for each grid-layout. The speed-up factors from CPU to GPU achieved are around 3 for layouts with many phase correlation operations (e.g. 625) and up to 30 for smaller layouts (e.g. 4). As FFTs greatly benefit from input sizes of power-of-two, the same images were padded with zeros to the nearest power-of-two dimension. The run-times were computed in the same way as before and are referenced with "P2" in the figures. It can be seen that the GPU implementations in all systems are much faster with this padding. Although slow GPUs can not keep up with the 100 fps as reported in [33], due to the much weaker hardware (our GPU NVIDIA Geforce 8600M GT has 32 stream processors with 475 MHz, whereas [33] use a NVIDIA Geforce 8800 GTX with 128 stream processors and 575 MHz). Medium and fast GPUs however accelerate the computation to around 200 fps. Interestingly the FFTW-library chooses between different internal optimization schemes depending on the input size. This leads to higher computation times for small variations in image size (e.g. see the peak at tile configuration 36 in Fig. 2.7). Hence the choice of the grid-layout is critical depending on the type of implementation and FFT-library used.

Accuracy

To evaluate a registration algorithm typically the root-means-squared (RMS) error between the estimated and the ground-truth homography is calculated [70]. This error measure reports by how many pixels the estimated registration differs on average from the ground-truth.

We introduce a new *dataset B* consisting of 10 sequences (i.e. 4184 frames in total) taken from different aerial platforms: 6 from our own flights, 3 from [169] and 1 from [88]. An overview of the dataset is depicted in Fig. 2.8. Multiple samples frames for each sequence are shown in appendix A.1. The dataset poses various realistic challenges such as view-point variations, noise/blur, varying overlap between frames, different sensor quality and type of scenery (e.g. rural vs. city like scenes).

The sequences used for evaluation were captured from a camera mounted on an air-

Fig. 2.8: Dataset B used for evaluation of registration accuracy.

plane and no additional sensors to generate the ground-truth were used. Therefore, the accurate, but slow multi-scale dual inverse compositional [12] registration (see section 2.3.2) was used, which was run on a manually selected area best suited for registration. All registered frames were validated by visual inspection. In cases of insufficiently exact registration a manual feature-based pre-registration was applied. It is important to note, that the ground-truth registration is not perfectly accurate due to lens distortions and nonplanarities of the scene. Hence, a small RMS error only shows that the estimated homography is "close" to the ground-truth homography, which was acquired in a "supervised" manner.

In Tab. 2.2 the results of the registration experiment are presented. Most sequences can be registered at sub-pixel accuracy, allowing for subsequent tasks such as motion detection and image fusion. Differences in accuracy are due to variation in scenery and camera motion. Details on the camera motion are presented in appendix A.1 for each sequences. All results were achieved with the constant settings while only varying the tile configuration. The results confirm previous intuitive considerations. Using smaller tiles increases the registration performance until a lower bound size is reached at which the tiles do not contain sufficient image information to allow for precise correlation. For most sequences tile numbers between 25-64 tiles performs well enough. Only small drops in

	4	5	8	9	16	25	25+	36	64	100	225	400	625
1	0.66	0.59	0.51	0.50	0.45	0.30	0.46	0.33	0.30	**0.29**	0.33	0.34	0.38
2	1.08	1.05	0.98	0.97	1.15	0.50	0.82	0.40	**0.36**	0.38	0.50	0.55	0.69
3	0.43	0.39	0.36	0.36	0.29	0.21	0.30	**0.20**	0.21	0.21	0.24	0.27	0.29
4	0.34	0.32	0.29	0.28	0.27	0.21	0.26	**0.19**	0.22	0.23	0.32	0.32	0.39
5	0.42	0.36	0.36	0.35	0.20	0.19	0.20	0.19	**0.18**	0.18	0.21	0.21	0.26
6	0.32	0.26	0.26	0.22	0.22	0.22	**0.19**	**0.19**	0.20	0.21	0.23	0.25	0.26
7	1.04	0.83	0.82	0.72	0.66	0.64	0.55	0.48	**0.45**	0.46	0.50	0.54	0.59
8	0.66	0.64	0.64	0.64	0.49	0.48	0.48	**0.47**	0.54	0.54	0.62	0.62	0.73
9	0.57	0.57	0.49	0.44	0.41	0.32	**0.30**	0.44	0.48	0.44	0.57	0.59	0.76
10	0.79	0.77	0.76	0.74	0.68	0.58	**0.57**	0.58	0.60	0.59	0.71	0.71	0.87

Table 2.2: Registration results in RMS, where the columns correspond to different tile layouts and rows correspond to the 10 evaluation sequences.

registration accuracy are observed for slight changes from the optimal tile configuration, e.g. using 36-tile instead of optimal 64-tile configuration. The α-parameter of the Blackman window was set to 0.16. These results indicate the robustness of the tiled phase correlation to parameter settings and various scene types.

Windowing

Furthermore, we investigate the impact of different windowing functions. In Tab. 2.3 the RMS errors for different window functions and parameter settings are listed which were computed using sequence 4 from *dataset A*. It is clear that windowing produces better results than not using any window function and that the choice of the function or settings is not critical.

no window	Blackman ($\alpha=0.0001$)	Blackman ($\alpha=0.16$)	Cosine	Gaussian ($\alpha=0.3$)	Gaussian ($\alpha=0.1$)	Tukey ($\alpha=0.9$)	Tukey ($\alpha=0.6$)
0.358	0.234	0.23	0.266	0.288	0.25	0.318	0.245

Table 2.3: Comparison of RMS errors for different windowing functions and parameters.

2.5 Conclusions

In this chapter we discussed the general problem of registration, in geometric and in photometric terms. We investigated three different types of registration, i.e. feature-based, intensity-based and frequency-based, and motivate why we consider each of them optimal for the respective image fusion techniques addressed in this thesis. Furthermore, we presented a new frequency-based registration method called *tiled phase correlation*. We demonstrated that the computational speed of this method can greatly benefit from a GPU implementation making it the preferred choice in some applications, e.g. aerial platforms with limited computing power, over current state-of-the-art methods based on local features. We presented benchmark results for comparing GPU-based implementations (i.e. achieving 200 fps) which favorably compare to the results reported in the literature for more widely used local-feature-based approaches (i.e. achieving 100 fps). In addition, for the application of registering aerial imagery the phase correlation was shown to produce accurate homographies (i.e. down to sub-pixel accuracy) comparable to state-of-the-art approaches such as a combination of the dual-inverse-compositional and feature-based methods.

3 Superresolution

Many Pixels Are Not Enough.

Over the past years, superresolution has tremendously evolved starting from simple deblurring coupled with interpolation up to very intelligent multi-image fusion techniques. In section 3.1 we motivate these fusion techniques and discuss relevant applications. In section 3.2 we provide an extensive overview of the vast amount of work on superresolution, which was published over the last decades. In this section we also present details of the *iterative back-projection* method, which is employed in chapter 4 to combine high-dynamic-range imaging and superresolution. Finally, in section 3.3 we present our approach on combining still high-resolution images with low-quality videos. The literature that is relevant to our algorithm is discussed in section 3.2.4.

3.1 Introduction

Cameras are becoming increasingly accurate in capturing the infinitely detailed real world. One popular, but not very precise measure for the amount of information that a digital camera can capture is the number of pixels. In the last decade the number of pixels has multiplied by roughly a magnitude from 320×240 (in 1990, *Dycam Model 1*, also known as *Logitech FotoMan*) up to 4896×3264 (in 2009, *Canon EOS-1D Mark IV*). Cameras with over 1 billion pixels have also been developed, yet only for very special and rare applications [121].

Although this improvement in resolution over the recent years sounds very impressive, images and videos we capture will in general never have enough pixels. We see the following reasons leading to this conclusion. First, even a "high-resolution" image (e.g. with 20 million pixels) is still far away from the level of detail that the human eye and brain is able to capture. Second, many mobile cameras still do not have the latest high quality imaging sensors. Surveillance cameras still operate with standard "low-resolution" video formats like NTSC and PAL (around 720×576 pixels). Many cameras like in hand-held devices like cell phones, which make up the largest share in digital cameras being sold, employ very cheap sensor due to cost restrictions. Hence, the images suffer from bad illumination, motion blur and noise. Third, spatial resolution is also constraint by the infrastructure. In surveillance settings higher resolution means that more data needs to be stored and transferred. Therefore, more storage and higher bandwidth is required, which can be quite expensive. Furthermore, many existing systems have to operate for a long time with a given resolution. Fourth, increasing the number of pixels in the imaging sensor is not enough. The sensor elements also have to become smaller in order to capture finer details or expensive optics have to be used to achieve the same effect. However, decreasing the size of a pixel is accompanied with an increase in shot noise as the amount of captured light is lowered. Lastly, an increased number of pixels also make fast read-out of sensors technically more difficult and more expensive.

An alternative to increasing the resolution of imagery on the hardware side is to employ software-based solutions. One class of algorithms that increase the number of pixels and the level of detail, i.e. the spatial resolution, is called *superresolution*. An example of the amount of resolution improvement is depicted in Fig. 3.1. Superresolution has many different applications. A very obvious one are forensic investigations. Like in many action movies, crime investigation often faces the problem of limited spatial resolution. The suspect is just too blurry and unclear to be identified. Other applications relate more to graphics. Here the overall quality of the whole image or video should be increased (e.g. sharpen a blurry image). Superresolution has also been applied to scientific imagery like satellite images. In this application the motion model is quite limited to planar motion

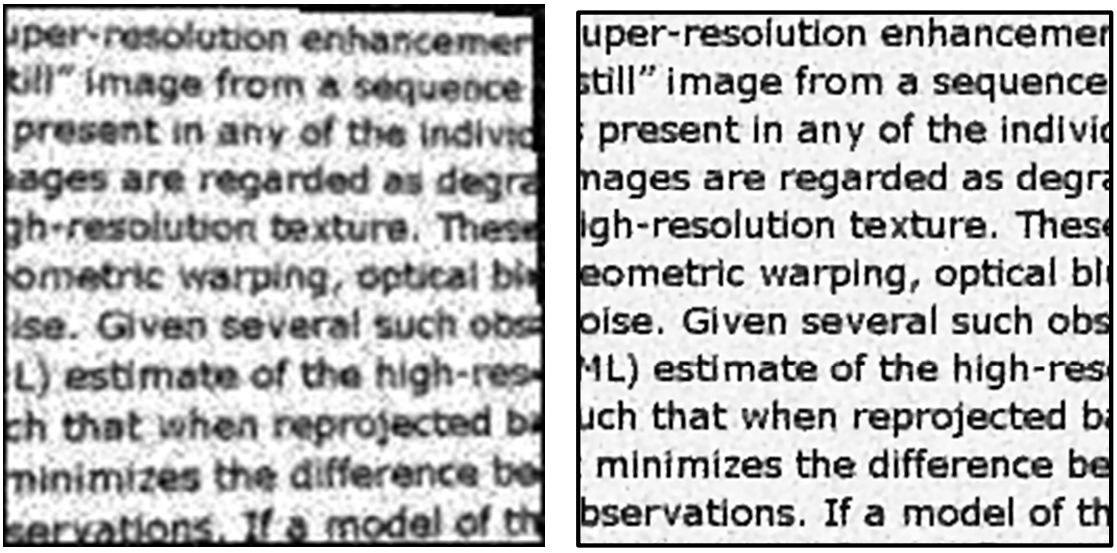


Fig. 3.1: An example how superresolution can improve the spatial resolution. Left: interpolated low-resolution input image. Right: superresolution result.

due to the great distance between the camera and the scene allowing for good resolution improvements.

3.2 Evolution of Superresolution Methods

Superresolution methods have evolved over a long history. Numerous methods with very different approaches have been proposed in the past. Many reviews [28, 173, 122, 31, 49, 21, 186, 127, 171, 130, 61] have been published, which summarize existing work, viewing at the problem of superresolution from various angles. In the following a chronological summary is presented, which highlights the most important corner stones in superresolution research.

Superresolution methods can be divided into many different categories. However, a very useful distinction can be made by looking at how they try to infer the missing information given multiple low-resolution images. One class of algorithms tries to "undo" the process of image degradation. These are called *reconstruction-based* methods as they try to reconstruct the true high-resolution image given multiple low-resolution images taken from

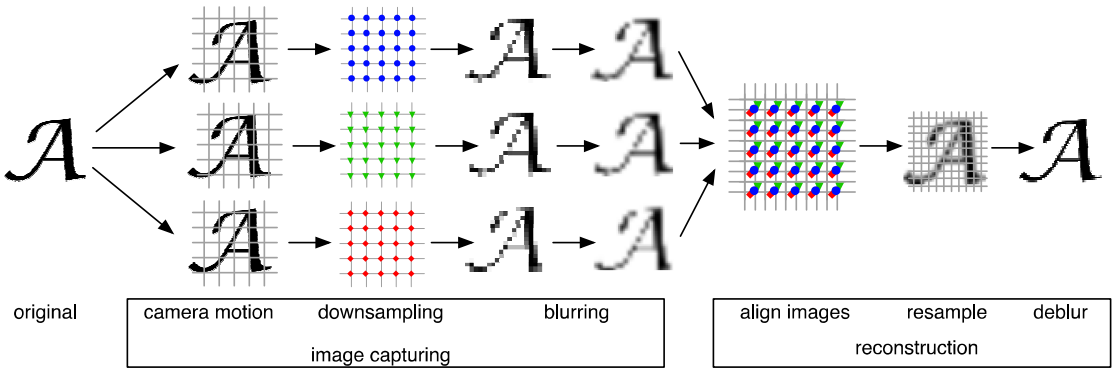


Fig. 3.2: Schematic for reconstruction-based superresolution.

the same scene with slightly different viewing angle. The reconstruction-based superresolution methods try to estimate the camera parameters used during capture. With these settings the unknown high-resolution image is reconstructed. This class of algorithms is discussed in the following section 3.2.1. An alternative class of algorithms tries to match the observed low-resolution images to previously learnt image content. These methods are thus called *learning-based* methods. Learning-based superresolution methods do not try to reverse-engineer the imaging process, but rather try to augment high-resolution image information with their known low-resolution appearance onto the low-resolution input image, e.g. via a patch-database with pairs of high- and low-resolution image patches. This class of algorithms is discussed in section 3.2.2. Yet another class of algorithms contains aspects of all those methods, which neither fit only one of the previous two classes and hence are referred to as *hybrid* methods. In section 3.2.3 this class of superresolution methods is discussed. All the methods discussed above address the superresolution applied to the spatial domain. In section 3.2.4 we discuss methods which extend superresolution to the temporal domain as well. This work directly relates to our contribution presented in section 3.3 where we combine still high-resolution images with low-quality videos.

3.2.1 Reconstruction-based Superresolution

In Fig. 3.2 the process of reconstruction-based superresolution is illustrated in more detail. During image capture various degradations are introduced into the image. The most notable ones are downsampling and blurring. Due to the discrete pixel grid of the camera sensor a true scene is discretized into a rather coarse image (see downsampling step in Fig. 3.2). These downsampled images are further degraded by blur introduced by motion, atmospheric effects and internal camera noise. How can these low-resolution images be combined to increase the resolution as shown in the right part of Fig. 3.2? It seems at first that during the generation of these low-resolution images, crucial information has been lost, which would be necessary to reconstruct the original high-resolution image. For example during down-sampling, multiple pixels in the high-resolution image are reduced into a single pixel in the low-resolution image. How could this many-to-one mapping be reversed? To answer these questions, it is helpful to consider a simple discrete image capturing process, i.e. integer down-sampling and no other degradations like blur, as illustrated in Fig. 3.3. On the left side of the figure a section of high-resolution image (i.e. the letter A) is enlarged to visualize the individual pixels (i.e. 4×4 pixel section) it consists of. A simple discrete image capturing process with an integer down-sampling factor of two generates multiple low-resolution images (i.e. 2×2 pixel section) as shown in the middle of the figure. Since the down-sampling factor is two, the low-resolution image consists of the same pixels as the high-resolution image except that every second in each dimension is lost. Hence, the 4×4 high-resolution pixel grid is reduced into a 2×2 pixel grid as highlighted by the letters in Fig. 3.3 ("zoom on original"). However, another low-resolution image, which was captured with a discrete integer offset, e.g. one pixel horizontally to the right as shown in green, captures high-resolution pixel values that were lost in the other two (i.e. green and blue). If these capture settings, i.e. down-sampling factor and offsets, were known, the high-resolution image could be reconstructed by simply reassembling the high-resolution pixel grid as shown on the right. Pixel values, which were not captured by any low-resolution image,

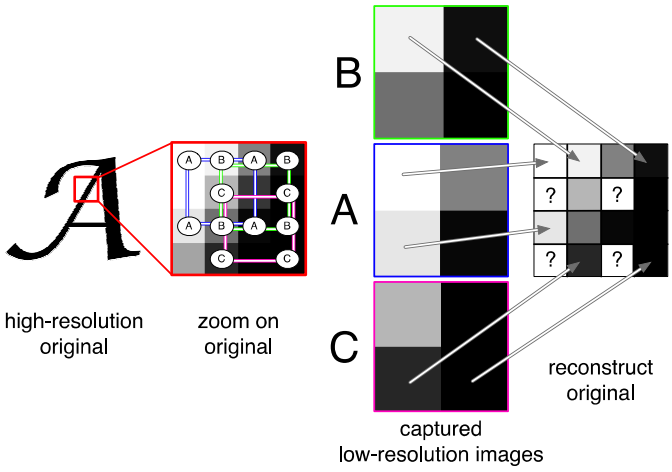


Fig. 3.3: Schematic illustration of reconstruction-based superresolution by image alignment.

need to interpolated from neighboring ones (i.e. pixel positions marked by "?" on the right of the figure. In other words the low-resolution images "scan" the high-resolution image at different pixel position, thus allowing the reconstruction. Although real image capturing is usually not discrete but rather continuous, the same principle holds.

Considering superresolution in more formal mathematic terms helps to understand the principle of the reconstruction-based approach. Let $\vec{h}$ denote the high-resolution image, linearized into a vector. We want to estimate $\vec{h}$ given a set of N low-resolution input images $\vec{l_i}$ with $i \in [0 \dots N]$. All the degradations mentioned above, which affect this high-resolution image and amount to the imaging or camera model can be formalized into a matrix M. Multiplying the high-resolution image $\vec{h}$ with this matrix simulates then the imaging process in mathematical terms

$$\vec{g_i} = \alpha \cdot D \cdot B \cdot T_i \cdot \vec{h} + \beta = M_i \cdot \vec{h} + \beta \tag{3.1}$$

where $\vec{g_i}$ denotes the simulated low-resolution image, D downsamples the image (by spatially averaging), B adds a global blur to the image and T_i geometrically transforms the image. In [28] a great discussion can be found on how this matrix M should be constructed. The parameters for the downsampling (D) are chosen by the user. In [8, 101] limits of possible values for this "resizing factor" are analyzed from practical

and theoretical view points. The blurring matrix B contains the size and shape of the point spread function (PSF), which can be either specified by the user or estimated simultaneously with the high-resolution image $\vec{h}$ [153, 130]. The same holds true for the geometric transformation T_i. However, if the registration parameters are also estimated along with $\vec{h}$ then a pre-registration is usually required[153]. The photometrical effects are captured by an affine model, i.e. additive (β) and multiplicative (α) factors, which represent brightness and gain changes. These can be estimated similarly to the geometric transformation parameters either prior to the superresolution estimation (see section 2.2.1) or simultaneously with the high-resolution image.

If we assume that the parameters of the matrices M_i have been given or estimated prior to the reconstruction, e.g. via registration or via user settings, then we can formulate the superresolution problem as a very sparse and large linear system just like Eq. 3.1:

$$\vec{g}_i = M_i \cdot \vec{h} \tag{3.2}$$

The number of elements, i.e. pixels, I in the observed low-resolution image $\vec{g}_i$ is usually a fraction of the total number of elements $d \cdot I$ of the high-resolution image $\vec{h}$ depending on the down-sampling factor d. Given enough input images however, i.e. if the number of input images N is greater than the down-sampling factor d, this system becomes an overdetermined system, which can be solved by standard means, e.g. directly via the pseudo-inverse or using an iterative scheme like least-squares [28]. Hence, if we were able to align multiple low-resolution input images to sub-pixel accuracy we can reconstruct these pixel at a denser grid.

The literature on reconstruction-based superresolution methods is very rich and the work on superresolution evolved over different, partly parallel lines of research. The first attempt was made by investigating the problem in the frequency domain [167, 164, 119, 22]. However, soon it became clear that a formulation in the spatial domain allows much more flexible algorithmic designs, which gave rise to a set of simple superresolution methods. In parallel some researchers tried to address the problem from a set theoretic point of view. These methods are based on projections onto convex sets (POCS) [154,

146, 160, 45, 124]. At last, the research mainly converged into considering superresolution as a formal optimization problem, which can be derived by statistical means. The latter approach is considered the state-of-the-art today. In the following we discuss the spatial domain-based methods and the optimization-based approaches in more detail as these are the ones we focus on in this thesis.

First Spatial-Domain-based Methods

First superresolution methods based in the spatial domain subdivided the problem into three steps: registration, upsampling and deblurring. Multiple low-resolution input images were registered to sub-pixel accuracy either by using a dense optical flow [65, 139] or by considering different degrees of freedom for a global alignment (2-DOF [170] or 3-DOF [90, 77]). The registered low-resolution images were merged, usually by a weighted average [170], which could include filtering for badly registered pixels [139]. This mainly eliminates noise in the observed low-resolution images, but does not reconstruct fine details that were lost due to the spatial averaging during downsampling. Therefore, the averaged upsampled image served rather as a first high-resolution estimate, which was then deblurred by standard methods [170, 139, 90] (e.g. Wiener filter) to generate a crisp and sharper looking final result image.

The fundamental step towards reconstruction-based superresolution defined in the spatial domain was made by work of Peleg et al. [125] and later significantly improved by Keren et al. [90] and Irani & Peleg [77, 78]. The core idea of these publications lays in the formulation of the superresolution as a generative reconstruction problem. A cost-function is defined based on the similarity between the observed and synthetically generated low-resolution images. Within an iterative optimization scheme this cost-function is minimized. During each iteration loop the synthetic low-resolution images are generated from the current high-resolution estimate. From the error of the cost-function an update for a new high-resolution estimate is derived. This update can be rather heuristic [125, 90] or based on a proper degradation model [77, 78] which the

author call *iterative backprojection* (IBP). The IBP of Irani & Peleg can be thus considered as a first sound reconstruction-based superresolution method. Various extensions to the IBP have followed like the substitution of a mean by a median to detect and exclude outlier pixels [194]. Another variation was proposed by Elad & Hel-Or [46] which radically limited the linear system of equations of the reconstruction to avoid an iterative solution. This simplified version was shown to be fast and yet effective. In chapter 4 we will extend this approach to work in the radiance domain, which allows to combine high-dynamic-range imaging with superresolution.

Maximum-A-Posteriori Approaches

Most of the early superresolution methods outlined so far, reassemble more or less maximum-likelihood approaches [28]. This means they seek a high-resolution image as an estimate, which maximizes the probability of being the solution given the observed low-resolution input images. The resulting image is often referred to as the *maximum-likelihood estimate* (MLE). In mathematical terms this can be written as

$$P(\vec{L}|\vec{h}) = \prod_i P(\vec{l}_i|\vec{h}) = \prod_i \prod_{\vec{x}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\vec{g}_i(\vec{x}) - \vec{l}_i(\vec{x})\right)^2}{2\sigma^2}\right) \tag{3.3}$$

where $\vec{L}$ contains the vectorized low-resolution input images $\vec{l}_i$ stacked on top of each other, $\vec{g}_i$ is the "simulated" low-resolution image defined by Eq. 3.1 and $\vec{h}$ is the sought high-resolution estimate. To find the MLE solution $\vec{h}$, which maximizes this probability, it is convenient to consider the negative log-likelihood

$$-\log P(\vec{L}|\vec{h}) = -\sum_i \log P(\vec{l}_i|\vec{h}) = \sum_i \sum_{\vec{x}} \log\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\vec{g}_i(\vec{x}) - \vec{l}_i(\vec{x})\right)^2}{2\sigma^2}\right)\right] \tag{3.4}$$

Since we are only interested in finding a solution $\vec{h}$, which maximizes this probability and minimizes the negative log-likelihood, we can omit the constant factors inside the

sum and simplify

$$\vec{h}_{\text{MLE}} = \underset{\vec{h}}{\text{argmax}} \left(P(\vec{L}|\vec{h}) \right) = \underset{\vec{h}}{\text{argmin}} \left(-\log P(\vec{L}|\vec{h}) \right) \tag{3.5}$$

$$\vec{h}_{\text{MLE}} = \underset{\vec{h}}{\text{argmin}} \left(\sum_i \|\vec{g}_i(\vec{x}) - \vec{l}_i(\vec{x})\|^2 \right) \tag{3.6}$$

$$\vec{h}_{\text{MLE}} = \underset{\vec{h}}{\text{argmin}} \|\vec{G} - \vec{L}\|^2 \tag{3.7}$$

where $\vec{G}$ contains the "simulated " low-resolution image defined by Eq. 3.1 stacked on top of each other and hence can be written as

$$\vec{h}_{\text{MLE}} = \underset{\vec{h}}{\text{argmin}} \|M \cdot \vec{h} - \vec{L}\|^2 \tag{3.8}$$

where the matrices M_i from the Eq. 3.2 imaging model were also stacked on top of each other to form the overall model M. It is easy to see that this convex cost function describes a least-squares solution to the linear system described in Eq. 3.2. Capel [28] analyzes in depth the behaviour of the MLE, such as its sensitivity to noise and errors in the model parameter estimation.

Although the MLE produces already notable enhancements in the resulting high-resolution image, its basic cost function can not overcome the imperfection of the imaging model and the ill-posedness of the problem. For instance when the parameters of imaging model are not absolutely accurate, the resulting high-resolution heavily overemphasizes noise as can be seen from many experiments in [27, 28, 130]. One way to reduce this effect is the inclusion of a regularizer.

$$\vec{h}_{\text{MLE}} = \underset{\vec{h}}{\text{argmin}} \|M \cdot \vec{h} - \vec{L}\|^2 + c(\vec{h}) \tag{3.9}$$

These additional constraints c "guide" the convergence process when the overdetermined set of equations is not well defined. Regularization is a very general mathematical tool that has been known for decades. Tikhonov (i.e. quadratic constraint term) or Total Variation (i.e. linear constraint term) regularizers have been applied to superresolution [73, 27, 122, 50, 118]. The importance of using additional constraints to aid the super-resolution estimation was already stressed in earlier work by Stark & Oskoui [154] and Elad & Feuer [45] in their work on POCS-based superresolution.

Although the regularizers serve their purposes very well, the justification is based on pure mathematical reasons without considering the superresolution domain. In other words, the goal of finding a high-resolution image and not any vector $\vec{h}$ that optimizes the cost function is not explicitly considered. However, an alternative, yet very similar formulation called *maximum-aposteriori* (MAP) estimation allows to interpret the constraining term in a more meaningful way and which is based on stochastic fundaments.

In the MAP formulation, not only the likelihood $P(\vec{L}|\vec{h})$ but also a prior probability $P(\vec{h})$ is considered. The probability models how likely the estimated high-resolution image is an optimal superresolution image. Using *Bayes' Theorem* a posterior probability can now be formulated as

$$P(\vec{h}|\vec{L}) = \frac{P(\vec{L}|\vec{h}) \cdot P(\vec{h})}{P(\vec{L})} \tag{3.10}$$

Various definitions of the prior probability are possible, ranging from very generic functions, i.e. the same as used for the regularizers can be used, up to very specific ones, e.g. a bimodal prior which specifically models black-and-white text images [42]. Similar to the MLE approach, the high-resolution image $\vec{h}$ is sought that maximizes the probability $P(\vec{h}|\vec{L})$

$$\vec{h}_{\text{MAP}} = \underset{\vec{h}}{\text{argmax}} \left(\frac{P(\vec{L}|\vec{h}) \cdot P(\vec{h})}{P(\vec{L})} \right) \tag{3.11}$$

Again, taking the negative logarithm helps to simplify the formulation to the following minimization problem

$$\vec{h}_{\text{MAP}} = \underset{\vec{h}}{\text{argmin}} \left(-\log \left(P(\vec{L}|\vec{h}) \cdot P(\vec{h}) \right) \right) \tag{3.12}$$

Using the same imaging model for $P(\vec{L}|\vec{h})$ as in Eq. 3.8 the standard MAP formulation results in

$$\vec{h}_{\text{MAP}} = \underset{\vec{h}}{\text{argmin}} \left(\|M \cdot \vec{h} - \vec{L}\|^2 + \lambda \cdot P(\vec{h}) \right) \tag{3.13}$$

where λ is a user-defined weighting factor which controls the impact of the prior on the final high-resolution image $\vec{h}_{\text{MAP}}$.

This class of superresolution algorithms was introduced to the community by early work of Hardie et al. [69], Cheeseman et al. [30] and Schultz & Stevenson [144] and

is now considered the standard for most state-of-the-art methods, at least for the core parts. Different methods mainly vary only in the type of prior employed. For instance Hardie et al. [69] use a L2 norm on the second order derivatives. Cheeseman et al. [30] use an average of the 4-neighbors for a pixel. Ng et al. [118] use the L1 norm on the second order derivatives. Schultz & Stevenson [144] and Capel [28] employs a Huber prior, which combines a L1 and L2 norm on the gradients of the superresolution estimate.

Although MAP with generic priors like the Huber prior work very well in practice, there exist more or less arguable limits on this general type of superresolution. Baker et al. [8] was one of the first to formally address this limitation. Lin & Heung-Yeung [101] extended this analysis and lifted it up to a much more theoretical and well founded level. However, it must be noted, that some conclusions from these investigations were already highlighted in earlier work of Elad & Feuer [45]: the key to successful superresolution lays in the strength of the prior. This lead to numerous creative ways to extend the power of the prior. For instance Fletcher et al. and Donaldson & Myers [56, 42] addressed specifically the superresolution of text and proposed especially designed priors. Where the former is merely a simple extension of the IBP to "sharpen" edges a bit, the latter actually extends a MAP with Huber prior by a bimodal prior which ensures that the superresolution estimate exhibits the typical bimodal intensity distribution (i.e. black and white) of text images. Further extensions on such priors are discussed in the section section 3.2.3 on hybrid approaches to MAP superresolution.

Being Robust

The superresolution is a complex reconstruction in which different algorithmic steps interact tightly with each other. The cost-functions for MAP or MLE formulations assume that the parameters for the imaging model are exactly known. For instance the registration parameters, which are required for the restoration step, are best estimated not using the given low-resolution images but using high-resolution images, which how-

ever are the desired output of the whole superresolution process. Typically this loopy dependency is ignored and these imaging parameters are estimated beforehand in form of image registration (geometric and photometric) and in form of an estimation of the point-spread-function. With real data these parameters are often difficult to compute to absolute accuracy. Therefore the uncertainties and errors made in the estimation of the parameters for the imaging model have to be handled by the restoration part in a robust way, for instance by applying strong priors.

Instead of treating the estimation of the imaging parameters and the restoration step separately, methods were proposed to solve these simultaneously. This idea was already presented by Hardie et al. [69], who employed a gradient descent optimizer that solved for the registration parameters and the superresolution image in an alternating fashion. This combination helped especially due to not very powerful registration methods, e.g. very often block matching was often used assuming just global translation [69]. In [179, 132] a similar alternating optimization scheme was employed, where also the point-spread-function was estimated on-the-fly additionally to the registration parameters. Tipping & Bishop [163] marginalize the superresolution estimate out of the MAP formulation to get a good estimate for the registration and blur. Because of the complexity this can be carried out only on small patches (e.g. 9×9 pixels) of the low-resolution estimates. Pickup [130] adapts this scheme, but marginalizes over the registration parameters instead. Given an initial registration and blur estimate, the superresolution solution fully includes the uncertainty of these estimates. Also in [35, 174] some pre-registration is employed, which is then refined in an alternating optimization scheme. They adapt their previous work on blind deconvolution [173], which recovers local blur kernels from multiple unregistered images, to include also downsampling in the local blur kernels [175]. The kernels also include the effect of a point-spread-function and small motion variations. The latest approaches [130, 174] represent probably the most sophisticated MAP superresolution methods using generic priors.

3.2.2 Learning-based Superresolution

One of the main goals of multi-frame superresolution is the enlargement of the input images. Increasing the size of a single images is achieved by sampling the smaller image at a finer pixel grid. For the positions where no intensity values from the small image are directly given, values have to be approximated or so called *interpolated*. This interpolation is usually done by treating the image as a 2D signal and approximating the missing pieces with generic functions like lines (i.e. linear interpolation) or higher order polynomials (i.e. cubic interpolation).

Around the same time when the multi-frame spatial-domain-based reconstruction-based superresolution methods became more popular, researchers like Candocia & Príncipe started to investigate, whether this interpolation can be learnt from other images. This does allow to fill in larger gaps, i.e. allow greater resizing factors, with more meaningful values [26]. With the seminal paper of Freeman et al. [58] this line of research gained much interest. Freeman et al. investigated how high-resolution patches can be augmented onto a low-resolution input image to make it look "visually plausible". The goal was not a true reconstruction, but to add missing high-frequency components to low-resolution input images by matching high-resolution image patches from a database to the low-resolution input image. This process is schematically outlined in Fig. 3.4. The input image was chopped into low-resolution image patches. The best matching low-resolution image patch in the database was selected and its corresponding high-resolution image patch was pasted into the high-resolution result image. Given a sufficiently large database and small patch sizes, the final image looked mostly smooth, even given the fact that the database was generated with generic images different from the one being processed. The layout and augmenting is modeled via a Markov Random Field (MRF) to avoid visual discontinuities by enforcing consistency with a neighborhood. This work was extended by other researchers in many different ways. Sun et al. [156] proposed another matching scheme based on derivatives (so called "primal sketches") to map high-resolution patches from the database onto the low-resolution images and included

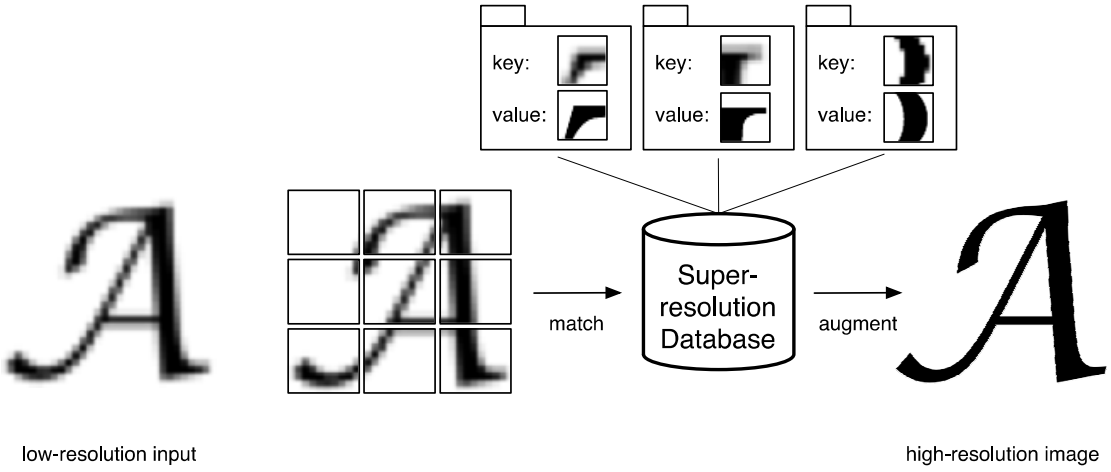


Fig. 3.4: Schematic for learning-based superresolution.

a reconstruction term, which should ensure that the resulting image does not diverge too much from the original input image. Tappen et al. also included a reconstruction term and employed a special prior for natural images [159]. Other modifications regard the image model [155], the formulation in the frequency domain instead of the spatial domain [82, 161], the estimation of additional parameters like of the point-spread-function [14], the application to video by explicitly including spatio-temporal constraints [20, 92] to avoid temporal flickering or the focus on specific domains like faces as explored by Liu et al. [103]. A latest trend in single image superresolution is the application of sparse coding techniques to the field of superresolution, which allows a compact representation of the patches (i.e. smaller databases) and much more efficient matching as not the patches itself are used for augmenting, but a co-occurrence prior [114, 185]. Similar to the reconstruction-based superresolution, the fundamental limits for learning-based methods have been investigated to see what resizing factors are achievable. However, results from the work of Lin et al. [102] are rather theoretic with little direct implications for practical applications.

The main focus for these single-frame learning-based methods are to produce visually plausible images, rather than a true reconstruction of the true scene. However, depending on the quality of the offline learned inference machinery, the resulting images

can exhibit strong artifacts, which make the images look worse than the input. Consequently, researchers tried to include reconstruction constraints to avoid this, such as the ones mentioned above [156, 159]. These single image methods can be extended to multi-frame methods, which then can be considered as a hybrid approach to multi-frame superresolution as they are both: reconstruction-based method (superresolution because of geometric displacement of the camera) and learning-based method (superresolution because of the inference using offline information). These hybrid methods will be discussed in the next section 3.2.3.

3.2.3 Hybrid Approaches To Superresolution

Although the two classes of algorithms described in sections 3.2.1 and 3.2.2 appear very different in their nature, they can be successfully combined. A pioneering paper in the field of such hybrid superresolution methods was published by Baker et al. [8]. They were the first to formally examine the limits of purely reconstruction-based superresolution. They concluded that certain high-frequency components get lost during the image degradation process (e.g. quantization into discrete intensity values), which prohibit a true reconstruction for larger magnification factors (i.e. > 2). According to their arguments generic priors alone like the huber prior do not help to recover this missing high-frequency information. The only way to recover the missing information is to make use of additional model-based constraints. Baker et al. included thus a so called *learning-based prior* to the standard MAP multi-frame reconstruction-based superresolution formulation, which augments or "hallucinates" high-frequency components onto the final image. This ensures that the superresolution MAP solution is close to a recognized class. In mathematical terms this approach is based on the standard MAP formulation (see Eq. 3.13). The contribution lays in the choice of the prior $P(\vec{h})$. Instead of using a generic function, which does not directly relate to the image content, they employ the following recognition-based prior:

$$P(\vec{h}) = \sum_k P(\vec{l} \in C_k) \cdot P(\vec{h}|\vec{l} \in C_k) \tag{3.14}$$

The first term $P(\vec{l} \in C_k)$ is the output of the classifier. The authors employ a pixel-wise nearest-neighbor classifier using a multi-scale gradient-based feature. This matches to each low-resolution pixel from the input a high-resolution pixel from a database. The second term is a cost function, which ensures that the gradients of the solution $\vec{h}$ are close to the high-resolution match. Depending on the class of images (e.g. faces or text) a specific database of high-resolution images has to be used. They evaluated their methods on face and text recognition and show that for large magnification factors (> 2) this hybrid approach works much better than methods using generic priors. A major drawback of their approach is the restriction that the low-resolution images from the input and from the database with the high-resolution image may only differ by translation in geometrical terms. Dedeoglu et al. [39] extended this approach to the application of video superresolution by enforcing spatio-temporal consistency in the learning-based prior formulation.

Along similar lines Capel [28] proposed to substitute the prior in the standard MAP formulation (see Eq. 3.13) using a PCA-based prior instead. Similar to the recognition-based prior of Baker et al. this ensures that the solution $\vec{h}$ is close to some offline stored high-resolution image data. Capel also proposed an alternative in which the MAP formulation is completely ported to the PCA subspace. This idea was also adapted by Gunturk et al. [66]. The main advantage of the latter approach lays in the reduction of the number of equations that need to be solved for MAP superresolution. Wang et al. [176] also proposes a learning-based prior for the MAP formulation. They employ a MRF-based approach similar to the one in [58] to enforce that the solution $\vec{h}$ is close to a learnt database of high-resolution patches. A last notable approach is work of Pickup et al. [131], which describe the image prior $P(\vec{h})$ by texture synthesis [44], which also samples from offline or learnt high-resolution images like the methods mentioned above.

3.2.4 Spatio-Temporal Superresolution

Superresolution methods can be applied to spatial as well as temporal dimensions [147] or to both [142, 147, 126, 178, 117, 92, 18, 67]. Shechtman et al. [147] discuss in-depth what the trade-offs are between temporal and spatial resolution enhancement, given multiple input videos or images with a certain spatial resolution and frame rate. A special input configuration is the combination of a high-resolution still image (i.e. input video with a very low frame rate, but high spatial resolution) with low-resolution video (i.e. input video with high frame rate, but low spatial resolution). The result is a high-resolution video with high frame rate. The applications include generating high-resolution stereo video footage [142] as used in IMAX films [54], video editing [18, 67], processing input from dual-mode cameras [126, 92]. A novel application presented in this chapter is to increase the resolution of a low quality video with automatically retrieved high-resolution images from large image databases in the Internet. Sawhney et al. compute stereo motion information between a low-resolution video (e.g. left view) and a high-resolution video (e.g. right view) and use this information to warp the high-resolution frames into the viewpoint of the low-resolution video and combine them to generate a high-resolution version of that view. Pelletier et al. [126] describe a camera, which produces simultaneously two types of frames: one with high resolution, which is read out at long intervals (i.e. due to longer exposure times) and one with low resolution, which is read out at high frame rate (i.e. due to short exposure times). They compute the motion for the low-resolution frames and use this information to synthesize the missing high-resolution frames from the ones taken at long intervals. Because of the longer exposure times, the high resolution frames are sensible to motion blur, which the authors compensate by applying a deblurring step. A similar input configuration and motivation was discussed by Watanabe et al. [178] and Nagahara et al. [117]. Both consider the application of a dual mode camera. The former combine the two inputs by a weighted average of the discrete-cosine-transform (DCT) spectra of the input images to transfer the high-frequency components from the high-resolution image to the

low-resolution frames. The latter synthesize the result image only by propagating the high-resolution pixels using motion information computed on the low-resolution video to generate the missing frames (i.e. similar like [126]). Kong et al. adapt the system from [156]. They key difference is the online generation of a patch database using the high-resolution image, which is in turn used to synthesize the high-resolution version of the low-resolution video frames. Bhat et al. [18] present a similar system, but significantly improve the spatial registration of the low-resolution video frames and the high-resolution still image. They employ a complex multi-view-stereo algorithm, which handles parallax effects. Another notable contribution is the image-based rendering step based on a Markov Random Field (MRF) formulation, which selects for each coordinate in the final video the best fitting still image from which the pixel values will be taken to reassemble the resulting frame. As opposed to most previous approaches, the fusion is performed in the gradient domain by constructing a space-time gradient field from which the final image is reconstructed. This last step ensures that the frames do not exhibit any spatial or temporal discontinuities. Recently Gupta et al. [67] substituted the complex multi-view-stereo registration of [18] with a powerful optical flow method [140, 141] to be able to handle dynamic scenes as well.

All methods mentioned above rely on image rendering techniques to reconstruct the high resolution, high frame rate video. None actually make use of redundant pixel information in the low-resolution frames to superresolve those in areas, where no pixel information from a high-resolution input is available. In this chapter we explicitly address this issue and present a method that combines high resolution still images with low resolution videos within a maximum-aposterior-based superresolution framework.

3.3 Fusing High-Quality Images With Low-Quality Videos

In this section we address how a reconstruction-based superresolution approach can be extended to combine the high temporal resolution and the high spatial resolution of the input. We focus on a special input configuration, where a high-resolution still image and

a low-resolution video is given, which is common for spatio-temporal superresolution problems. Many applications exists, such as dual-mode cameras that simultaneously capture images and videos or like movie post-processing systems that enhance videos without the need to retake a scene.

Our research is further motivated by a new application. Recently a lot of research has been directed towards the question: What can be done with "brute-force vision" using very large amounts of data? Image retrieval methods have been shown to succeed on collections of images with sizes over a million. Various applications such as object recognition, 3D geometrical arrangement of images showing the same scene or inferring missing image regions can benefit from large image databases. Motivated by this research we propose an alternative use of image information stored in large pools such as the internet. Given an input video, we can utilize corresponding still images stored at much better quality to improve the overall quality of the video. In section 3.3.1 we discuss this application scenario in more detail.

In contrast to the state-of-the-art methods, we propose a novel hybrid superresolution scheme to smoothly incorporate the high-frequency components from additional database sources into a standard MAP superresolution process. On those areas where hallucination of details fails, i.e. no additional information is available, a standard MAP-estimation of the high-resolution image with a generic prior is performed. In that respect the proposed approach can be considered as a novel extension to recognition-based MAP using a model-based prior (see methods described in section 3.2.3). It combines the strength of generic priors with specific knowledge about the scene. This allows to combine high-resolution input at very low-framerate (i.e. still image) and input with high-frame rate but low resolution (i.e. video). The final result has high-resolution at high frame rate. An overview of our approach is given in section 3.3.2 and discussed in detail in the sections thereafter.

3.3.1 Application Scenario

Much attention has been paid to the vast amount of information available through images in online databases like Flickr, Yahoo Image and Google Image Search in the past years. Methods for using image information contained in large image collections have been shown in work of [71, 152, 166, 128]. Given the technology to handle such databases especially in terms of retrieval many new applications arise. Hays & Efros [71] utilized the countless information of Flickr for scene-completion in a "brute-force vision" scheme. Snavely et al. [152] demonstrated an automatic and very robust sparse 3D reconstruction method, which formed the core of a 3D image browser. Torralba et al. [166] analysed what amount of data is actually needed for non-parametric "brute force vision" methods to work and what benefits arise using large databases for various vision tasks like classification, detection, etc. Philbin [128] showed, in the context of object retrieval, efficient search in those computationally difficult accessible large databases.

This motivates the following application scenario: Assume we have a low-quality input video from a famous site for instance (e.g. Notredam Cathedral in Paris) that we wish to enhance. Using the above methodology one can retrieve a best matching image showing that same building in a similar view. We query the Flickr image database with certain keywords for a scene (e.g. "buckingham palace") and download about 500 images via the Flickr API. Usually, it is easy to manually select a best fitting image. We also assume that these still images are often available at much better quality than the input video. The goal is then to incorporate the details from the matched still image into the input sequence.

3.3.2 System Overview

We propose to use a combination of a reconstruction-based superresolution algorithm (MAP) with high-resolution prior information to merge both, the low-resolution image sequence and the high-resolution still image. Most likely the video and the still image

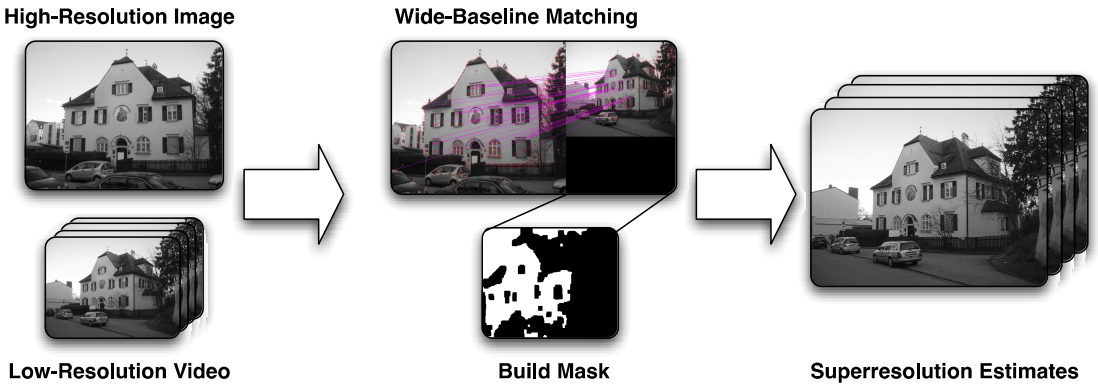


Fig. 3.5: Outline of the proposed method: matching, masking, superresolution.

were taken from different viewpoints. Thus a wide-baseline matching needs to be applied, in order to align the still image to one of the low-resolution frames (e.g. the first frame). Since we compare high- and low-quality images, which were possibly taken by different cameras, we employ a masking to identify regions which overlap and hence can be merged. We then apply a MAP-based superresolution using a generic prior on the unmasked parts. For the overlapping regions we constrain the superresolution result with the additional high-resolution information provided by the still image. In Fig. 3.5 the proposed method is outlined. Thus the method can be summarized into 4 steps:

1. initial registration of video frames and still image
2. masking
3. refined registration using masks
4. hybrid superresolution

which we will address in the following sections.

3.3.3 Initial Registration

In order to merge pixel information from both inputs, the still image needs to be spatially aligned with the video. The high-resolution image $\vec{p}$ and each low-resolution frame $\vec{l}$ can

be recorded from quite different view-points resulting in wide-baseline between them. As discussed in section 2.3.1, feature-based registration is the most suitable choice for estimating a homography, H, that geometrically aligns these images. We employ SURF features [13] computed on each image and the gold standard feature-based registration as described in section 2.3.1. To cope with the wide-baseline scenarios, the estimation is repeated multiple times (e.g. 3-5) using increasingly restrictive matching criteria and outlier thresholds. Using 8-DOF homographies is sufficient as we assume a static, planar scene, e.g. camera zooming and rotating around its center or far distance between scene and camera.

3.3.4 Masking

The two inputs might have been taken at different time points, under varying viewing angle and by cameras with different quality. Hence, besides structural and dynamic changes in the scene (e.g. a person walked into the scene or a parked car has moved) there will also be differences in global illumination, local reflections, image quality and even some imperfect alignment in the initial registration step. Therefore, only certain regions will overlap in both images which can be merged. Areas of moving objects or other structural changes in the scene need to be excluded from the fusion process. We automatically generate a binary mask specifying regions, which are identical in both inputs and which are not. A good similarity measure needs to be found which is sensitive enough to distinguish between subtle differences in textured areas and structural changes in the scene. Besides a good initial geometric registration, a global photometric registration (contrast and brightness) is performed to reduce illumination differences. For simplicity no explicit temporal coherence is enforced in the current implementation. Therefore, the robustness of the employed similarity measure is crucial for visually stable results (otherwise ghosting-artifacts may appear as shown in section 3.3.7).

Various similarity measures like sum-of-squared differences (SSD), mutual information (MI), normalized cross-correlation (NCC) and histograms of oriented gradients (HOG)

can be employed. Following consideration have to be made for a good choice:

SSD is too sensitive to mis-alignment or illumination changes and not sensitive enough to similarly textured, but different areas – a single pixel difference does not give enough information.

MI is computed on a small region around each pixel. Generally mutual information is a very robust measure used for alignment of images coming from different sensors in medical imaging for instance. Hence, it is insensitive to global differences in illumination. However, it is difficult to estimate meaningful joint and marginal probabilities for small areas (10×10 pixels). Furthermore, on homogeneous regions marginal entropies are very low resulting in an overall low mutual information value independent of whether these regions are similar or not.

NCC is also computed patchwise. It gives very low errors on well aligned structured areas (e.g. border of windows) but is still sensitive to local illumination changes (in Fig. 3.6 sky and homogeneously textured wall of house have almost same error as mis-aligned section of the roof because of slight illumination differences).

To measure the difference in structure and not in illumination, the measure should only consider high-frequency content. We adapt HOG features [36] computed on a dense grid to measure image similarity. The use of histograms achieves some robustness to noise via binning. They robustly represent image structures in a similar way to SIFT [105].

Some examples of high- and low-resolution input images and their various dissimilarity measures are shown in Fig. 3.6. The HOG measure gives the best mask for the two inputs. To generate a binary mask, the dissimilarity image needs to be binarized via thresholding. Erosion and dilation operations are applied to refine the mask.

3.3.5 Refined Registration

Using the mask from the previous step, a refined registration is performed on the unmasked areas. In order to process a whole video sequence, the registration with the

Fig. 3.6: Top: Aligned sample input images (Left: first video frame, Right: still image). Bottom: different dissimilarity images between inputs (Left-Right: SSD, MI, NCC, HOG).

still image is performed for all video frames. The binary mask is obtained using the first video frame and then aligned to the other video frames using the homographies from the last registration step. For the refined registration we employ the dual inverse compositional [12] approach, which is discussed in section 2.3.2. This method serves as a compact framework for photometric registration (as applied in previous steps) and simultaneous geometric alignment. Afterward, the mask building process is repeated with slightly more sensitive parameter settings to obtain the final mask (see top of Fig. 3.6 for an example).

3.3.6 Superresolution With High-Resolution Prior

We employ the standard imaging model (see Eq. 3.1) and recapitulate in the following:

$$\vec{g}_i = \alpha \cdot D \cdot B \cdot T_i \cdot \vec{h} + \beta = M_i \cdot \vec{h} + \beta \tag{3.15}$$

α and β photometrically deform the image, D downsamples the image (by spatially averaging), B adds a global blur to the image and T_i geometrically transforms the image (for simplicity we restrict the transformation to homographies and omit lense distortions).

The blur and downscale parameters have to be set manually. The transformation parameters (photometric and geometric) are computed by registering the frames $\vec{g}_i$ of the video. Since the inter-frame motion is usually small, we employ again the same intensity-based method [12] as in the mask refinement step and which is discussed in section 2.3.2.

We employ a MAP superresolution formulation as described in Eq. 3.13:

$$F_{standard} = \sum_i \|M_i \cdot \vec{h} - \vec{l}_i\|^2 + \text{constraints} = \|M \cdot \vec{h} - L\|^2 + \lambda \sum_{\vec{x}} \rho(\nabla(\vec{h}, \vec{x}), \delta) \quad (3.16)$$

To constraint this ill-posed problem, we use a generic prior, which is defined by a Huber-function $\rho(.)$ applied to the norm of the gradient of the superresolution estimate:

$$\rho(\nabla(\vec{h}, \vec{x}), \delta) = \begin{cases} \left\|\nabla(\vec{h}, \vec{x})\right\|^2 & \text{if } |\nabla(\vec{h}, \vec{x})| < \delta \\ 2\delta|\nabla(\vec{h}, \vec{x})| - \delta^2 \end{cases} \quad (3.17)$$

$\nabla(\vec{h}, \vec{x})$ is the magnitude of the gradient of $\vec{h}$ at position $\vec{x}$. The Huber-function combines the smoothing Tikhonov regularization with the L2-norm ($\|.\|$) and the edge-preserving total variation regularization with the L1-norm ($|.|$) via a fix threshold δ. This type of generic prior has been successfully applied in previous approaches and shown to be very robust [144, 28, 130].

The main goal of this part is to incorporate available prior high-resolution information into the superresolution estimation process in corresponding areas of the input. For this a binary mask was constructed in the previous step that indicates whether a pixel of the input has a corresponding high-resolution pixel in the high-resolution prior image or not. In the unmasked areas we want to use a generic edge preserving prior (e.g. Huber-prior) and in masked areas the solution shall be close to the corresponding high-resolution prior. Since the high-frequency are most responsible for sharp details perceivable by human eye (similar argument were used for instance in [58, 9]) and the low-resolution content from the input sequence shall be preserved to obtain a smooth result, the following additional gradient based prior is added to the cost function (see Eq. 3.16):

$$\|W(\nabla(\vec{h}) - \nabla(\vec{p}))\| \quad (3.18)$$

∇ computes the gradients (vertical and horizontal) and W denotes the mask. Therefore, the following final cost-function is subject to minimization:

$$F_{combined} = \|M \cdot \vec{h} - L\|^2 + \lambda \sum_{\vec{x}} \rho(\nabla(\vec{h}, \vec{x}), \delta) + \gamma \|W(\nabla(\vec{h}) - \nabla(\vec{p}))\| \tag{3.19}$$

The first term is the reconstruction constraint from the MAP formulation (Eq. 3.16). The second term refers to the generic image prior, the Huber-function (Eq. 3.17) and the third term contains the *high-resolution prior* (Eq. 3.18) using the high-quality input image $\vec{p}$. Using experimentally found values for the weights (λ, γ) for the priors and for the threshold δ, a smooth augmention of the additional high-resolution prior is achieved.

Because of the large numbers of equations, we employed the very fast limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [104] for optimization, but any non-linear optimizer like conjugated gradient descent could be used instead.

Usually, only a few iterations are needed (30-100) until all residuals are minimized as shown in Fig. 3.7. Similar to [132] we initialize the optimizer with a MLE estimation (just the reconstruction constraint from (Eq. 3.19) is optimized for a few iterations) which gives a much sharper starting point than the average image which is usually used [28]. We also allow for a registration refinement as suggested in [132], by iteratively solving equation (Eq. 3.19) and registering the low-resolution input L to the current superresolution estimate $\vec{h}$. For registration we use the dual inverse compositional approach from [12]. Usually, after the first two iterations the registration refinement stops updating the warping parameters. Note that no explicit temporal coherence is enforced. However, for the standard MAP superresolution temporal smoothness is achieved by combining multiple input images to one high-resolution image.

3.3.7 Results

The proposed algorithm was tested on various real sequences. The evaluation is performed in a qualitative measure comparing interpolation, standard MAP and the presented method. The first video shows the “Notre Dame Cathedral”. The camera is

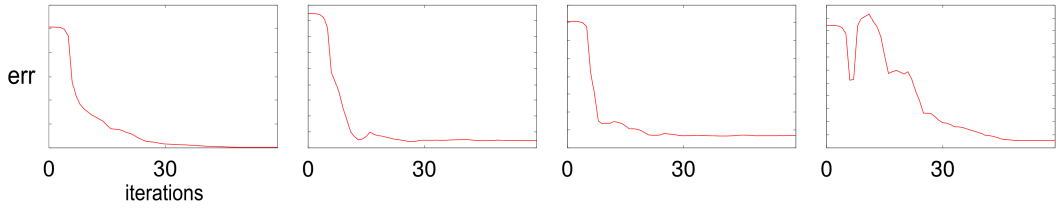


Fig. 3.7: Individual residual errors along the iterations for a typical computation (Left-Right: total cost, reconstruction term, Huber-prior, high-resolution-prior).

Fig. 3.8: First frame of input video and high-resolution image for "Notredam" sequence.

mostly rotating around its center justifying the use of homographies for the registration of the frames. We used the Flickr API to download 500 images labeled with "notre dame". The most suitable high-resolution image was manually selected from this set, although retrieval methods such as [152, 71] could automate this step. The first frame of the video and the corresponding high-resolution image from Flickr are depicted in Fig. 3.8.

Because the video and the still image were taken from different viewpoints, the buildings on the left side of the cathedral are slightly moved. This part is highlighted by the mask

Fig. 3.9: Left: Bicubic interpolation of low-resolution input, Middle: MAP superresolution using Huber-prior, Right: MAP superresolution with Huber-prior and high-resolution prior.

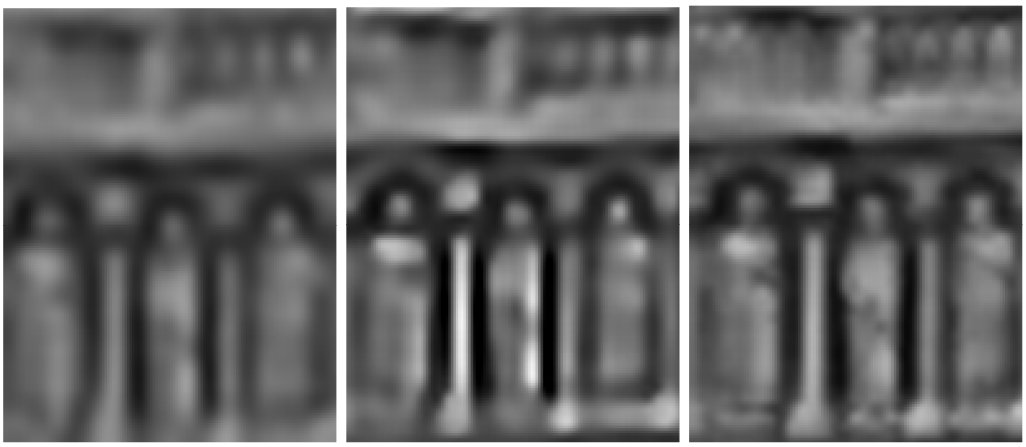


Fig. 3.10: A zoom into images from Fig. 3.9 (marked by yellow rectangle in left image).

so that no information of the high-resolution still image is included and only standard MAP superresolution is performed. However, very fine details on the cathedral can only be recovered by the additional high-resolution image. The results of standard MAP superresolution on the whole image (10 frames were used to generate 1 high-resolution image) and the proposed combinations are shown in Fig. 3.9 and Fig. 3.10. The standard superresolution is capable of emphasizing edges and improving the quality of the image. Fine structure however can only be made visible by including additional knowledge of the details, for instance by querying the internet for a high-quality shot and including this information into the superresolution estimation process.

Another example shows a poster of a car (see Fig. 3.11). The superresolution enhanced

Fig. 3.11: Left: high-resolution still image, Middle: low-resolution video frame, Right: MAP superresolution with Huber-prior and high-resolution prior.

images show fine details which could have never been recovered by standard superresolution alone. The still image is smoothly included into the reconstruction process (for example the lighting is also adjusted). In cases where the masking is incorrect, e.g. due to missing temporal coherence, ghost-edges appear as the MAP estimation tries to reconstruct the parts in the scene which are not really present (in Fig. 3.11 mask reaches over left border of the poster). Areas outside the mask (everything else than the poster) are processed with standard MAP superresolution and also show some quality improvement over the low-resolution input.

The third sequence shows a movie poster behind a window (see Fig. 3.12). It is straightforward to find high-quality still images of movie poster in the internet. Because of very strong reflections in the window, also parts of the poster are masked out (for instance parts of the headline text). However in regions where the high-resolution prior is applied a very strong quality improvement can be noticed. In the other regions slight enhancement by standard MAP superresolution is achieved.

Further results presented in figures 3.13, 3.14 and 3.15 show similar improvements as the previous results.

Fig. 3.12: Left: high-resolution still image, Middle: low-resolution video frame, Right: MAP superresolution with Huber-prior and high-resolution prior.

Fig. 3.13: Left: Bicubic interpolation of low-resolution input, Middle: MAP superresolution using Huber-prior, Right: MAP superresolution with Huber-prior and high-resolution prior.

Fig. 3.14: Left: Bicubic interpolation of low-resolution input, Middle: MAP superresolution using Huber-prior, Right: MAP superresolution with Huber-prior and high-resolution prior.

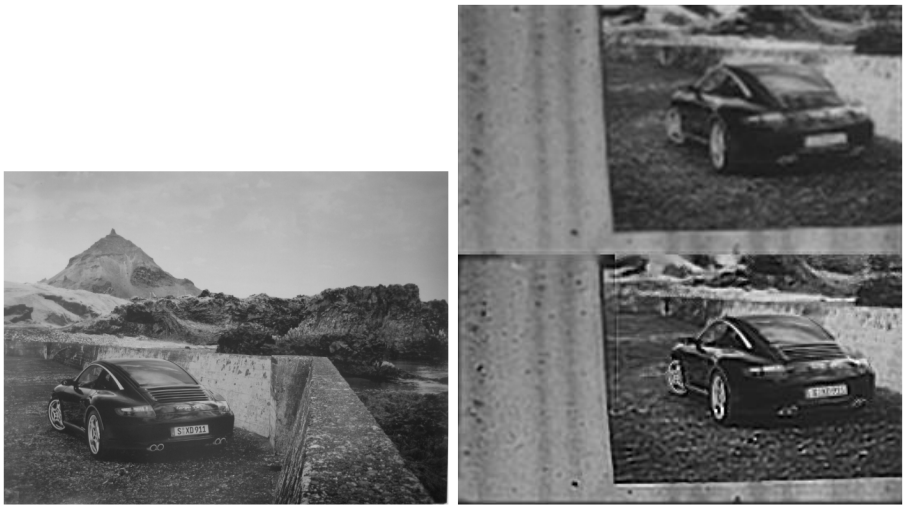


Fig. 3.15: Left: high-resolution still image, Right-Top: bicubic interpolated low-resolution video frame, Right-Bottom: MAP superresolution with Huber-prior and high-resolution prior.

3.4 Conclusions

We motivate the use of high-resolution information in form of still images available from the internet as an additional prior for MAP-based superresolution on videos. The results demonstrate that it is possible to smoothly include the high-resolution images into the superresolution estimate. With robust registration the two inputs can be aligned and a masking procedure highlights parts that can be merged. Unmasked areas are enhanced using standard MAP superresolution. However, accurate registration and masking of matching regions between the two inputs is crucial to avoid ghosting artifacts. A more complex registration method like the one used in [18, 67] could help to improve this point. Furthermore temporal consistency is not explicitly enforced, which also leaves room for improvement.

4 High-Dynamic-Range Imaging

Cherry Picking Best Pixels To Assemble HDR Images.

Similar to superresolution discussed in chapter 3, high-dynamic-range imaging is also a multi-frame fusion method to generate enhanced images. Although many solutions have been proposed for each of the two research fields independently, little attention has been paid to how these two are related. Therefore, in this chapter we present a framework that combines both enhancement methods and produces high-dynamic-range and high-resolution images from low-resolution, low-dynamic-range videos. In section 4.1 we motivate high-dynamic-range imaging and discuss relevant applications. In section 4.2 we provide a literature review of high-dynamic-range imaging. Because existing methods typically require a complex setup for recording of many differently exposed images, we propose an algorithm, called *Minimal-HDR*, which minimizes this requirement for the fusion step to only two input images. This new method is presented in section 4.3. Finally, this approach is combined with superresolution into a single framework in section 4.4.

4.1 Introduction

Camera sensors are typically far more limited in terms of dynamic range, i.e. ratio between the darkest and brightest parts, than the human eye. Similar to many other human sensors, the eye is extremely adaptable and can detect very weak light sources (e.g. star or moonlight with as little as $0.1\frac{cd}{m^2}$) as well as very bright ones (e.g. outdoor

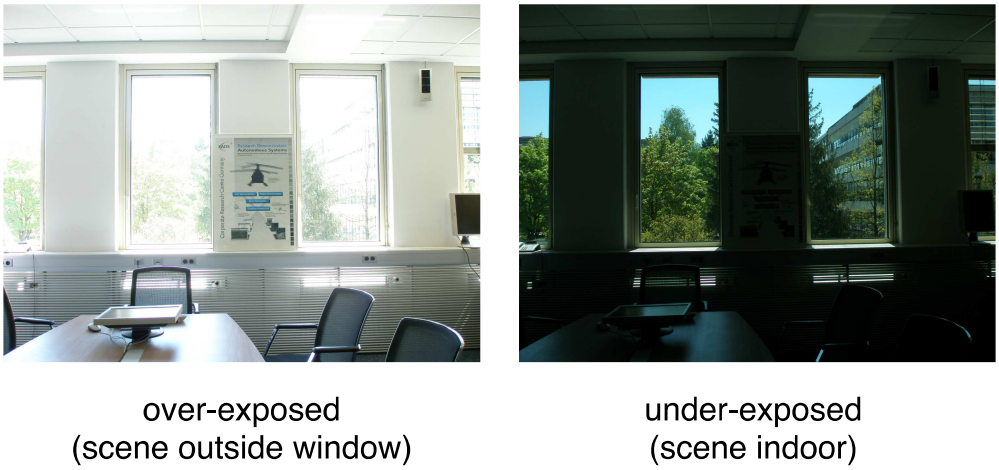


Fig. 4.1: Typical scene with larger dynamic range than the camera can capture, which leads to under- and over-exposed areas.

sunlight with up to $1,000,000 \frac{cd}{m^2}$) [53, 135]. A typical real world scene may expose a dynamic range of about $10,000 : 1$, which can be easily captured by a single view of the human eye. Only very recent imaging sensors are also capable of capturing such large dynamic ranges. Some more extreme, but also very typical natural scenes, e.g. an indoor photography depicting bright light sources such as a window on a sunny day, may even show a dynamic range of $100,000 : 1$. However, most consumer cameras still employ sensors with smaller dynamic range, which often leads to over- and under-exposed areas in an image as illustrated in Fig. 4.1. Therefore, adaptation of exposure time and aperture, automatically or manually, is vital to fully capture a natural scene.

Even though image sensors will continue to improve in the future, there are still applications in which hardware-based image enhancement is not practical. For instance large surveillance systems at airports and train stations usually employ a significant number of cameras, which may easily exceed thousands. For such application scenarios where the replacement of cameras entails changes in other infrastructure components (e.g. data storage, data transmission, power supply or operating software) the hardware layer is considered irreplaceable for a long time period and software-based enhancement is the only acceptable solution. *High-dynamic-range* (HDR) fusion is then a suitable

algorithmic solution. Multiple images that have been captured with different exposure settings are merged to generate an image with larger dynamic range than any of the input images by itself.

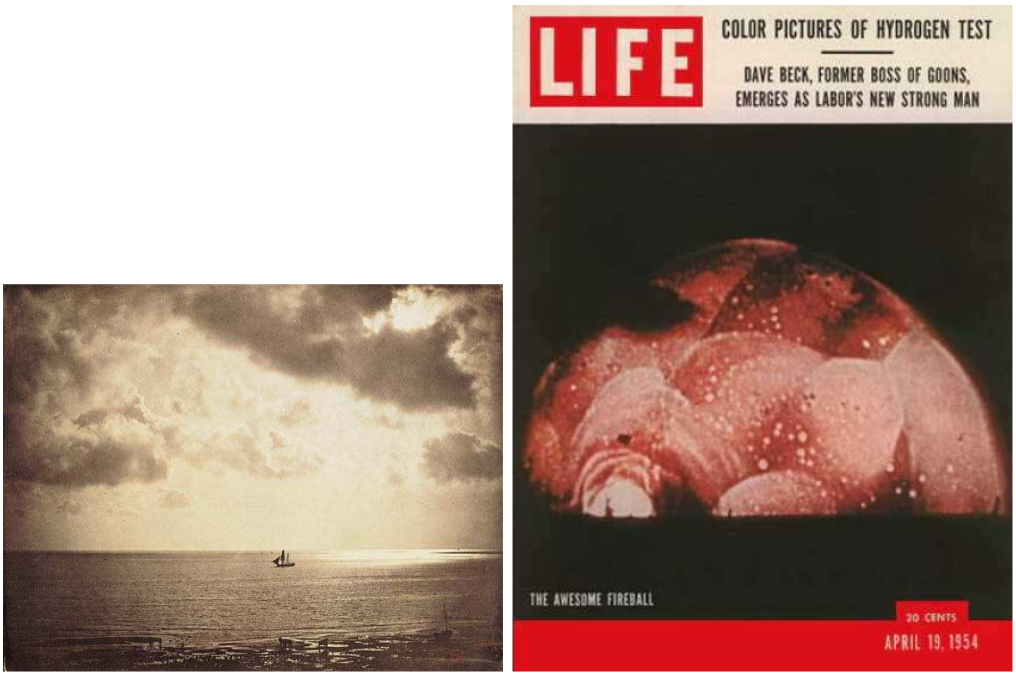


Fig. 4.2: Left: Famous photography *Brick au clair de lune* by Gustave Le Gray taken around 1856/1857. Right: Famous photography of the atomic bomb *Ivy Mike* by Charles Wyckoff presented on the cover of *Life Magazine* in 1954.

4.2 Existing Approaches

Possibly the earliest work on high-dynamic-range imaging was carried out by Gustave Le Gray around 1850. Back then, the films used were far more limited in dynamic range than the camera sensors used today. To be still able to capture bright and dark parts in the sky and on the horizon of the sea, he physically combined two differently exposed images captured with different exposure times [137]. In Fig. 4.2 (left) one of his images is depicted. Later Charles Wyckoff invented a film with multiple layers, which differed in sensitivity to light [111]. Images captured with this special film were

then printed in pseudo-color to visualize all the details. A famous example was depicted on the *Life Magazin* in 1954 (see right image in Fig. 4.2). With the development of digital imaging sensors these very old concepts of image fusion for HDR imaging were rediscovered.

In 1993 Steve Mann discussed for the first time concepts for image fusion, which could be used for HDR imaging [109]. The algorithmic detail was published later in a pioneering paper, which suggested to reconstruct the camera response curve from a set of differently exposed images and then fuse this set to recover the HDR image [111]. A few years later, Debevec & Malik presented a similar approach [38], which mainly differed in the reconstruction of the camera response curve. Many other methods have been proposed until now [110, 136, 64, 135, 157].

To better understand the dynamic-range limitations of current imaging sensors and the principles of high-dynamic-range fusion, it is helpful to recapitulate the imaging process. At first glimpse, photographs taken by recent digital cameras look like a true image of captured scene. However, at a closer look, much information of the real world scene have been lost during the capturing process. In other words they have been digitized or quantized into a very compact description. A real world scene with many continuous properties, such as radiance, is generally captured in form of a 2D array of intensity values laying within a rather limited range (e.g. $[0 \dots 255]$). For some applications such as photo manipulations, this quantization is a very useful property as values can be easily transformed (e.g. via copy/paste or via filters). For other applications, which rely on the realistic capturing of the scene, e.g. reconstruction of the scene via shape from shading, this property is rather a curse.

The digital camera is a quantizer in many respects. In Fig. 4.3 the photometric quantization process of a typical digital camera is schematically illustrated. Incoming light of the scene passes the lens and is already quantized by the optics. The aperture (i.e. the width of the shutter) and the shutter/exposure time (i.e. the time how long the shutter is opened) control how much light will hit the sensor during the capturing of

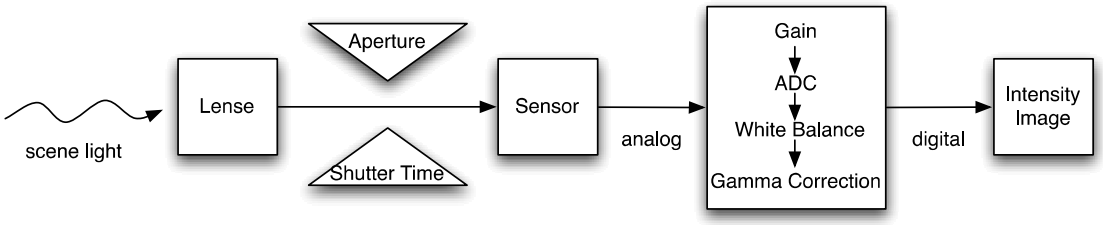


Fig. 4.3: Illustration of the photometric capturing process.

an image. Most cameras have only a few, rather discrete settings for those parameters. Any incorrect choice of those settings, either by the automatic exposure control or by the user, may spoil parts of the image, due to too much or too little light passing through. Once the light reaches the sensors, it is accumulated for a short time interval, also called *exposure time*, of the capturing process. This step also involves a form of quantization as some sensors, depending on their type and quality, are more sensitive to varying amounts of incoming light, than others. The sensor converts the incoming light into analog charges, which are read out and passed through a series of internal filters where the signal is further quantized into digital values. The analog-digital-conversion (ADC) within this pipeline is probably the most dominant source of quantization although very often the digital output is further processed by a JPEG-compressor, which also considerably quantizes the image [157]. By the end of the imaging pipeline much information from the original true scene has been lost due to the different levels of quantizations.

Many applications such as shape from shading or high-dynamic-range image fusion require rather "undigital" images, the radiance values of the scene [64], as Mann & Picard formulate it [111]. To recover the radiance values from the captured digital and quantized intensity images, a function called *camera response curve* or *characteristic curve of the film* needs to be known. This mapping function f summarizes the imaging pipeline explained above and specifies how radiance values X are mapped to intensity values I

$$I = f(X) \tag{4.1}$$

If we were to model the camera correctly, we would need to consider each step of the imag-

ing pipeline described above and specify the involved quantization parameters. However, given the complexity and diversity of those imaging pipelines and the loss of most of the parameters involved, usually a single mapping function f is assumed. The camera response curve of most digital cameras is a non-linear function. Such a non-linearity is often built in on purpose for instance to allow better visualization on media with limited dynamic range (e.g. printed media) [38].

All the standard approaches to high-dynamic-range imaging roughly follow a common procedure. First, the camera response curve is estimated for the camera only once, which can be considered as a calibration step. Second, multiple images with different exposure times are recorded from a scene with high-dynamic-range. Third, these images are transferred into the radiance domain using the inverse of the camera response curve in relation Eq. 4.1. Fourth, a final radiance image is generated, typically using a weighted average of the input radiance images. Fifth, the final radiance image is converted from floating-point values, i.e. high bit depth, to a displayable range, i.e. 8-bit. This step is referred to as *tonemapping*. In our work presented in sections 4.3 and 4.4, we also perform the estimation of the camera response curve and the tonemapping. Therefore, section 4.2.1 briefly summarizes a popular estimation technique, which we adopt. In section 4.2.2 we discuss the tonemapping technique we employ.

4.2.1 Camera Response Curve

The camera response curve is a mapping function, which is specific to each camera and can be estimated once in form of a calibration process. However, the imaging pipeline discussed above (see Fig. 4.3) depends also on environmental factors such as temperature, requiring a more or less frequent re-calibration of a given camera. Since the manufacturers rarely provide the response curve for a camera, many researchers proposed different strategies for estimating it outside the laboratory using images taken by the camera [64]. Different algorithms can be coarsely categorised into *chart-based* and *chart-less* methods. As the name indicates, the first type of algorithms requires the capturing

of a chart with colored patches of known reflectance. The challenge is to ensure uniform illumination of the chart which may be difficult in some cases (e.g. outdoor scenarios). Using the known reflectance values of the color patches and the corresponding intensity values from the captured images, a mapping function can be specified in terms of a look-up table. The second category of algorithms allows a more practical approach as no additional hardware (e.g. color-chart) is required for the calibration, but only the ability to change certain camera parameters (i.e. exposure/shutter time and/or aperture) between different captures. Multiple images captured with varying, known exposure settings can then be used to recover the underlying camera response curve. Because most of the cameras allow the control of the exposure settings we use a chart-less estimation algorithms.

One of the first approaches for chart-less recovery of the response curve was proposed by Mann & Picard [111]. In this work the response curve is model as a parametrized gamma curve

$$f(X) = \alpha + \beta \cdot X^{\gamma}$$

The parameters α, β, γ are estimated using a regression on a 2D intensity histogram H which they call *comparagram*. For two images I_1, I_2 capturing the same scene using different exposure settings, an entry $H(a, b)$ of the comparagram holds the number of intensity pairs at coordinates (x, y) where $I_1(x, y) = a$ and $I_2(x, y) = b$. In this formulation, only two input images are considered for the recovery of the response curve. Another well known alternative of parametrized response curve recovery was proposed by Mitsunaga & Nayar [115]. In this work the authors approximate the camera response curve by n-order polynomials. In their experiments they find $n = 10$ produces sufficiently accurate results. For applications were only limited information about the exposure settings is known (e.g. only coarse estimates), their method proves to be advantageous over others as no exact exposure time values need to be provided.

However, in applications where the camera settings are fully controllable, e.g. by setting exact exposure times, which is possible in most modern digital still-image and video

cameras, there is no reason to just approximate the camera response curve by some arbitrary parametrized model. Instead the full, parameter-free response curve should be estimated which can be easily described in terms of a look-up table [38, 110, 136, 64]. For each of the possible discrete intensity values (e.g. 256 for 8-bit images) a corresponding exposure value is stored. Debevec & Malik were the first to propose such a non-parametric approach [38] and it is our method of choice. The authors exploit a physical property known as *reciprocity*

$$I = f(E \cdot t) \tag{4.2}$$

This relates the captured intensity values (of a low-dynamic range image I) to the true radiance values E and exposure time t. f is the camera response function which can be considered as a non-linear photometric warping function. Reformulating the equation from above by taking the inverse of the response function and the natural log results in

$$\ln\left(f^{-1}(I)\right) = \ln(E) + \ln(t)$$

Considering the log-inverse of the camera response curve as a look-up table ($\ln\left(f^{-1}\right) = G$), then this can be written as

$$G(I) = \ln E + \ln t \tag{4.3}$$

where the outputs of the discrete bins (the bins correspond to the low-dynamic range values, e.g. $[0 \dots 255]$) are unknowns. However, for each exposure pixel, one known low-dynamic range pixel of the input image I exists, as well as the known exposure time t. Hence, when capturing multiple low-dynamic range images at different exposure times, an overdetermined set of equations can be constructed. To solve this system, the authors suggest the following cost-function based on a reconstruction term and a smoothing regularizer using all $N \cdot K$ pixels:

$$O = \sum_{i=0}^{N} \sum_{j=0}^{K} \left[w(I_{ij}) \cdot (G(I_{ij}) - \ln E_i - \ln t_j)\right]^2 + \lambda \sum_{v=L_{min}}^{L_{max}} \left(w(v) \cdot \frac{\partial^2 G}{\partial^2 v}\right)^2 \tag{4.4}$$

where E_i is the *i-th* radiance value and I_{ij} is the *i-th* pixel of the *j-th* low-dynamic range input image with the corresponding exposure time t_j. Solving this least-squares

problem, simultaneously estimates the look-up table G corresponding to the log-inverse of the camera response function f and the radiance E. At the ends of the low-dynamic range this cost-function is not well defined. Therefore, a weighting function is applied to give intensity values at those critical ranges less weight. The authors suggest the use of a hat-function (i.e. a triangle) with its peak at the intensity value 128 with unit weight. To robustly reconstruct the true radiance image, all measurements of the low-dynamic range input images are combined. The authors suggest a pixel-wise weighted average

$$\ln E_i = \frac{\sum_{j=0}^{K} w(I_{ij}) \cdot (G(I_{ij}) - \ln t_j)}{\sum_{j=0}^{K} w(I_{ij})}.$$

4.2.2 Tonemapping

Pixels in HDR images (radiance domain) lie in a range of values far greater than what can be visualized. Tonemapping methods compress these radiance maps into the intensity domain with the focus on maintaining as much detail in the final image as possible while still retaining a realistic look. We employ the gradient domain fusion proposed by Fattal et al. [51] as it produces very good results and allows to boost weak details in the final image even further than other existing methods [99, 97]. The key idea behind this method is to generate an attenuation matrix which indicates for each gradient pixel in the high-dynamic-range image whether it should be amplified or suppressed. A new gradient field is computed from the existing gradients $H(x, y)$ according to the formula [51]

$$G(x, y) = \nabla H(x, y) \cdot \varphi(x, y)$$

The attenuation factor φ modifies the existing gradients $H_k(x, y)$ at different resolution scales k of an image and is computed as [51]

$$\varphi_k(x, y) = \frac{\alpha}{\|\nabla H_k(x, y)\|} \left(\frac{\|\nabla H_k(x, y)\|}{\alpha} \right)^{\beta}$$

From this gradient field G an image is reconstructed using the Poisson equation

$$\nabla^2 I = \operatorname{div} G$$

where ∇^2 denotes the Laplace-operator of the image I that is to be recovered and $\operatorname{div} G$ represents the divergence of the modified gradient field. The parameters α and β control the amount of attenuation of large gradients and magnification of small ones. The reconstructed image is then post-processed by automatic adjustment of brightness, contrast and saturation. Note that the gradients are computed on the radiance values.

4.3 Minimal High-Dynamic-Range Imaging

The existing approaches discussed in section 4.2 propose ways to estimate the intrinsic photometric parameters of the camera from differently exposed images and show how to combine many of them to generate HDR images. Unfortunately, in a practical scenario the acquisition of many differently exposed images takes too much time. This can decrease the accuracy of the image alignment as many scene changes can occur during the acquisition time. Fortunately, even only two images with properly chosen exposure times often contain sufficient complementary information which already represent a dominant fraction of the true dynamic range of the scene (see Fig. 4.1 and 4.4). However, in existing methods the fusion is based on simple weighted average in the radiance domain, which produces unsatisfying results if only two input images are used. This is due to the fact that only one of the images contains the best appearance of a specific pixel and averaging with the corresponding one in the other image only degrades the quality.

Beside the standard approaches, few researchers have addressed the problem of generating high-dynamic-range imaging using a minimal number of input images, i.e. only two extreme exposures. Kang et al. [89] proposed a system which performs HDR fusion on video data. They automatically acquire only two differently exposed images to allow for real-time image acquisition. After a complex offline registration step both inputs are merged with a weighted average to generate an HDR image. Eden et al. [43] address the problem of automatically generating HDR panorama images from input images with large geometric and photometric variations. In contrast to the previous work the final image is reassembled from the input image pixels. They employ a graph-cut

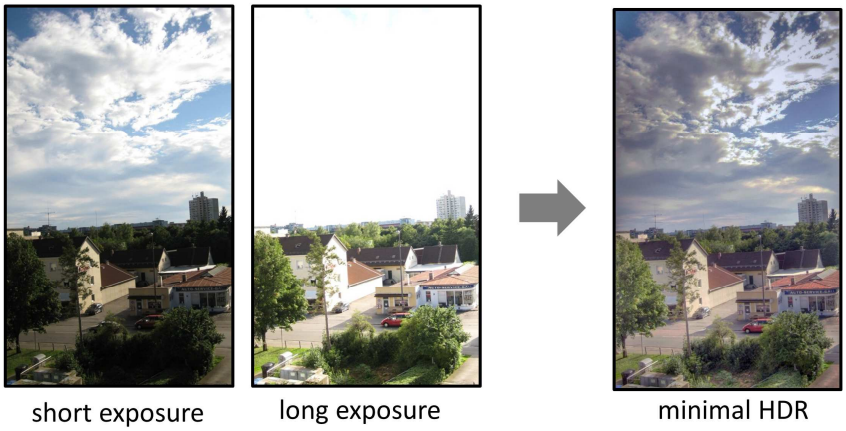


Fig. 4.4: Example for Minimal-HDR: a long exposed image and a shortly exposed image are combined to generate a minimal HDR image.

algorithm to decide which input image contains the best pixel information. We adopt the approach introduced by EDEN ET AL. [43]. A labeling process decides for each pixel in the result image from which input image to copy the intensity information. A simple method would be to generate pixel-wise labels depending on which of the two input intensity values is less saturated. However, this produces very inconsistent masks, which lead to visible seams if the photometric registration (e.g. the camera curve estimation) is not absolutely accurate. To avoid this problem the labels in the neighborhood also need to be considered for every pixel location (e.g. 4-connected pixel neighbors). Thus two neighboring pixels are assigned to different labels only if this does not introduce a visible seams. Unlike [43] our cost function in the labeling process makes use of the camera response function, which accommodates the photometric relations much more accurately.

Similar labeling problems have been faced in other domains, for instance image segmentation, image rendering (e.g. seamless merging of overlapping images), dense stereo and object extraction (e.g. intelligent scissors). To solve the binary labeling problem a

cost-function is formulated as follows

$$F = \sum_{i,j} V(\vec{b}_i, \vec{b}_j) + \sum_i D(\vec{b}_i) \tag{4.5}$$

where $\vec{b}$ denotes the binary label mask, V is a seam-cost and D is a data-cost function that specifies how well the current label fits to the image data at pixel position i. Clearly, the data- and the seam-cost depend on the two input images. The data-cost favors the label of the image with a less saturated pixel value at this location. This aims at generating a final image with as little saturated areas as possible. The data-cost is formulated as

$$D(\vec{b}_i) = \frac{\partial g}{\partial l}\left(\vec{l}_i^{\vec{b}_i}\right) \tag{4.6}$$

where $\frac{\partial g}{\partial l}$ denotes the deviation of the log-inverse g of the camera response function and $\vec{l}_i^{\vec{b}_i}$ denotes the intensity value of image with label $\vec{b}_i$ at the pixel location i. Since g is a mapping from low-dynamic range into high-dynamic-range, steep parts of the curve indicate that a small change in intensity values is mapped to a large change in radiance values. In these areas the dynamic range of the sensor is insufficient to capture the dynamic range of the scene, which results in image saturation. Therefore the deviation of the log-inverse g is an appropriate measure for saturated pixels.

The seam-cost function should favor consistently labeled regions to prevent the final label from scattering. Furthermore, the cost-function should allow neighboring pixel locations (e.g. i and j) to be assigned different labels only when the corresponding intensity values do not introduce a seam. The edge-cost function is formulated as follows

$$V(\vec{b}_i, \vec{b}_j) = |\vec{e}_i^{\vec{b}_i} - \vec{e}_i^{\vec{b}_j}| + |\vec{e}_j^{\vec{b}_i} - \vec{e}_j^{\vec{b}_j}| \tag{4.7}$$

where $\vec{e}_i^{\vec{b}_j}$ is the radiance at position i of the image with label $\vec{b}_j$. If the two neighboring labels $\vec{b}_i$ and $\vec{b}_j$ are equal this cost is zero, hence favoring large areas with homogeneous labels. If the neighboring labels are different then the cost can be considered as a seam measure. The value depends on how similar the two input images are at these neighboring pixel locations. We used the standard graph-cut method described and implemented in Szeliski et al. [158] to solve this discrete optimization problem of binary labeling

with a cost-function F. The resulting label mask determines the source image for each pixel in the final image (see Fig. 4.5 for an example). To generate the final HDR image, radiance values from the input images are transferred to the final image composite. The input images are photometrically aligned in the radiance domain per definition, thus hardly any seams are visible.

Different input images do not have to be perfectly registered for the HDR fusion, because in contrast to previous methods the final image is reassembled from the input images and not generated via a weighted average as done in [38, 110, 136]. The main camera motion needs to be compensated, e.g. using a global 8-DOF homography. Small dynamic scene motion, e.g. a pedestrian passing by or a tree moving in the wind, does not introduce any artifacts since the label mask generation ensures that all pixels for this moving object are taken from only one of the input images.

Fig. 4.5: Example mask (right) corresponding to input (left) and results shown in Fig. 4.15: in black areas pixels are transferred from the image with short exposure time and in white areas the image with long exposure time is used instead.

As the result images are judged by subjective taste, it is difficult to evaluate by quantitative means. Hence, we analyze them only by qualitative means. An example result is depicted in Fig. 4.4. Two images with short and long exposure are fused to generate a minimal HDR image, which clearly contains more dynamic range than any input image by itself. Similar results with different scenery are shown in figures 4.6, 4.7 and 4.8. The images were captured with the camera *Canon Ixus 40*. The camera curve used to generated these results is depicted in Fig. A.13.

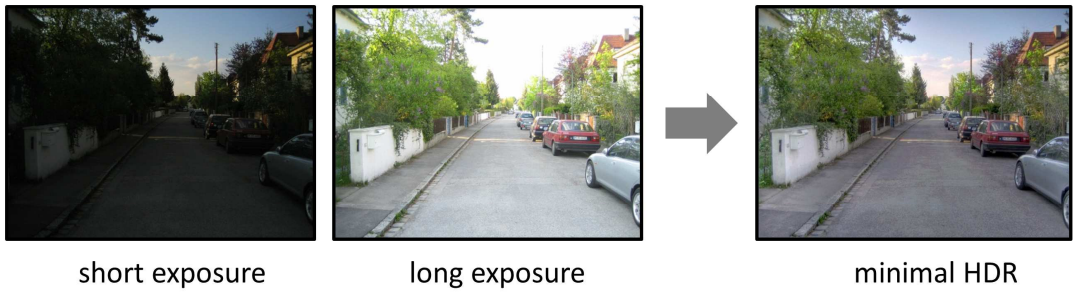


Fig. 4.6: Example for Minimal-HDR: a long exposed image and a shortly exposed image are combined to generate a minimal HDR image.

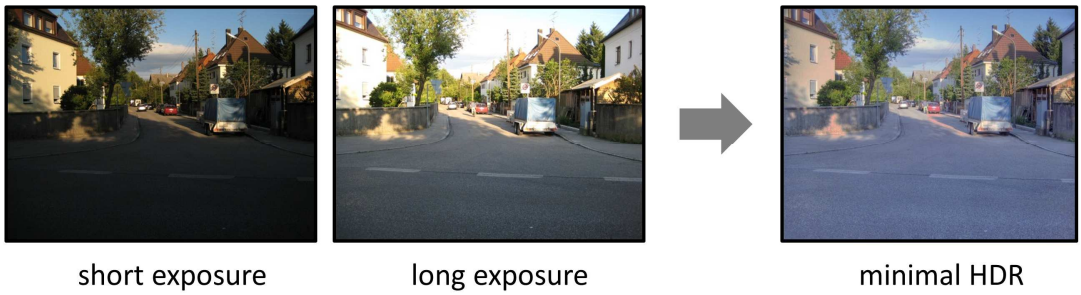


Fig. 4.7: Example for Minimal-HDR: a long exposed image and a shortly exposed image are combined to generate a minimal HDR image.

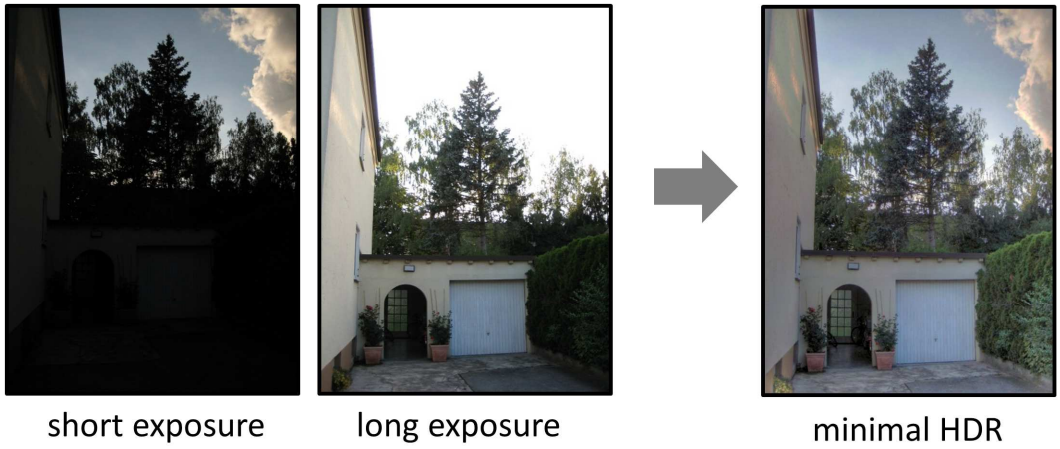


Fig. 4.8: Example for Minimal-HDR: a long exposed image and a shortly exposed image are combined to generate a minimal HDR image.

4.4 Combining Minimal High-Dynamic-Range Imaging With Superresolution

Both image fusion methods presented so far in chapter 3 and section 4.3 are very similar as they fuse the variations of multiple input images to create a result image with better quality. However, they fundamentally differ in the type of those variations. Superresolution works best with input images that can be geometrically aligned with sub-pixel accuracy and show little photometric diversity. This provides redundant information to make the reconstruction of high-resolution images possible. High-dynamic-range (HDR) fusion works best with input images showing only little geometric but large photometric variations. Such images introduce complementary information necessary for increasing the dynamic range of the final image. Despite very active research in the two fields discussed above, little attention has been paid to the relation of both and to how to combine redundant and complementary information in a proper manner.

In this section we propose such a system that utilizes superresolution and high-dynamic-range fusion to generate substantially enhanced images specifically aiming at practical applications such as surveillance cameras. The most related work was done by Gevrekci & Gunturk [62]. They consider the problem of performing superresolution on input images with large photometric variations. Two different methods are proposed to handle such input. The first method uses the camera response curve to photometrically align input images in the intensity domain. It is demonstrated that the method performs better than the typically used affine photometric registration. The second method transfers the input images into the radiance domain where images are photometrically aligned by definition. The superresolution estimation is then also performed in the radiance domain hence producing high-resolution high-dynamic-range images. In both cases the authors consider the problem of how to photometrically register the input images for superresolution rather than how to introduce these exposure variations in a controlled way to perform HDR image fusion. In that respect the high-dynamic-range fusion is not

explicitly addressed in their work. Choi et al. [32] use the same maximum aposteriori (MAP) approach as Gevrekci & Gunturk [62] to estimate a superresolution image from input images showing large photometric variations. However, simultaneously to the superresolution estimation they also compute the camera response curve, although only as a parametrized polynomial. Both of the above approaches perform superresolution fusion in the radiance domain. However, they rely on each input image to capture both, photometric and geometric variation. We argue that these two should be acquired in two separate steps, so that both methods can be optimized independently to achieve best fusion effects. This for instance allows to minimize the number of input images.

In section 4.4.1 we give an overview of our method. In section 4.4.2 we outline how we combine superresolution and high-dynamic-range imaging. In section 4.4.3 we provide details on the acquisition of the input images. Section 4.4.4 discusses the necessary modifications of the iterative-back-projection algorithm introduced in section 3.2.1. Finally, we present the results of our HDR-SR fusion method in section 4.4.5.

4.4.1 System Overview

The complete system includes several steps and is illustrated in Fig. 4.9. First, a calibration is performed to model the photometric properties of the camera, that is the camera response curve. Second, an automatic image acquisition process is implemented for capturing multiple input images with the variations necessary for proper image fusion. The input images are taken under different views and with varying camera parameters such as different exposure times. For the fusion steps, photometric and geometric registrations are performed to establish correspondences between pixels in the input images. Finally, the image fusion with the aim to increase the resolution and to maximize the dynamic range is carried out.

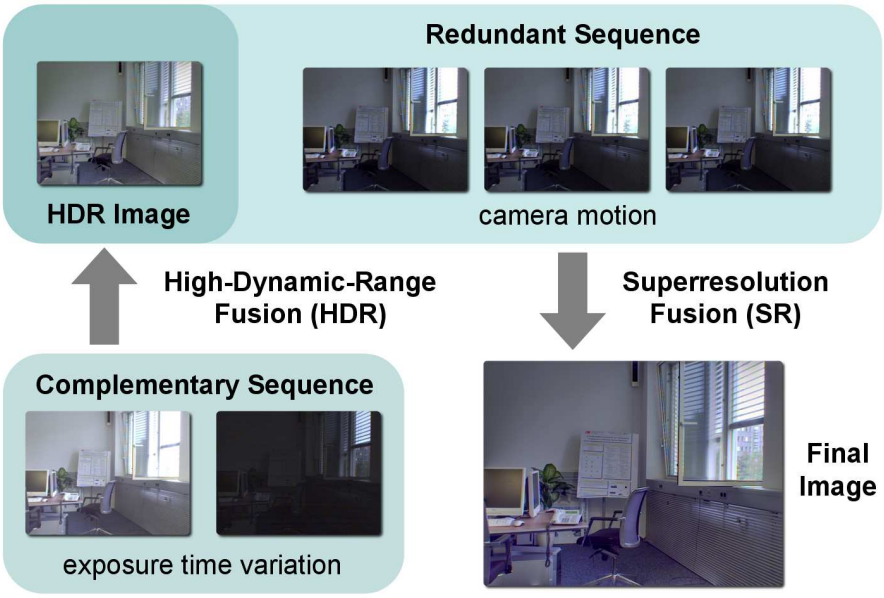


Fig. 4.9: Overview of the system: the HDR result is enhanced by SR to produce the final image.

4.4.2 Image Fusion Scheme

Multiple images need to be fused to generate results that exceed the physical limitations of the sensor, that is to generate images with higher resolution and with a larger dynamic range. Prior to image fusion, a geometric alignment of the input images is required to compensate for the camera motion. We assume scenes with a camera motion that can be described by a 8-DOF homography. This assumption holds for planar scenes, cameras which are far away from the scene and for images taken by a camera rotating around its center like PTZ surveillance cameras. Since superresolution requires sub-pixel accurate alignment, only static parts of the scene can be enhanced in terms of resolution. We employ the dual inverse compositional intensity based approach proposed by Bartoli [12] and discussed in section 2.3.2. After the input images have been aligned to each other, a Minimal-HDR image is computed using the approach discussed in section 4.3. This image is used as an initialization of the superresolution algorithm. The superresolution then enhances only those areas for which the SR input images contain

meaningful pixels (e.g. non-saturated areas). For the saturated parts the HDR image already contains the best information from the differently exposed inputs and hence it is left unchanged. To improve the saturated areas even further, multiple differently exposed input images would have to be used for the superresolution. However, this would significantly reduce the processing speed and complicate the acquisition process. Fig. 4.10 visualizes the contribution of each fusion method to the final image.

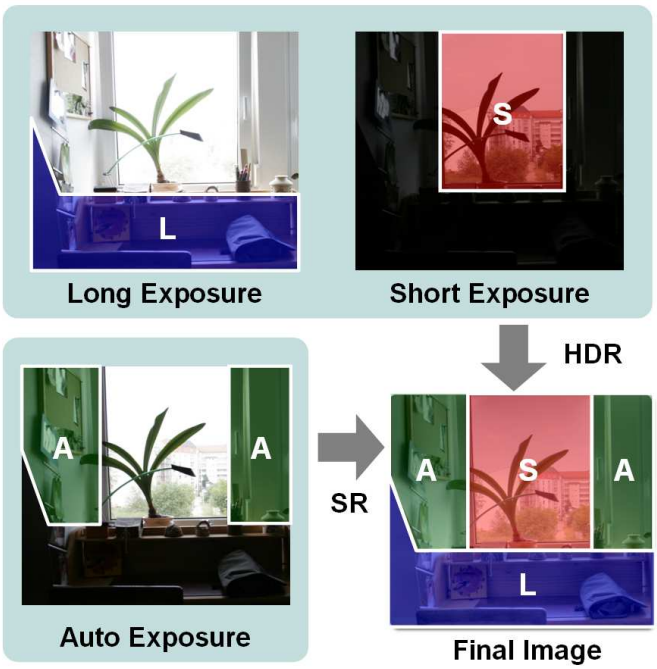


Fig. 4.10: Visualization of HDR and SR contribution to the final image.

4.4.3 Controlled Image Acquisition

For the fusion methods two input sequences are acquired consecutively. The images for the HDR fusion need to capture different intensity ranges. This variation is controlled by taking one image with an exposure time below the auto exposure settings and a second one with an exposure time chosen above these settings. Depending on the camera type this is achieved by programmatic means (e.g. when using a webcam) or manually (e.g. when using a digital handheld camera). In the latter case all necessary information

about the camera settings can be extracted from the EXIF-tags. For the superresolution input, multiple images with auto-exposure settings are recorded. These must be taken from slightly different viewpoints to capture enough information of the scene. Many surveillance cameras are based on pan-tilt-zoom units allowing for programmatic control of capturing such images.

4.4.4 Superresolution In Radiance Domain

Some areas of the previously computed HDR image can also be captured with auto-exposure settings. A number of such low-resolution views can then be used to increase the resolution in these areas via superresolution. Many of the previously proposed superresolution methods [49] could be adapted to enhance an HDR image. We extend the iterative back-projection algorithm (IBP) [78], which was discussed in section 3.2.1, because it can be efficiently implemented allowing for processing of typical consumer images (e.g. resolution 1024×768). A direct extension of the IBP is to apply the algorithm to radiance images rather than to the low-dynamic-range images. We use the upscaled precomputed low-resolution high-dynamic-range image as the initial superresolution estimate and run the IBP algorithm to perform superresolution on radiance values for the non-saturated areas.

In the original IBP algorithm [78] all images are considered to have a low-dynamic-range. Given multiple low-resolution images $\vec{l}$ representing slightly different views, the image rows can be concatenated into vectors and stacked onto each other to form an input image matrix L. Furthermore, assume that registration parameters which map all input images to one reference image are given, as well as the blur of the camera system (defined by the pointspread function of the camera) and the desired enlargement factor. Using these parameters, one can construct a system matrix M that maps a superresolution image $\vec{h}$ onto low-resolution images: $M \cdot \vec{h} = L$. If the true $\vec{h}_{true}$ is used then the observed low-resolution images L and the generated low-resolution images $M \cdot \vec{h}_{true}$ should be equal. One way to solve this system of equations is to use iterative

back-projection. The estimated image $\vec{h}_n$ is refined using the following iterative scheme: $\vec{h}_{n+1} = \vec{h}_n + kB^T(L - M \cdot \vec{h}_n)$, where B is the back-projection matrix that upscales and blurs the difference between the observed and generated low-resolution images. Usually, the initial superresolution estimate is the upscaled average of the input images. This iterative update can be reformulated by substituting the low-dynamic range superresolution estimates $\vec{h}_n$ with high-dynamic range superresolution estimates $\vec{q}_n$ and by using the relation between low-dynamic and high-dynamic-range images shown in Eq. 4.2

$$\vec{q}_{n+1} = \vec{q}_n + kB^T \cdot W \cdot (\exp(g(L) - \ln t) - M \cdot \vec{q}_n) \tag{4.8}$$

where g is the look-up table corresponding to the camera response curve and t is the exposure time of the low-resolution and low-dynamic-range input images. It is impossible to reconstruct the true radiance $\vec{q}$ for a given input image in under- or over-exposed areas. Hence, a difference between the observed low-resolution radiance map $\exp(g(L) - \ln t)$ and the generated low-resolution radiance map $M \cdot \vec{q}_n$ can be caused not only by spatial resolution differences, but also by saturated homogeneous areas. The superresolution estimation then falsely tries to compensate for those differences leading to noisy artifacts. Therefore, to ensure that the superresolution is performed only on non-saturated areas of the initial HDR image, a weighting matrix W is applied during the update. The weights are estimated based on the intensity values of the low-resolution input images and are close to zero at saturated regions or one elsewhere. An example for the weighting matrix is shown in Fig. 4.11.

In summary, our hybrid method combines high-dynamic-range information obtained by Minimal-HDR and high-resolution information obtained by superresolution. The superresolution step enhances only the areas in the HDR image where the low-resolution, low-dynamic range input images provide additional information (non-saturated areas). The parts in the HDR image corresponding to saturated areas in the SR input are left unchanged.

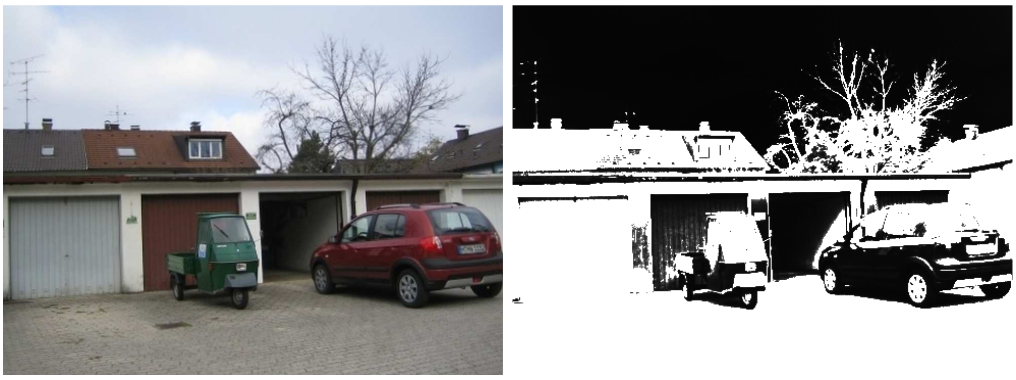


Fig. 4.11: Example mask (right) corresponding to input (left) and results shown in Fig. 4.15: in white areas superresolution is performed.

4.4.5 Results

In this section we demonstrate the performance of the proposed system on images captured with different types of cameras. The images were acquired with an *AVT Dolphin F-145C* (professional video camera), a *Canon Ixus 40* (consumer digital camera), a *Canon Ixus 70* (consumer digital camera) and an *Axis 233D* (professional surveillance camera). The camera curves used to generate the results are depicted in section A.2. In Fig. 4.12 an office scene with a dynamic range exceeding the limitations of the sensor is shown. The images were taken with the *AVT* camera. The left image shows a low-resolution view with auto exposure settings (standard view). The result of the proposed method is shown on the right side of Fig. 4.12. Very saturated parts in the view using autoexposure (e.g. the window) cannot be enhanced by superresolution as hardly any structure is visible (see right side of Fig. 4.13). These pixels are filled in from the HDR image. In areas that were captured with auto exposure settings, the resolution of the image is clearly enhanced as can be seen in the zoomed in views shown in Fig. 4.13. For the superresolution 10 input images are used to increase the resolution by factor 2. Fig. 4.14 shows the superresolution enhancement over the bicubic HDR image. In Fig. 4.15 an outdoor scene was captured with the *Canon Ixus 40* camera. Similar to the previous example, the scene contains very bright areas (e.g. sky) and very dark areas (e.g. opened

Fig. 4.12: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed method (right).

garage), which can not be captured by a single image. Two differently exposed images are merged using the mask shown in Fig. 4.5. The resulting HDR image provides the pixels for the areas with extreme illumination (e.g. sky and opened garage). Superresolution leads to improvements on all other parts, which can be noticed for instance at the license plate in Fig. 4.16. The weighting mask which indicates regions in the HDR image to be enhanced by superresolution is shown in Fig. 4.11. In Fig. 4.18 an example is shown for images taken with the *Axis* camera. The overall quality of this type of camera is usually very poor, introducing color artifacts and jagged edges. Still, an HDR image can be computed which contains less saturated areas than the standard view (left side of Fig. 4.18). Because of low quality of the input images only minimal resolution enhancement is achieved (right side of Fig. 4.18). Further result images compared to the low-resolution input images are depicted in figures 4.17, 4.19 and appendix A.3. In all examples the improvements in terms of both, resolution and dynamic-range, are clearly visible. These images were recorded with the *Canon Ixus 70*.

Fig. 4.13: Zoomed views: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated: left part of left image and right part of right image) and proposed enhancement.

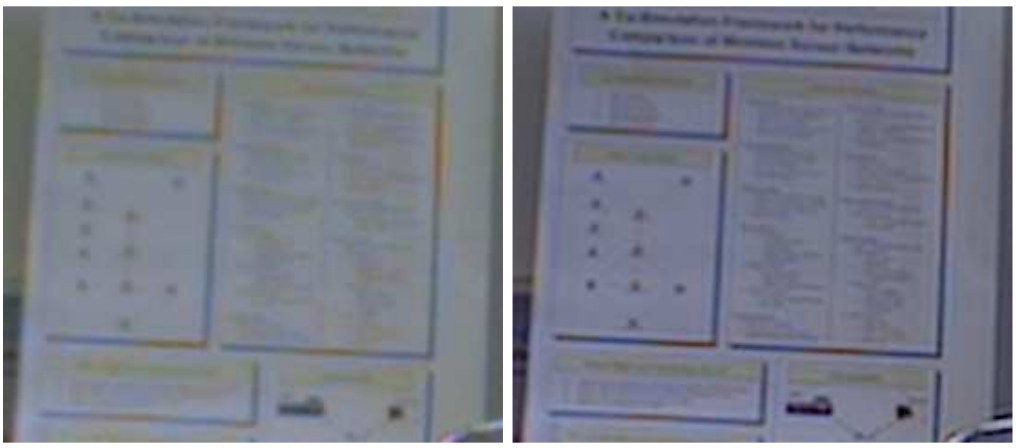


Fig. 4.14: Zoomed views: Comparison of low-resolution bicubic interpolated HDR image (left) and proposed enhancement (right).

Fig. 4.15: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

Fig. 4.16: Zoomed: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

Fig. 4.17: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

Fig. 4.18: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

Fig. 4.19: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

4.5 Conclusions

In this chapter we proposed a new high-dynamic-range imaging approach which only uses two input images recorded with extremal exposure times. We further presented a methodology to combine this Minimal-HDR fusion with superresolution to be used for image enhancement. The methodology builds on a controlled acquisition process that separates the input information into a complementary (image sequence with varying exposure time) and a redundant part (image sequence with camera motion). A two-stage fusion scheme of HDR followed by SR ensures that the most appropriate information contained in either one of these parts is synthesized into the final enhanced image. As the number of required input images is minimized, this system is especially useful in applications where acquisition time is an important factor. The performance of the system is demonstrated on various examples using different types of cameras.

5 Advanced Filter Methods

Seeing Less Is Sometimes More.

Many image enhancement algorithms that increase the level of image details are motivated by the hope that they improve the performance of subsequent computer vision tasks. However, little research has been carried out to investigate the quantitative influence of these enhancement methods on the performance of common computer vision applications. Therefore, in this chapter we present a benchmark which analyzes the impact of advanced gradient-based filtering techniques on image retrieval and scene recognition. We motivate this research in section 5.1 and discuss related work. In section 5.2 we briefly discuss the suitable filtering techniques. In section 5.3 we give details about the implementation of the two computer vision applications. Finally, in section 5.4 we discuss the evaluation protocol and the results on the benchmark datasets.

5.1 Introduction

The underlying assumption of enhancing images prior to retrieval and recognition steps is that the more detail an image contains and the more realistic it looks, the better the performance will be. We argue that altering an image to enhance it can be considered as a form of image filtering, which has a direct impact on the image gradients and hence most types of feature extraction. As most state-of-the-art recognition and retrieval techniques use gradient-based features, applying such filtering as a pre-processing step can significantly change their performance. Although much progress has been made in image

recognition and retrieval over past decades due to intensive studies of feature extraction methods, image representation and machine learning techniques, little research has been carried out on the quantitative influence of such filtering used for pre-processing. Previous works include only basic filtering methods (e.g. blurring) employed in the context of very specific tasks such as face recognition [72, 63, 94] or character recognition [75]. In these studies the filters are applied to make the recognition performance more robust to illumination changes and other noise effects. Some open-source implementations of feature extractors also apply blurring as an initial step, but such pre-processing steps are never discussed in terms of quantitative performance gain in the respective papers. Besides these simple filtering techniques, there exists however a wide variety of more advanced image filtering techniques (e.g. bilateral filtering, cartoon-style or image-based rendering) in the domain of computer graphics which are not commonly used.

In this chapter, we therefore investigate the quantitative differences such advanced image filtering techniques can generally make in respect to the performance of feature extractors and subsequent computer vision applications. We present a performance evaluation of a number of image enhancement or modification techniques in the context of common computer vision tasks such as scene recognition and logo retrieval. To our knowledge this is the first quantitative evaluation of such image filtering for pre-processing. Because image filtering is a data-driven or pixel-wise local process, it is to be expected that the influence of image filtering also depends on the image content. We therefore evaluate using different datasets consisting of images of various categories. For scene recognition we evaluate using the well-known Pascal VOC 2007 dataset [47] which consists of 20 different object categories (e.g. natural scenes, man-made objects, etc.). For logo retrieval we evaluate using a dataset of 30 different logo classes (e.g. Volkswagen, BMW, Coca Cola, etc.) which consists of real images of these logos captured in normal life (i.e. the images were taken from personal and professional photographs downloaded from Flickr).

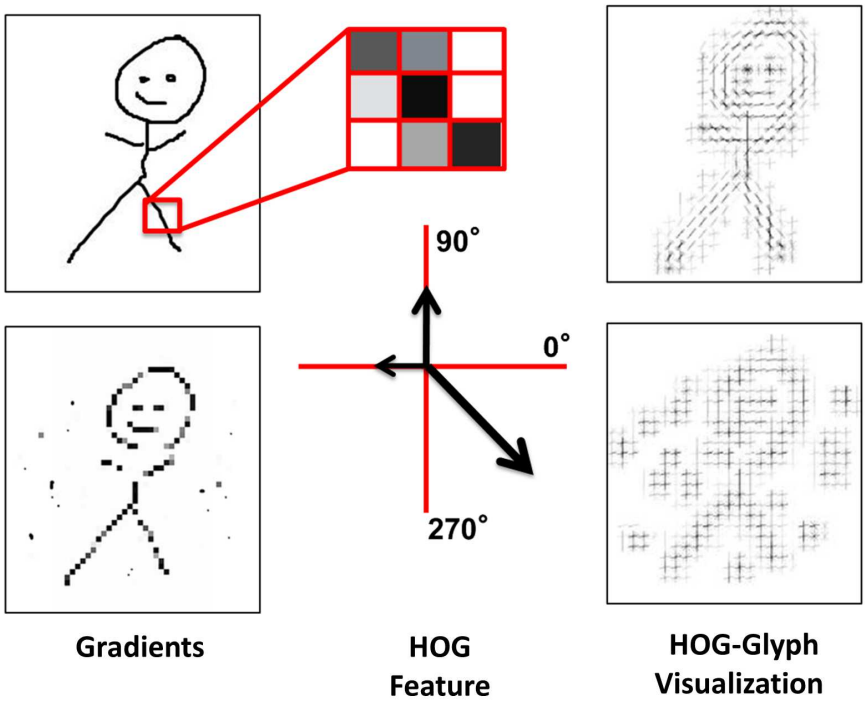


Fig. 5.1: Glyph visualization (right) of HOG features computed on input gradients of a stick-men (left).

5.2 Filtering Techniques

The type of normalization and hence the effect on subsequent steps of the recognition system, in particular feature extraction, strongly depends on the way the filter modifies the image. We focus on three categories of filters:

1. Boosting gradients
2. Suppressing gradients
3. Enhancing color

These types are motivated by the fact that the most successful features in computer recognition applications are based on gradients [47] (e.g. Harris, Hessian, HOG, SIFT, DAISY) and color (e.g. Color-SIFT). For instance reducing the number of weak short

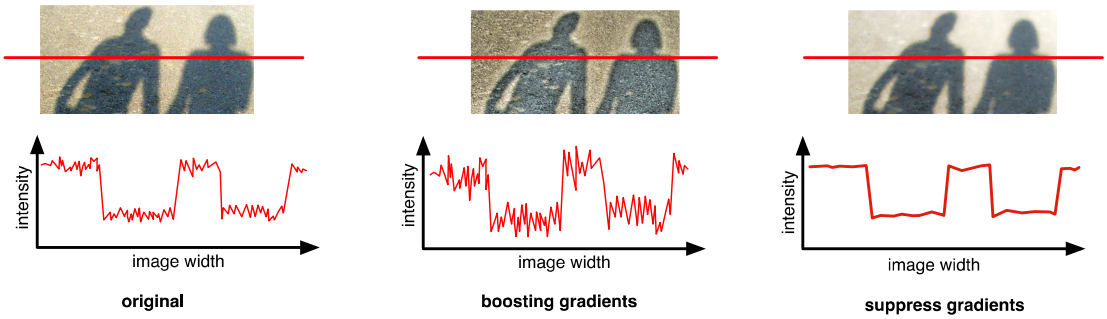


Fig. 5.2: Intensity values (diagrams on bottom) along the line scans across the image (red) are shown for different filters: original (left), boosting (center) and suppression (right) of gradients.

edges, generates cleaner gradient histograms. In Fig. 5.1 two stick-men figures are used as input gradients to compute histogram-of-oriented-gradients (HOG) features. Visualizing these histograms using the glyph presentation indicate that less noisy HOG features can be computed, i.e. capturing the true object, when using fewer and cleaner input gradients. Intuitively, the well defined HOG features (top right) are better suited for machine learning tasks than the clutter ones (bottom right). For each of the filter categories we consider different methods, which we discuss in the following.

To better understand the difference of gradient suppression and gradient enhancement an illustration is given in Fig. 5.2. Intensity values along line scans of an example image are shown. The left plot shows the original intensity values, the center one shows the intensity values after tonemapping, the right one shows the intensity values after applying an abstraction filter. Filters such as tonemapping boost weak gradients and keep the dominant ones unchanged. The filters suppressing gradients such as abstraction filters keep dominant gradients but significantly smoothen small gradients. The choice of the amount of boosting and suppression is clearly arbitrary and depends on the image content and the subjective taste of the user. In Fig. 5.7 examples of the filtered images are depicted for each dataset.

5.2.1 Boosting Gradients

In the computation of HOG or SIFT the strength of the gradient is used to weight the corresponding bins in the histograms. Hence, boosting important gradients can increase their importance in the image descriptors. Furthermore, boosting certain gradients will attract feature detectors to these locations, which can have significant impact on further processing steps. We consider two different variations to increase the strength of gradients:

1. convolution with sharpening kernels
2. tonemapping [51]

The first is a fairly simple and well-known filter. The second one is a more complex filtering technique which is based on a compression, where weak edges are boosted and strong ones are reduced. In the following we discuss the two filters in more detail.

Sharpening The very standard approach to "boost" gradients is a sharpening filter and consists of the following steps:

1. Smooth image with Gaussian
2. Compute Laplace filtered image (3×3) from the smoothed image
3. Add the Laplace-filtered image to the original one using a global weight
4. Rescale intensity values to fit into defined intensity range

Tonemapping was originally designed for compression of high-dynamic-range images which have pixel intensity values larger than 8-bit [51]. These 16- or 32-bit values have to be mapped to the 8-bit range used by most displays. The compression has the goal to boost weak edges and suppress very large ones so that all of these gradients result in

the 8-bit intensity value range. The key idea lays in recovering the final image from a modified gradient field by solving a Poisson equation. [51] was the very first paper which demonstrated such a gradient domain reconstruction. In section 4.2.2 we discussed the details of this method, which can be applied in the intensity domain, except that the input gradients are computed on the low-dynamic range intensity values instead on the high-dynamic range radiance values. The effect of tonemapping (TMO) is visualized in the right column of Fig. 5.7.

5.2.2 Suppressing Gradients

Eliminating only weak gradients results in smooth image segments and cartoon-like stylization. In such images feature detectors extract interest points mainly on dominant image structures. These interest points tend to be more stable under different variations (e.g. pose variations) than if extracted from the original image. This can help a learning process to focus on the important image structures, leading to a better visual recognition despite the loss of information. For instance a car has dominant edges on the boundaries, wheels and joints between core elements such as the doors. However, it also has weak gradients on these core elements due to reflections, textures etc. Eliminating these weak gradients helps to "focus" the feature extraction just on the dominant elements. A similar approach is naturally taken by cartoonists, who first draw the outline of their objects. Very often cartoon-like outlines are sufficient for a human to recognize the object class. We consider four different variations of such abstraction techniques or suppression filters:

1. Gaussian blurring
2. Median filtering
3. Bilateral filtering [165]
4. Weighted-least-squares filtering (WLS) [48]

The first three are often used as pre-processing filters. However, the impact of these preprocessing steps for recognition are rarely discussed or evaluated. The fourth filter is an advanced edge-preserving filter which is superior to the standard bilateral filtering technique.

Blur is a standard procedure to smoothen images and make detection a bit more robust. Standard convolution filter using a Gauss function are applied to an image as described in standard literature [1].

Median is, similar to blurring, a very simple filtering technique which filters out small gradients. This is also a common filter described in standard literature [1].

Bilateral filter is an extension to median filtering as it blurs the image depending on the gradients within a local neighborhood. Different versions for this edge preserving filter exist (see references in [48]). We employ the method described by Tomasi & Manduchi [165] which combines gaussian-based spatial- and intensity-domain weighting to filter a given pixel location. This smoothens homogeneous areas with small intensity variations and preserves edges.

Weighted Least Squares Filter is a very recent extension to bilateral filtering which provides a much better user-controllable parameter set [48]. In this formulation the resulting image is found to be as close to the original image as possible yet under the constraint of smoothing only non-significant gradients which is formulated as a regularizer. The image is obtained via an iterative optimization procedure (i.e. weighted least squares) which minimizes the following cost function C [48]:

$$C = \sum_{(x,y)} \Big([I(x,y) - O(x,y)]^2 + \lambda \left[u_O(x,y) \left(\frac{\partial I}{\partial x} \right)^2 + v_O(x,y) \left(\frac{\partial I}{\partial y} \right)^2 \right] \Big)$$

The first term of the cost function ensures that the resulting image I is visually similar to the original input image O. The second term acts as a regularizer which suppresses gradients along x- and y-direction in the resulting image at all locations where the input image O contains weak gradients. These locations are controlled by the spatially varying weights u and v:

$$u_O(x,y) = \left(\left| \frac{\partial O(x,y)}{\partial x} \right|^{\alpha} + \epsilon \right)^{-1}$$

The formulation for v is analogous to u, except the partial derivative is along y direction. Manually chosen parameter α controls the strength of the smoothing effect. The constant ϵ prevents division by zero. All other locations which contain dominant gradients are left unchanged. The parameters λ and the weight functions u and v control the amount of smoothing. The effect of this weighted least squares (wls) filtering is visualized in the center column of Fig. 5.7. The solution can be computed quickly by converting the problem into a matrix formulation and solving with a standard numerical least squares optimizer. The approach can be easily extended to a multi-scale scheme improving the results even further.

5.2.3 Enhancing Colors

Using color in image descriptors was reported to significantly improve the recognition results [182]. Pictures are often taken with sub-optimal color settings due to simplistic auto-exposure controls and auto-white-balancing. This is further elevated by small, low quality displays used in cameras which do not present the image as it does appear on high quality displays used in notebooks. That is one reason why the user is often not aware of having taken an image with sub-optimal color settings. A post-processing step can help recover or improve the contrast of the image if all details and structures are captured without saturation. A histogram normalization step, which we call *colorboost*, can be used to equalize the colors within an image and bring out much better detail. We use the method described in [74], which consists of the following steps:

1. Find optimal black and white points of gray-scale and color-channel histograms.
2. Rescale gray-scale histogram using optimal black and white points to achieve contrast normalization.
3. Eliminate flat areas in gray-scale histogram to make best use of all 255 bins.
4. Find best saturation factor that balances the color-channels.
5. Rescale color-channel histograms.
6. Recompute color-values of image using new color-channel histograms.

The advantage of this method is that it is completely parameter free and produces excellent results. The filtered color images also show better contrast when converting them to grayscale. As all images in our benchmark datasets are colored, we can therefore evaluate this color-normalization also for descriptors which only use grayscale images.

5.3 Recognition Applications

A straightforward approach to investigate the impact of image filtering techniques, could either consist of a simple toy application (e.g. simple nearest neighbor search within a pool of features) or some heuristic measurements on the feature vector (e.g. variations of individual feature dimensions or intra/inter-class variance). Furthermore, the glyph visualization of gradient-based features could be qualitatively analyzed, such as shown in Fig. 5.3. An input image (top center) is modified by different filter methods, i.e. TMO (left) and WLS (right) and HOG features are computed for all images. The features are then visualized using the glyph representation shown (bottom row). However, in our opinion conclusions drawn from such experiments cannot really be generalized to other realistic applications.

We therefore propose to evaluate the impact of image enhancement on the performance of typical computer vision tasks which use gradient-based descriptors (e.g. SIFT). We

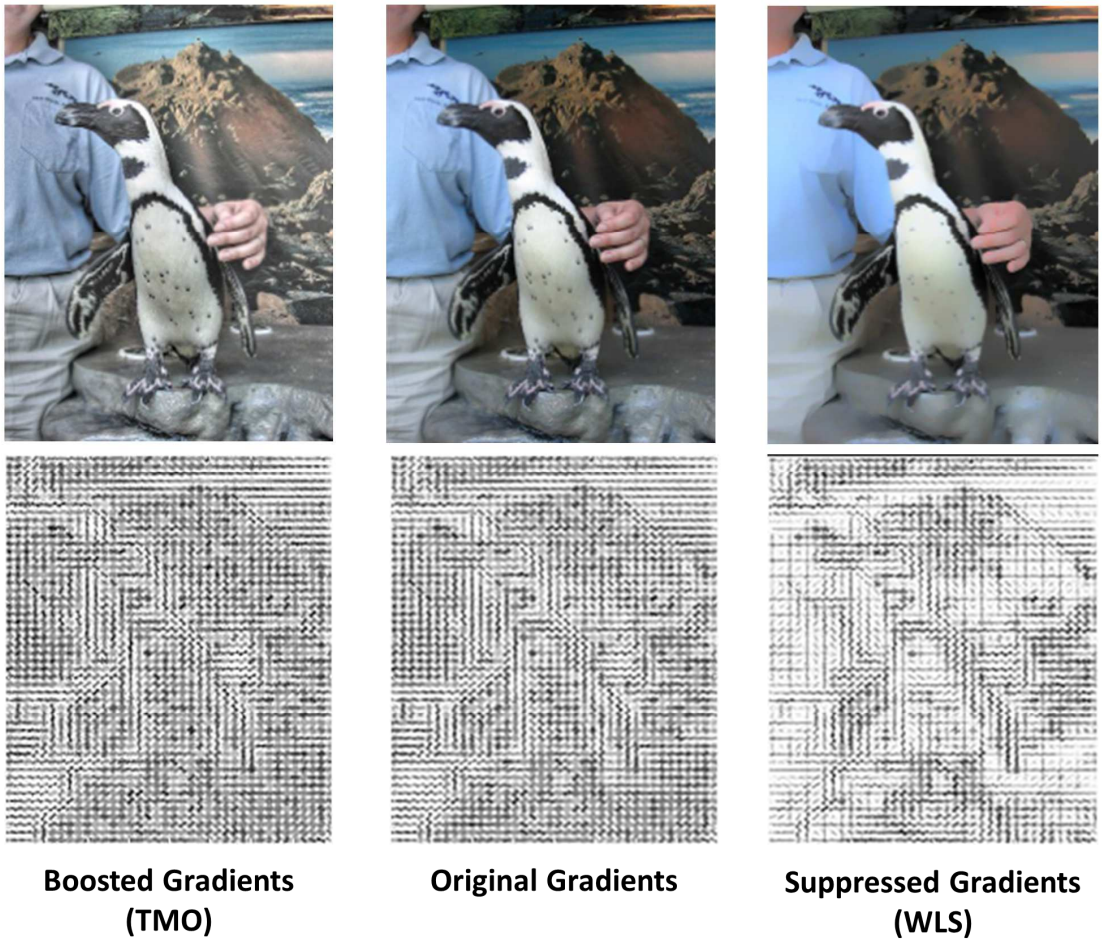


Fig. 5.3: Comparison of glyph visualization of HOG features computed on filtered and unaltered images.

considered two different applications: scene recognition and image retrieval. Both applications share a search task or matching step based on features that are computed. In the first case this matching step is based on a learning process, whereas in the second case a distance measure is used. In the following a brief summary of the implementation of each application is given.

5.3.1 Image Retrieval

We employ a standard bag-of-word retrieval engine such as the ones presented in [150, 81, 128] which has become the state-of-the-art for fast scalable retrieval tasks. For the sake of completeness a brief description of the method is given.

Every image in the database that is to be searched in, is reduced to a single vector $\vec{d_i}$ ($i \in [1 \dots N]$, where N is the size of the database), which represents a compact description for each image. Building image representation is one of the key components of the retrieval and recognition approach. We describe a baseline method in the following paragraph below. The retrieval process itself is summarized in the following. A query image for which the most similar database images are to be found, is also described by a vector $\vec{v}$. The retrieval is then a simple matrix multiplication $D \cdot \vec{v} = \vec{s}$ where $D = [\vec{d_1}, \dots, \vec{d_N}]^T$. Vector $\vec{s}$ now contains the retrieval scores for each image. The smallest scores indicate the database images that best match the query image. For efficiency reasons these vectors and matrices are implemented as sparse vectors and matrices since many of the entries are zero. This increases the retrieval speed (i.e. < 1 second) even for large databases (e.g. $N > 1$ million). The description vectors $\vec{d_i}$ are called *tf-idf vectors* (Term-Frequency-Inverted-Data-Frequency). The more compact and descriptive the descriptors are, the quicker and more accurate search can be performed in a large database. The remainder of this section explains how $\vec{d_i}$ is computed.

In the bag-of-words scheme of [150], feature descriptors (e.g. SIFT) are computed at sparse locations (e.g. using Hessian- or Harris-Laplace detectors) in each image of the database. All descriptors are then clustered (e.g. using K-Means or using random clustering) into k clusters (e.g. $k = 1$ million). All cluster centers form a *codebook* (see Fig. 5.4). Using the codebook the computation of the tf-idf vectors can be done in three steps:

1. compute bag-of-words for all images (i.e. database and query images)
2. compute idf weights (using bag-of-words of database)

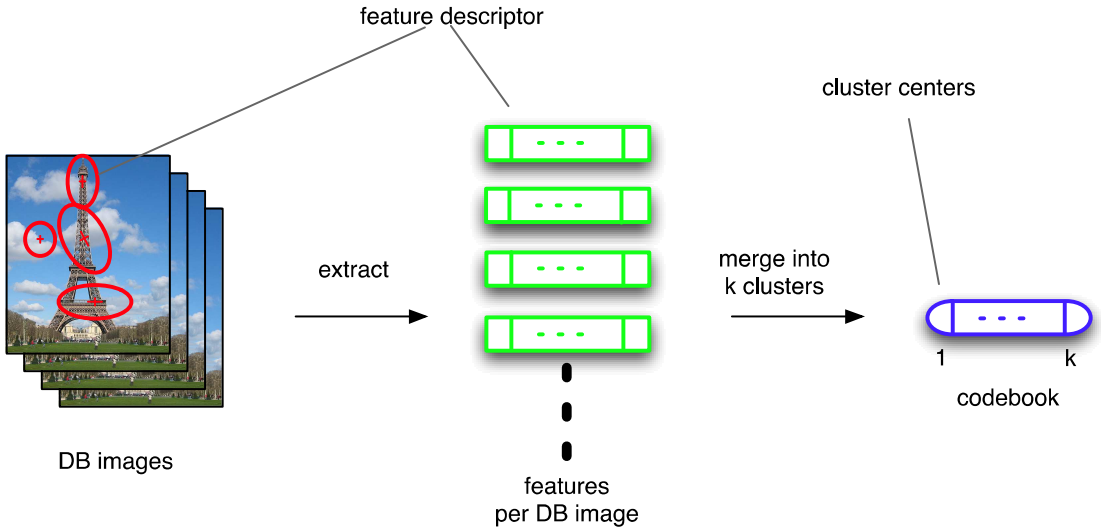


Fig. 5.4: Schematic illustration for generating codebook.

3. compute the tf-idf vectors

The bag-of-word representation is a histogram (i.e. a vector) which summarizes for each codebook entry (i.e. cluster center) how often a feature was assigned to it (using nearest neighbor assignment). All bag-of-words corresponding to the database images are summed up and each element is converted into an idf-weight (inverted-data-frequency weight) using $w_i = \log \frac{N}{b_i}, \forall i \in [1 \dots k]$, where b_i is the sum of all i-th bag-of-word histogram bins and w_i is the idf-weight. The more often a codebook cluster center is assigned to, the less discriminative it is and therefore gets a low weight. These idf-weights have to be computed only once for the database. Each bag-of-words histogram (for all the database images and a query image) can now be converted into a *tf-value* (term-frequency value) by dividing each bag-of-word histogram bin h_i by the total number of codebook entries M that were assigned at all. Multiplying the tf-values with the idf-weights, results in a tf-idf-vector $\vec{d_i}$ where each dimension j is computed as

$$\vec{d_i}(j) = \frac{h_j}{M} \cdot w_j$$

The computation of the *tf-idf vectors* is summarized in Fig. 5.5.

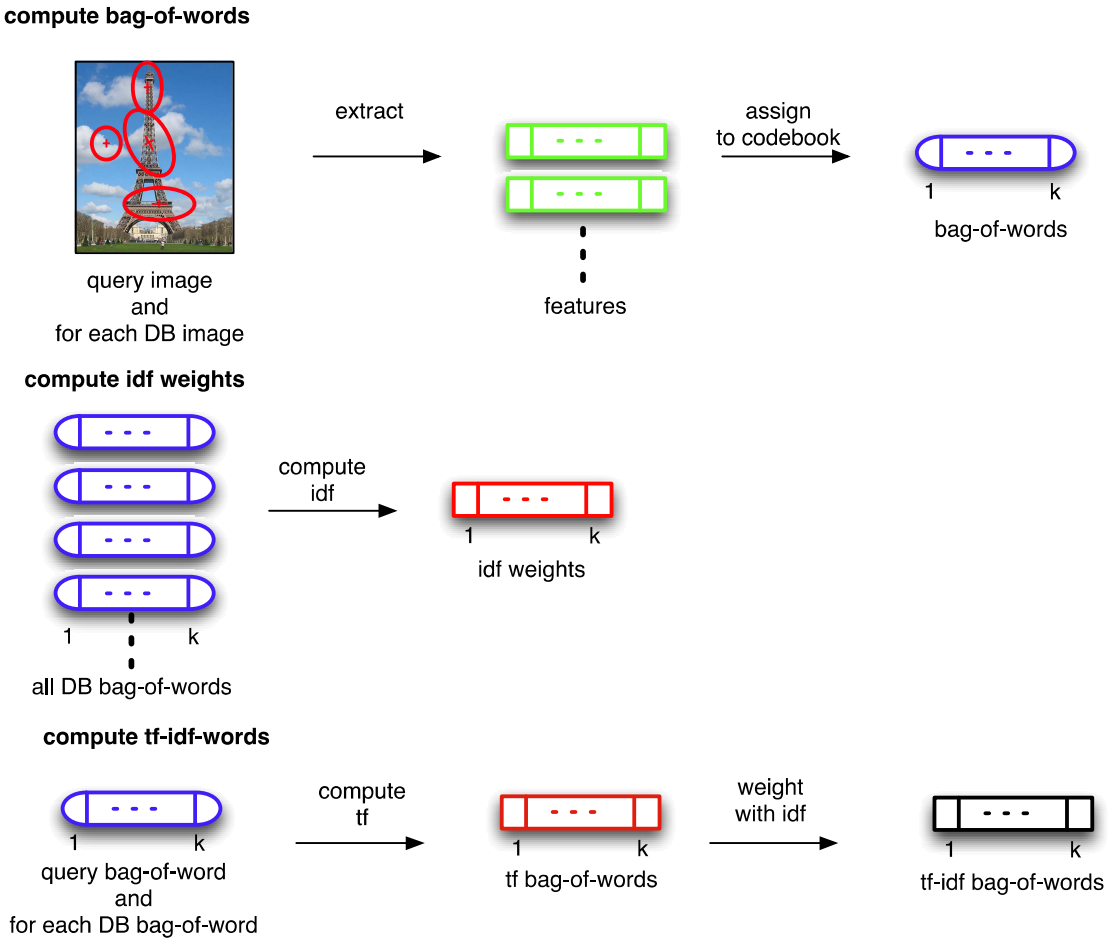


Fig. 5.5: Schematic illustration for computing tf-idf image vectors.

Many powerful extensions have been proposed in the past like enforcing geometric consistency of the matched image features [128], increasing the number of matches by query expansion [128] or by improving the bag-of-words (i.e. substituting the k-means assignment by a richer description [81]), just to mention a few. However the baseline approach is sufficient to demonstrate the benefit of using pre-processing filters. It can also be considered as a core part of the extensions mentioned above, which makes it a representative approach for image retrieval. Therefore we will use this baseline approach to investigate the effect of image filtering.

5.3.2 Scene Classification

Scene recognition is similar to the image retrieval task described in section 5.3.1. The main difference lays in the type of the result: the scene classification results in a label of a class the classified image belongs to where as for the retrieval the result is a list of similar images. The state-of-the-art algorithms for both applications share many parts. We use a standard scheme that has proven very successful [183, 24, 182]. It consists of the following steps:

1. compute local image descriptors (e.g. SIFT, CSIFT, etc.) on a uniform dense grid
2. generate a codebook from the features of all the training images using feature clusters
3. compute a histogram of visual word occurrences for every image
4. compute spatial pyramid match kernel [96] from the histograms
5. train SVM with χ^2 kernels
6. classify new image using combination of multiple kernels

The first three steps use the same methods as described in section 5.3.1 and are illustrated in Fig. 5.4.

5.4 Results

In this section we evaluate the impact of the image filtering techniques on the recognition applications. For classification and retrieval different benchmarks have been established in the literature. We evaluate the impact of the image filtering techniques using standard evaluation protocol of the respective datasets used for the two applications. For the scene recognition the Pascal datasets are considered as the gold standard. We use the Pascal VOC 2007 dataset [47] which contains images of 20 different object categories (e.g. natural scenes, man-made objects, etc.). The Pascal benchmark is considered a challenging test which allows to draw general conclusions about the impact of the filtering techniques.

For the image retrieval task, we collect an own dataset of logos. Typical retrieval datasets (e.g. the Oxford Building dataset [128] or the Flickr1M dataset as used in [81]) either address the scalability of the retrieval task or they are designed for the retrieval of a very specific image, e.g. near-duplicate detection. In the first case the goal is to show that the retrieval engine is capable of finding many images that are similar to the input image but not necessarily belong to the same category. In the second case the datasets contain images of specific objects (e.g. buildings such as in the Oxford building dataset [128]) which have been taken from different viewpoints and the goal is to find all instances of the object shown on the input image.

We would like to generalize the retrieval task further by allowing more variation to the retrieved objects but still ensure that the objects are similar. We therefore consider the task of logo retrieval. Logos can have different appearances (e.g. a painted logo or printed logo) and yet coreespond to the same logo label. An example of this type of variation is shown in Fig. 5.6 for the *peace logo*. There exist a few datasets for logo retrieval however they are either too simple (e.g. only contain synthetic images [80]) or not consistent (e.g. logos have the same label despite its design change over time [85]). We therefore provide a new logo retrieval dataset with 30 different logo classes which makes it comparable in

Fig. 5.6: Illustration of typical variations for logos.

terms of size or variations to the existing datasets (e.g. the Flickr27 logo dataset with 27 logos classes [85]). The results of the evaluation of each application will be individually addressed in the following sections.

5.4.1 Image Retrieval

The retrieval evaluation is done on our benchmark dataset with images that present particular challenge to the descriptors due to various rendering methods (e.g. logo is painted on a wall or carved out of metal) which introduce more appearance variations. In such cases, image filtering is especially expected to aid the matching process. The dataset consists of 30 logos classes from well known brands (e.g. Coca Cola). For each logo 10 random images were pooled out of 1000 images downloaded from `www.flickr.com` using the logo name as the search query. For all 300 images of the dataset the occurrences of the logos are labeled. In Fig. 5.7 two examples of the logo images are shown. An example image for each logo is depicted in Fig. A.20. The retrieval task is to use each labeled logo as a query and retrieve all the other ones with the same label. We use the same protocol for the generation of the index and evaluation of the retrieval performance as in [150]. All filter settings were constant for all images and were chosen prior to running the experiments. In Tab. 5.1 the summarized mean-average-precision (mAP) values are listed separately for the two interest point detectors (Harris-Laplace and Hessian-Laplace) used in the experiment. For each query image an AP value [47] is generated

which is then averaged (mAP value) across all queries belonging to the same logo label. We further average these mAP values over all logo labels to generate a single score for each filter. We can observe that gradient suppression filters, in particular median and wls, improve the retrieval by up to 8%. The performance gain depends on the type of interest point detector, but the general tendency is the same. It is important to note, that the overall performance of $mAP \approx 45\%$ is not very high compared to systems with geometric verification or query expansion [128]. However, in this experiment we are interested in relative performance differences between the filtered and the original images. Although the overall performance across a collection of 30 very different logos consistently improves by using wls filtering, we noticed that certain logo types benefit more than the others. Car logos (e.g. Porsche) which do not vary as much in their rendering form (e.g. car logos are usually printed on badges and not other material like T-Shirts) improve by 58.1% (mAP for "Porsche" logo using original images is 36.2% and 94.3% using wls filtering). Prior to all experiments, we set the filter parameters manually without focusing on increasing the performance but purely on visual appearance to achieve clearly visible filtering effects.

5.4.2 Scene Classification

Improvements in scene classification are evaluated using the evaluation protocol from Pascal VOC 2007 dataset [47]. More specifically the "average-precision" (AP), which is the area under the precision-recall curve [47] is computed for each scene class (20 categories in total) for the original images and for all filtered ones. In this experiment we considered four different filters (blur, colorboost, bilateral, wls). The filters were applied to each training and test image and evaluated separately with constant settings for all experiments. The results are summarized in Tab. 5.2 and compared to a recent alternative state-of-the-art approach [184].

For 16 out of 20 classes in Tab. 5.2 filtered images produce better results than the original ones. Gradient suppression (e.g. bilateral or wls filters) in particular improves the AP

Fig. 5.7: Sample images (original and filtered) from the logo dataset (top 2 rows) and Pascal VOC 2007 (bottom 2 rows).

filter name	Harris (diff)	Hessian (diff)
original	32.4	38.4
bilateral	35.5 (+2.9)	39.5 (+1.1)
blur	33.7 (+1.3)	39.8 (+1.4)
colorboost	33.3 (+0.9)	41.4 (+3.0)
median	35.4 (+3.0)	44.2 (+5.8)
sharpen	29.7 (-2.7)	36.1 (-2.3)
tonemapping	31.9 (-0.5)	39.4 (+1.0)
wls	40.4 (+8.0)	46.9 (+8.5)

Table 5.1: Mean-Average-Precision (mAP) listed for each filter and interest point detector. Behind each mAP score, the difference to the original (top row) is given.

performance by up to 6%. This can be explained by the elimination of weak, noisy gradients using abstraction filters such as bilateral filtering. For instance, many of the images in the class "bird" were captured with background such as vegetation and nature, which contain many fine detailed gradients that are irrelevant for the classification. Focusing the descriptors on dominant gradients (e.g. stems from trees and not the leaves, bird shape and not the feathers) helps to discriminate these images. In Fig. 5.8 a zoom on the original and bilateral filtered image of the class "bottle" is depicted. This class increases by 4.7% when suppressing gradients. The zoom shows that especially noisy compression artifacts are reduced by this filtering

We note that an automatic choice of the best performing filter would be required for practical applications. However, in this experiment we are more interested on the quantitative performance differences, which indicate how much mAP can be gained by a good choice of image filtering for preprocessing. The filter parameters were manually chosen prior to all experiments without focusing on increasing the performance but purely on visual appearance to achieve clearly visible filtering effects.

class	filter	original	diff	filter name	GS-MKL [184]
aeroplane	64.9	64.4	+0.5	bilateral	79.4
bicycle	56.2	52.9	+4	wls	62.4
bird	43	37	+6	bilateral	58.5
boat	55.5	52.5	+3	colorboost	70.2
bottle	19	14.3	+4.7	bilateral	46.6
bus	43.4	43.1	+0.3	colorboost	62.3
car	69.4	68	+1.4	bilateral	75.6
cat	45.4	46.4	-1	colorboost	54.9
chair	42.4	41.6	+0.8	bilateral	63.8
cow	23.9	21.8	+2	wls	40.7
table	31.9	29.5	+2.4	bilateral	58.3
dog	35.8	36.1	-0.3	colorboost	51.6
horse	64.6	65.2	-0.7	colorboost	79.2
motorbike	52.6	49	+3.6	wls	68.1
person	78.7	77.8	+0.9	bilateral	87.1
plant	22.6	18.6	+4	bilateral	49.5
sheep	26.6	28	-1.4	bilateral	48.8
sofa	33.7	32.6	+1.1	blur	56.4
train	64.3	63.2	+1.1	bilateral	75.9
tv	39.9	39.2	+0.7	colorboost	54.5

Table 5.2: Comparison of recognition performance (AP) on VOC 2007 using best performing filter and original images. In the fourth column the difference of AP between filtered and original images are given. For comparison best state-of-the-art results from [184] are listed.

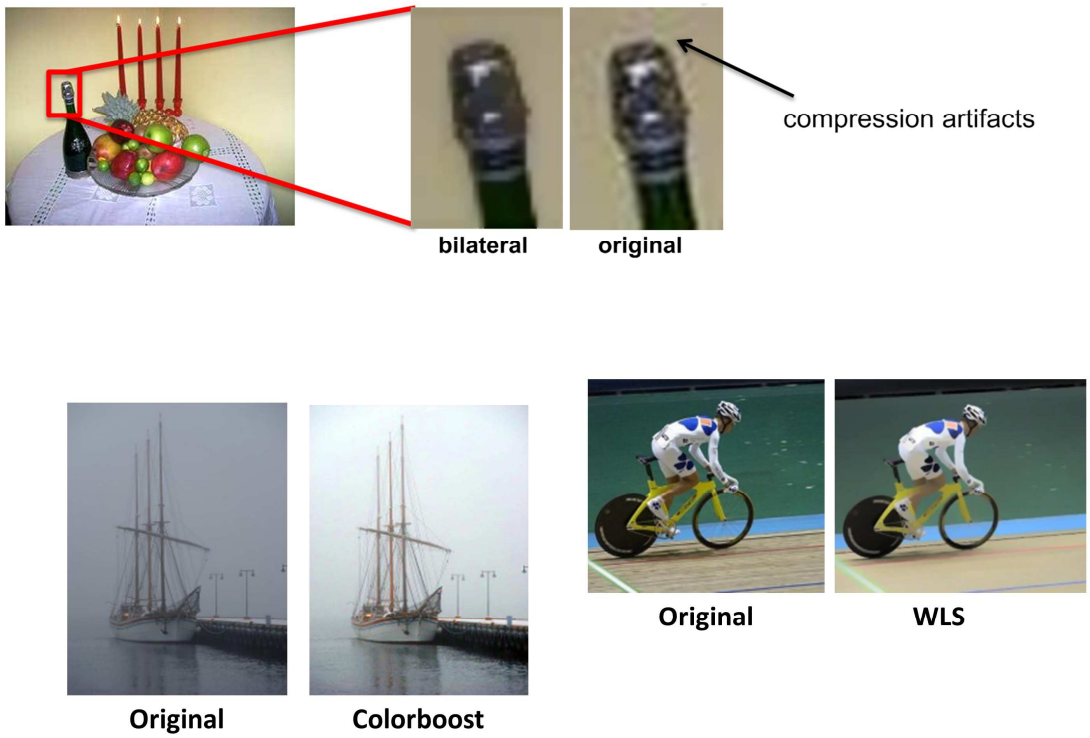


Fig. 5.8: Zoom in on image from class "bottle".

5.5 Conclusions

The results from the evaluation indicate that image filtering significantly improves the logo retrieval by up to 8% mAP and the scene classification performance by up to 6% AP. Furthermore, we observe that the amount of improvement and the type of best performing filter depends on the image category, e.g. natural scenes, synthetic images. However, in most cases image abstraction, i.e. median, bilateral and WLS filtering, improves the performance compared to using original images. For the retrieval task we also notice that certain types of logos benefit more from filtering than others.

6 Moving Object Detection

Image Registration Is Never Perfect.

In this chapter we propose approaches that address the two main challenges in multi-frame motion detection. First, we demonstrate that the novel image registration method, presented in section 2.4, robustly copes with a large variety of aerial imagery and registers images with sufficient accuracy. Second, we present a simple yet efficient filtering step to reduce incorrect motion hypotheses that arise during background subtraction, making the extraction of moving objects more accurate. We evaluate the registration and motion detection on a large dataset. In section 6.1 we introduce the general problem of motion detection. In section 6.2 we discuss the related work. The outline of our approach is presented in section 6.3. Details to the registration and filtering parts are given in sections 6.4 and 6.5 respectively. Finally, in section 6.6 we present results for our motion detection approach.

6.1 Introduction

The detection of moving objects is a functionality which is used in many surveillance and video analysis applications [133, 68] and can be realized with image fusion techniques. Given multiple image frames, which are temporally related, the task is to identify regions in the images that have moved within the temporal window used for analysis. Only pixel changes that are due to foreground object movement are of interest while changes due to noise or background motion (e.g. tree waving in the wind) should be discarded. In our

Fig. 6.1: Examples images taken from aerial platforms. Images can be recorded from low (left) or high (center) altitudes resulting in different viewing angles and object sizes. Viewing objects from far distances and fast camera motion may produce artifacts such as motion blur in the images (right).

application we are interested in the detection of moving objects from aerial platforms. Common challenges are a large range of possible displacements of a moving object between two recorded video frames, the diversity of the camera motion, which has to be compensated, and various types of noise (see Fig. 6.1 and Fig. 2.8 for a few example images).

In general the displacement in pixels of a moving object within video frames is influenced by 3 factors: velocity of the moving object, velocity of the camera and frame rate of the camera. To separate a moving foreground object from its background, the relative displacement across consecutive frames between the foreground objects and the background is important. Too little displacement makes it difficult to detect a motion change event. With too much displacement it is challenging to link the motion changes to the same object while processing multiple frames. The velocity of the object influences this displacement of the foreground object relative to the background. The velocity of the camera does not influence the relative displacement, but it limits the number of consecutive video frames that capture a specific scene point. If multiple images are used to detect moving objects, a camera plane flying too fast might not capture the targeted scene point with enough images. The framerate of the camera has the same effect, i.e. some frames can be skipped. It also influences the displacement of the moving object

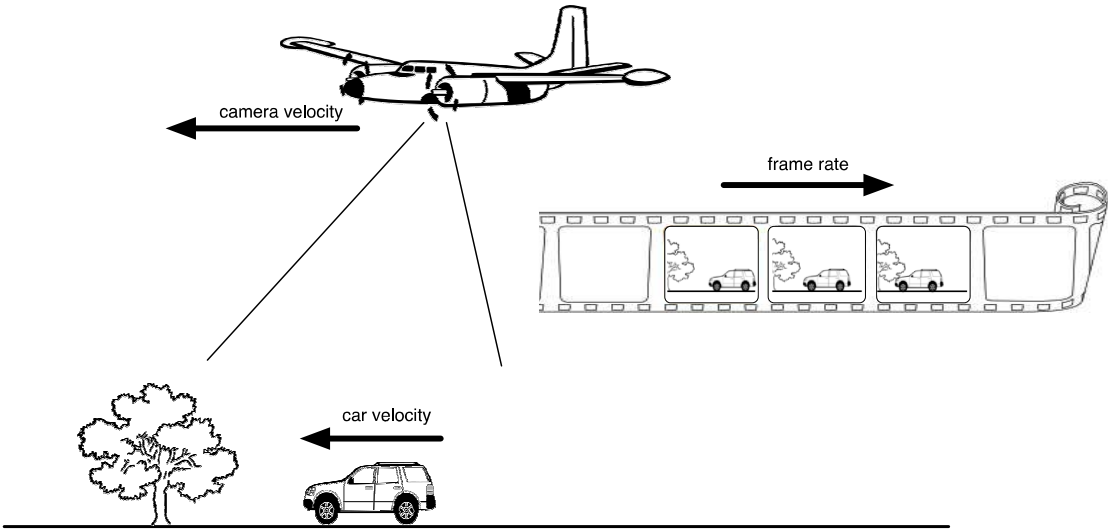


Fig. 6.2: Illustration of influences on the displacements of background and foreground. The displacement of an object is influenced by the camera and object velocity as well as by the frame rate used for recording.

within two neighboring frames. These aspects are visualized in Fig. 6.2.

Simple approaches are based on *frame-differencing*, where two consecutive frames are registered to compensate the camera motion and subtracted from each other to detect motion in the scenery. Thresholding this difference image results in a binary image indicating pixel candidates which correspond to motion. Several heuristic binary operations are usually carried out (e.g. dilating, erosion) to eliminate the ones which result from noise. A connected component analysis combines the pixel candidates to regions with semantic meaning, e.g. belonging to a single moving object. The basic outline of such a simple approach is shown in Fig. 6.3.

Such simple approaches are not very robust and cannot detect slowly moving objects. However, they are very popular due to their fast real-time implementations. More advanced approaches such as *multi-frame motion detection* build up a background map from multiple frames and subtract the current frame from this background image. This allows the detection of fast and slowly moving objects. However, these methods impose two main challenges: accurate registration, which is usually time consuming to

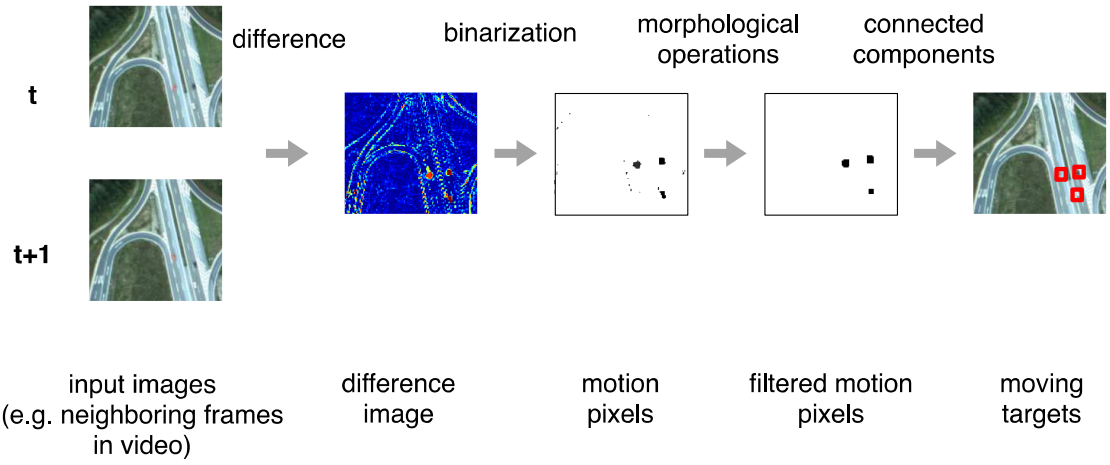


Fig. 6.3: Outline of a simple moving target detection.

compute and extraction of true positives from the difference images, which usually requires tedious, manual parameter tuning. In the following we present solutions to those challenges.

6.2 Existing Approaches

Moving object detection has been of great interest to the computer vision community. Especially due to the numerous applications in the surveillance domain, such as detecting object removal in museums, detecting humans accessing restricted areas or counting traffic on highways, many different methods have been proposed of which a few already were industrialized into products [6]. The fundamental algorithms behind these methods can also be used for other application domains such as medical or forensic image analysis. The surveys in [133, 68] provide a good overview of the broad spectrum of applications and the methods used.

As an alternative to *frame-differencing* described above, *multi-frame moving object detection*, also called *background-subtraction*, is much more powerful due to its sophisticated background modeling using a temporal sliding window of consecutive frames. This approach allows to handle small local changes due to varying illumination or movements of

background objects, e.g. shadows, trees waving in wind. The key challenge lays in the construction of the background, for which many methods have been proposed [129, 123]. As we are focusing on aerial camera platforms, compensation of camera motion is necessary. Registration methods based on local features are widely considered as the gold standard for general applications [157] and in aerial imagery [187]. However, in our application we consider more than 10 images to be registered for every processed frame resulting in many registration operations. Hence, we require a registration which works very fast (e.g. >100 fps), a speed that state-of-the-art feature-based approaches do not currently achieve especially if robustness across a large variety of aerial imagery has to be maintained. There are additional challenges related to aerial platforms, e.g. small objects, motion blur, varying camera motion as well as constraints that can be applied, e.g. scene planarity assumption for registration. The last chapter in [19] provides an overview of the latest state-of-the-art methods for generic motion detection. In [3] a well known moving target detection system called *COCOA* was presented with the application to aerial platforms. It contains different registration methods, i.e. intensity- and feature-based and identifies motion pixels using frame-differencing or background-subtraction. A simple blob-tracker associates different moving objects over time. General moving object detectors make no assumptions about the type of objects, e.g. only cars, to be detected. This flexibility comes at the cost of higher false-alarm rates. However, if the context information about the type and location of objects (e.g. vehicles and roads) or their movement (e.g. direction and speed) can be made, including these as constraints improves detection. In [100, 180] road maps are used to exclude false detections and [2] uses the motion context of neighboring cars to redetect temporally hidden cars. Another assumption that may help to improve the detection is the motion continuity. Tracking methods which use either appearance or statistical models can be included into the moving target detection as a post-processing step to filter out false positives or to identify objects which would otherwise have been missed [181]. Although, in general, the planarity assumption of the scene does hold for the aerial moving target detection, in some cases the altitude may not be high enough causing global registration to fail due to par-

allax effects. A few approaches [87, 189, 187, 41] have addressed this issue by identifying motion pixels which were generated by parallax rather than by truly moving objects.

6.3 System Overview

Our multi-frame motion detection system can be summarised into four main operations: image registration, background subtraction, filtering of motion candidates and extracting the moving objects. Fig. 6.4 illustrates the processing pipeline. In the following we give an overview of each step. The registration and filtering are the main operations, which we describe in detail in sections 6.4 and 6.5.

Image registration For each video frame we consider a temporal window $\mathbf{F}^i$ of size $2k+1$ centred at frame F_i. The first operation is to register all frames $\mathbf{F}^i = \{F_{i-k}, \ldots, F_{i+k}\}$ to the centre frame which is discussed in section 6.4. Using the centre frame as the reference reduces drift errors, which quickly accumulate when concatenating the pairwise homographies. The resulting set of registered images $\mathbf{R}^i = \{R_{i-k}, \ldots, R_{i+k}\}$ is free of any camera motion and the only differences among them are due to moving objects and noise.

Background Subtraction Next, a background image B_i is computed by fusing the registered set of frames with a pixel-wise median. Subtracting every registered image R_j from background B_i and thresholding produces binary maps H_j, which highlight motion seeds, i.e. pixels (x, y) where the frames with moving objects differ from the background. These maps also form a set $\mathbf{H}^i = \{H_{i-k}, \ldots, H_{i+k}\}$.

Motion candidate filtering Not all of the motion seeds actually originate from a moving object but are due to parallax, noise and other artifacts. To remove the noisy seeds we propose an efficient two-stage filtering. First, we enforce a motion continuity constraint for each seed across the temporal set of maps $\mathbf{H}^i$ by using a voting scheme.

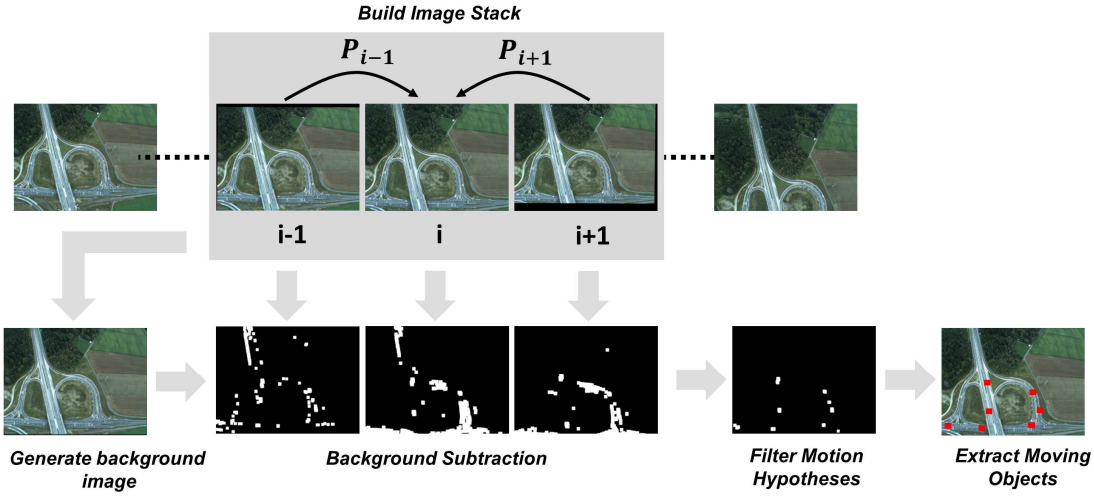


Fig. 6.4: Schematic outline of the moving object detection process.

Second, a forward-backward tracking of the seed's intensity pattern is performed in the registered frames to validate the candidates. The filtering is discussed in detail in section 6.5.

Moving object extraction Finally, morphological operations are applied to filter out remaining noise. We first apply the erosion to eliminate small motion hypotheses, followed by the dilation and connected component analysis to close the gaps within the moving objects. These are straightforward operations, which provide the final moving object detection results in form of a bounding box per object.

6.4 Registration

In contrast to approaches operating on a frame-to-frame basis, multi-frame object detection makes it possible to extract even slowly moving objects. The lower bound for the velocity of the moving objects to be detected is decreased by increasing the number of frames in the temporal window. However, if a temporal window with more than 10 frames are to be registered, the registration method needs to be extremely fast to achieve

real-time processing. We therefore consider algorithms that benefit from hardware accelerated implementations and inherently allow parallization. Futhermore, due to varying backgrounds in aerial imagery, a method capable of registering a large variety of different images is needed (see Fig. 2.8 for images we are considering). This ranges from images with very small structures (e.g. aerial imagery of a dune), where it is very difficult to find reliable features, to images containing very repetitive patterns (e.g. aerial imagery of a forest). Although local feature-based approaches are more often used for homography estimation, given the challenges discussed above, we demonstrate that the tiled phase correlation presented in section 2.4 is more suitable for fast registration of aerial images.

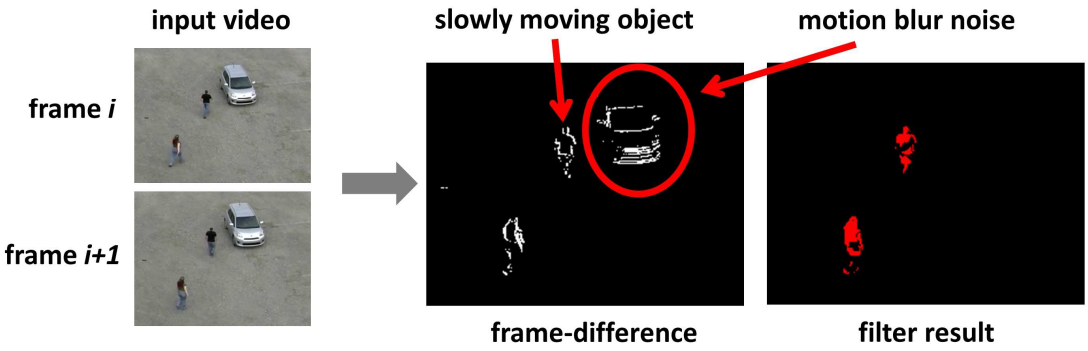


Fig. 6.5: Illustration of filter effect using exemplary result.

6.5 Motion Candidate Filtering

Moving object detection purely based on a difference between two registered frames is not robust. For instance, due to motion blur the contour of static objects will be detected as motion since the blur is not compensated by the homography-based registration. Furthermore the detection of slowly moving objects is very challenging as they overlap significantly between the neighbouring frames resulting in differences only along the contours. An example is presented in Fig. 6.5. Similar effects are caused by parallax and subtle background motion, e.g. a tree waving in a wind. We therefore propose an efficient two-stage filtering approach that addresses the issues discussed above. It is

applied to every motion map H_i, which is the centre of stack of maps $\mathbf{H}^i$ resulting from the set of registered frames with subtracted background. Each pixel in H_i with a value above a threshold is considered as a seed for a moving object candidate and is validated or rejected by the filter, which is explained in the following section.

6.5.1 Voting Filter

In the first stage of the filter, a motion continuity constraint is enforced by a voting scheme. As illustrated in Fig. 6.6, given the coordinates of a seed we place a 2D window as a search area centred on these coordinates in every map of temporal set $\mathbf{H}^i$ to capture all seeds in neighbouring frames that may belong to the same moving object. The size of the window increases with the distance from centre frame H_i in both time directions to allow for a certain motion speed of the object. It is straightforward to calculate the increase based on the expected maximum speed of the objects. A naive approach would be to use every seed s_n in such constructed windows to vote for the seed in frame H_i. The accumulated score would however be unreliable due to noise and other moving objects with trajectories within the larger windows. Ideally, one would like to accumulate votes from the seeds belonging to only one trajectory passing through the seed in frame H_i. To address that we build a 2D voting space V_i of size equal to the furthest window size in frame H_{i+k}. Each seed s_n from the constructed windows votes in V_i space for its own location with a footprint $v(s_n)$ of a size that is inversely proportional to the size of the window. Since the size of the window in nearby frames is small it is likely that these seeds belongs to the same object, therefore footprint $v(s_n)$ is large. However, the further and larger windows may contain more seeds that originate from other objects therefore their footprint in the voting space should be small as illustrated in Fig. 6.6. With this approach, every moving object with a trajectory of seeds in the temporal window will generate one local maximum in the voting space. The largest local maximum is then thresholded to validate the seed in frame H_i.

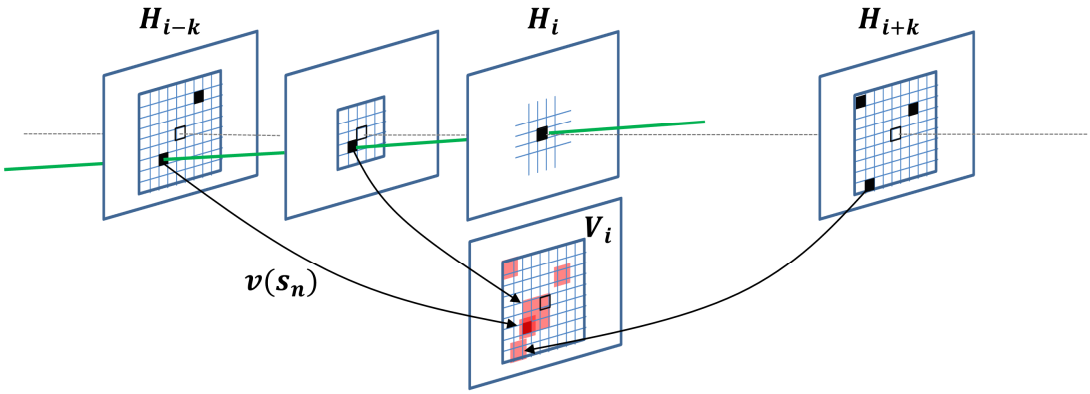


Fig. 6.6: Illustration of voting scheme.

6.5.2 Forward-backward Filter

The second stage of the filter processes the remaining seeds by checking consistency of the trajectories obtained with forward-backward tracking based on optical flow. Note that only very few seed positions have to be processed at this stage, allowing to quickly compute the optical flow. As demonstrated in [84] it is a very efficient and reliable approach to reject erroneous points. Forward-backward trajectories are estimated in both time directions from the seed and only those that differ by small margin m are kept.

6.6 Results

We use the *dataset B* presented in section 2.4.4 for evaluating the proposed moving object detection. An overview of the 10 sequences is depicted in Fig. 2.8. Multiple samples frames for each sequence are shown in appendix A.1. Other existing public aerial sequences, e.g. [169, 15, 143], are not suitable for a benchmark of moving object detection. They either do not show enough variation in terms of view-point changes, image quality and types of scenery or they do not contain annotated moving objects. All

sequences in *dataset B* contain moving objects of different sizes which have been manually labeled to obtain 11379 ground-truth detections distributed across 4184 frames in total.

In order to generate correct background models and motion hypotheses, accurate registration is needed. In section 2.4.4, we show that the tiled phase correlation achieves sub-pixel accurate alignment on the dataset (see Fig. 2.8). The most computationally demanding part in the motion detection pipeline (see Fig. 6.4) is building up the stack of registered images for each frame as more than 10 homographies have to computed. In section 2.4.4 we show that the tiled phase correlation can gain significant speed-up from a hardware-accelerated implementation on a GPU achieving up to 200 fps.

Given this fast and accurate registration, in the following we evaluate the overall performance of the motion detection and the effect of our proposed filtering scheme versus using standard frame-differencing. In the evaluated system the size of the temporal window was $k = [10, \ldots, 20]$, the size of the window in the motion map H_{i+k} as well as in the voting space V_i was 20-50 pixels per dimension depending on the velocity of the moving objects. The forward-backward margin was $m = 2$ pixels.

Tab. 6.1 compares the results for the motion detection using standard frame-differencing method (FD) to our proposed filtering technique (Our). Average recall and average precision with the overlap criterion [47] was used in this experiment. We set the overlap threshold very low at 5% to account for variations in the size of the ground-truth labels. Ground-truth annotation of a very small, e.g. 5×5 px, object is often inaccurate, where one pixel can be more than 4% of the whole object. A small overlap threshold only indicates low accuracy in size estimation, but still allows to identify the true positives of small moving objects with sufficiently high localization accuracy. Typically the users are interested in the location of the moving object, therefore a small overlap threshold is acceptable. The same parameter settings for registration and extraction of objects were used for both methods.

Sequence	1	2	3	4	5	6	7	8	9	10
AP (FD)	70.9	49.9	73.8	35.3	20.2	91.9	64.2	82.5	86.5	64.8
AR (FD)	**85.1**	**70.3**	**66.7**	**90.4**	**55.3**	17.5	48.3	25.1	**64.6**	**52.2**
AP (Our)	**95.9**	**86.9**	**96.3**	**98.0**	**68.7**	**96.8**	**97.5**	**95.8**	**97.4**	**95.5**
AR (Our)	68.8	54.9	31.8	76.0	48.7	**84.2**	**55.1**	**97.6**	56.6	50.3

Table 6.1: Comparison via avg. precision (AP) and avg. recall (AR) of our proposed filtering (Our) and standard frame-differencing (FD) using the same settings for registration and object extraction.

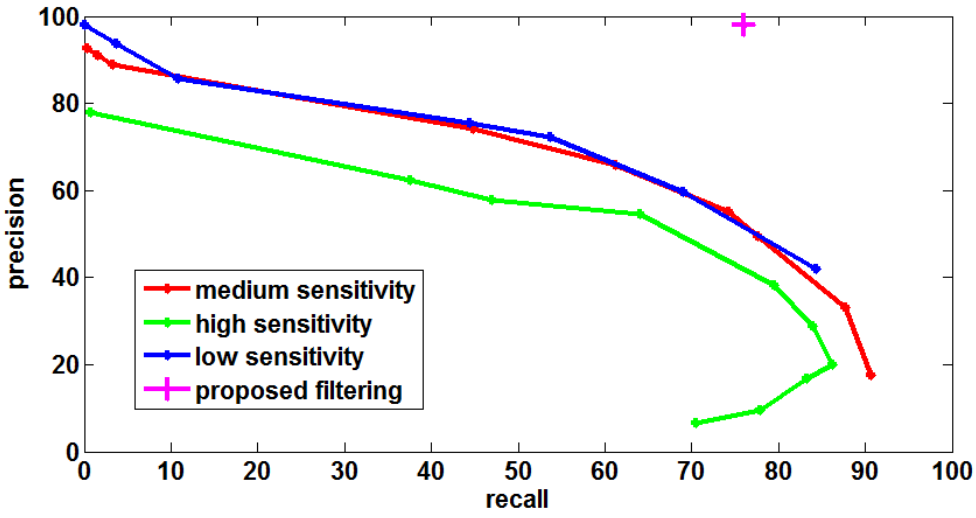


Fig. 6.7: Precision-Recall curves (red, green, blue) for different settings using frame-differencing. Precision-recall for our proposed filtering (magenta).

The critical factor of a motion detector is a low false-positive rate. Otherwise, too many false-alarms occur and the system is not useful in practical scenarios. Hence, we are favoring for our application a system which sacrifices recall for a higher precision. The precision scores, which indicate the false-alarm rate, are significantly higher for our proposed filtering as reported in Tab. 6.1. In Fig. 6.7 different settings were

evaluated for the frame-differencing on sequence 4 producing a precision-recall-curve. Three increasingly sensitive intensity thresholds were used to produce three curves (blue, red, green). For each curve the size filter for extracting bounding boxes from connected components in the difference image was progressively increased. If too small objects are considered as a valid, then the bounding box of an object is fragmented into many small ones which have too little overlap with the single ground-truth bounding box reducing the recall in addition to the precision (see bend of the green curve at precision 20% and recall 86%). Our filtering approach achieves a much higher precision-recall score (see magenta cross in Fig. 6.7) than any other settings of the frame-differencing. This precision-recall value was taken from Tab. 6.1. In some cases, both the recall and the precision are lower for frame-differencing (see sequence 7 in Tab. 6.1). In particular slowly moving objects are difficult to detect when using frame-differencing only. For example sequence 6 (see Fig. A.6) shows slowly walking people, which are detected by our method with a much higher recall.

The following figures 6.8, 6.9, 6.10 and 6.11 visualize the difference images and the detections comparing the frame-differencing method (left column) and our approach (right column). The binary maps (top row) were computed using the same intensity thresholds for both approaches. Similarly, the bounding boxes of the moving objects were extracted using the same parameters. The yellow boxes indicate false positives, whereas red boxes highlight true positives. According to these results, our proposed filtering scheme significantly reduces the registration errors in the binary maps, causing much fewer false positives. In appendix A.5 additional results, leading to similar conclusions, are presented. For each sequence of the *dataset B* results of intermediate steps, i.e. background construction, background subtraction, filtering and bounding box extraction, are shown.

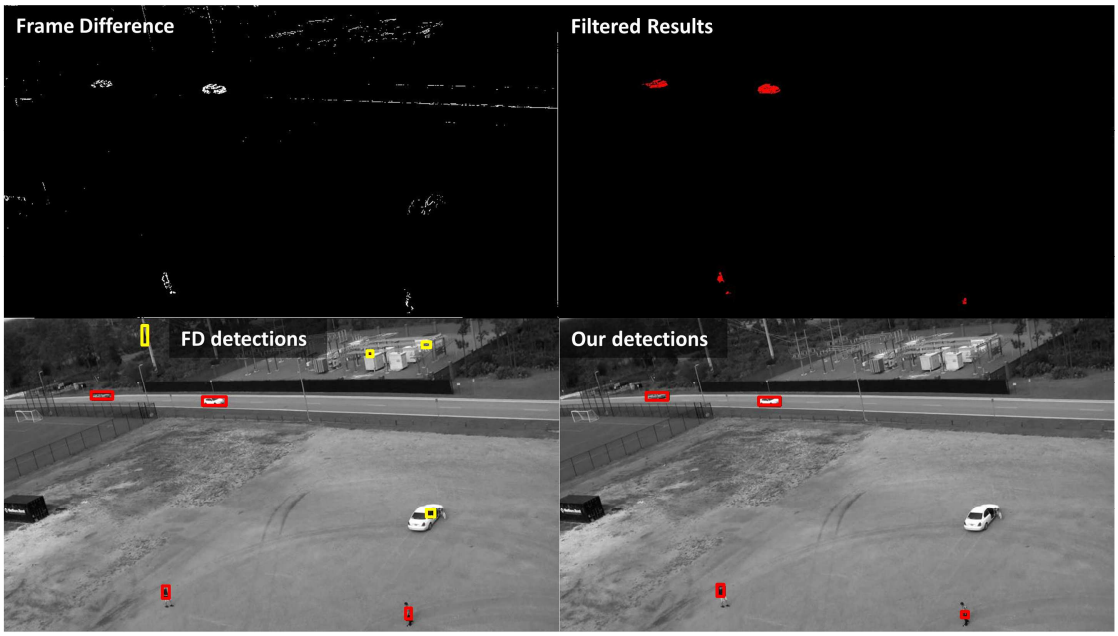


Fig. 6.8: Binary maps and detections of sequence 1 for frame-differencing approach (left) and our approach (right).

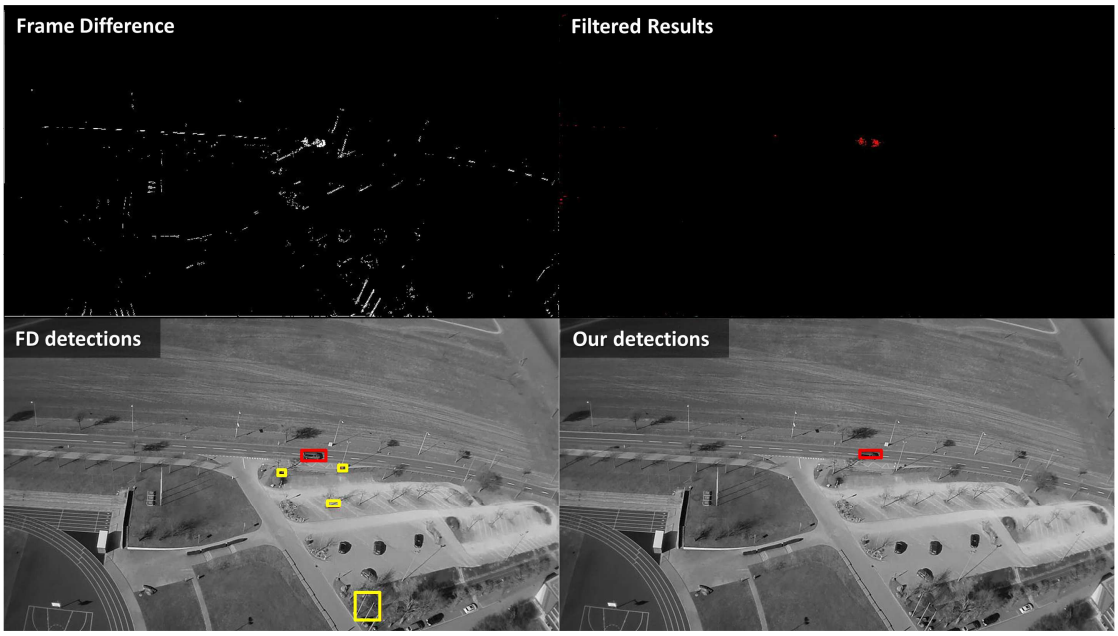


Fig. 6.9: Binary maps and detections of sequence 4 for frame-differencing approach (left) and our approach (right).

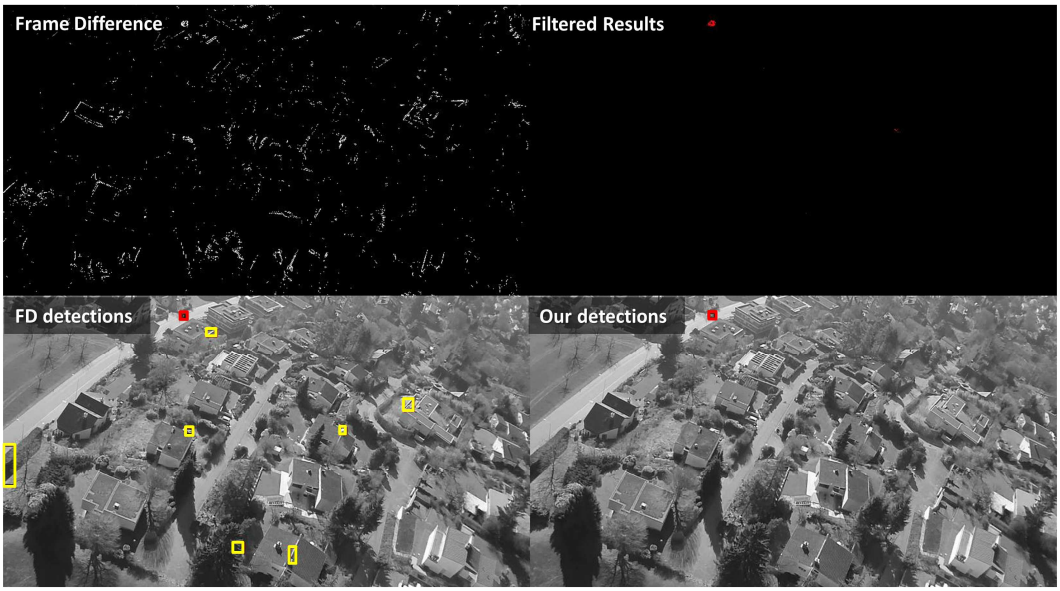


Fig. 6.10: Binary maps and detections of sequence 5 for frame-differencing approach (left) and our approach (right).

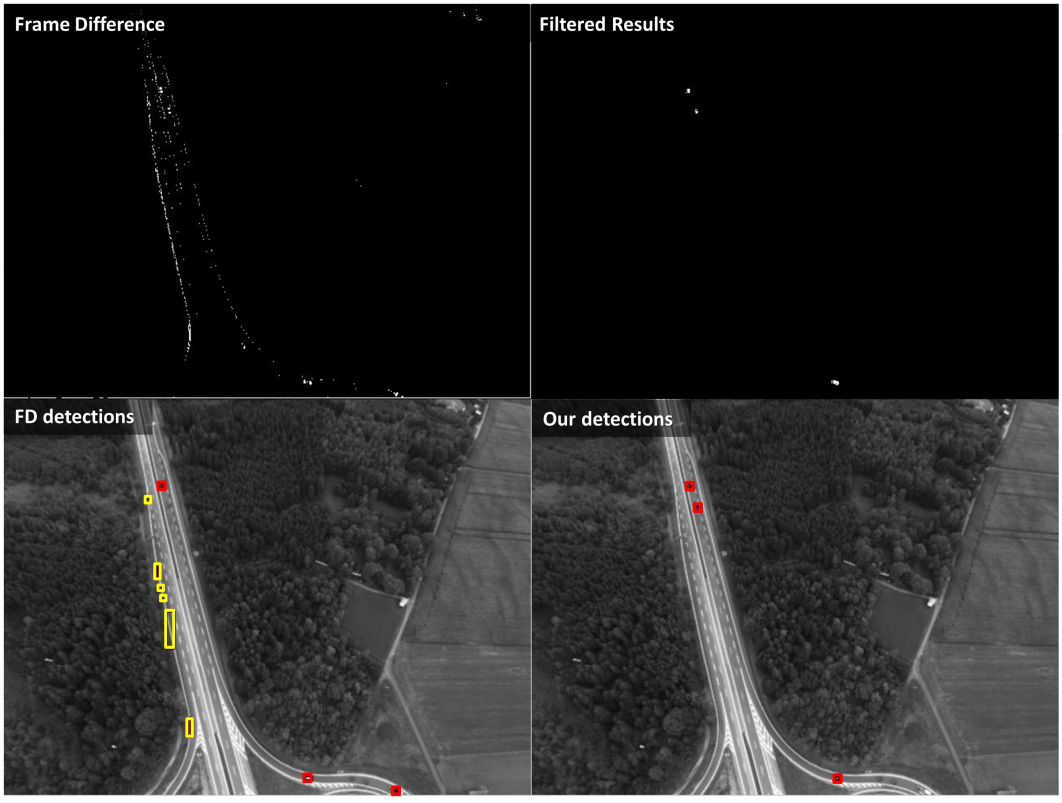


Fig. 6.11: Binary maps and detections of sequence 7 for frame-differencing approach (left) and our approach (right).

6.7 Conclusions

In this chapter we addressed the main challenges in multi-frame motion detection. We demonstrated that the presented tiled phase correlation is able to register more than 10 images per frame at real-time while maintaining sub-pixel accuracy. Furthermore, we showed that registration errors and noise artifacts can be filtered out using the proposed two-stage filtering scheme. As opposed to applying complicated 3D registration methods to avoid parallax errors, our filtering is less complex and can be easily parallelized. We evaluated this algorithm within a motion detection framework using a challenging dataset. The results indicate that especially the precision is significantly increased, while maintaining high recall.

7 Conclusions

Many Images Are Better Than One.

In this thesis we investigated various algorithmic solutions using image fusion to overcome hardware limitations of camera sensors. We considered limitations in terms of spatial resolution, dynamic range and temporal information. We demonstrated that these can be overcome by applying super-resolution, high-dynamic-range imaging and motion detection, respectively. As mobile cameras are not static, they impose the challenge of compensating the ego-motion. Therefore, we investigated for each of the image fusion techniques the required image registration steps and proposed algorithms that are best suitable in our opinion. In the following section 7.1 we summarize the contributions and in section 7.2 we provide outlooks for further research.

7.1 Summary

In chapter 3 we presented a solution to enhance low-resolution videos using multi-frame reconstruction-based superresolution. We showed that high-resolution images, which are retrieved from the Internet, can be used as strong priors within a maximum-a-posteriori formulation. We demonstrated that this superresolution framework increases the resolution of low-quality input videos taken with mobile cameras. We motivated that these high-resolution images can be retrieved from the Internet using powerful retrieval engines such as the ones discussed in section 5.3.1.

In chapter 4 we presented a new high-dynamic-range imaging approach, which we call *Minimal-HDR*. Similar to superresolution, high-dynamic-range imaging is also an algorithmic solution to overcome the physical limitation of a sensor. However, instead of fusing redundant images, rather complementary image information is combined to produce an enhanced image. In order to benefit from both fusion techniques, we presented an approach that combines the iterative back-projection superresolution algorithm and the minimal high-dynamic-range imaging method into a unified framework.

Relating multiple consecutive frames from a video allows to extract pixels belonging to moving objects. This task requires many registration operations, hence demanding very fast registration methods. In chapter 2 we presented a fast algorithm, called *tiled phase correlation*, to compute 8-DOF homographies based on phase correlation. We demonstrated that the method benefits from parallel implementations using graphic cards. Furthermore, evaluation using a dataset with aerial imagery shows that the approach can register images robustly with sub-pixel accuracy. This algorithm was successfully applied for the motion detection approach presented in chapter 6.

Despite accurate registration, errors in the motion extraction process due to parallax and sensor noise are inevitable and will always generate false-alarms. In chapter 6 we presented an efficient filtering scheme, which reduces such wrong detections. We evaluated the approach using a dataset with many different aerial videos. The results show that the proposed algorithm significantly increases the performance of moving object detection. Especially the precision is increased while maintaining high recall, which means that much fewer false-alarms are triggered, making our solution more user friendly than other state-of-the-art approaches.

Many image enhancement algorithms are motivated by increasing performance of subsequent computer vision tasks. In chapter 5 we investigated how much the performance of two common applications, image retrieval and scene recognition, can be increased when applying image filters as a pre-processing step. We evaluated standard and advanced gradient-based filtering techniques using state-of-the-art benchmark datasets. The re-

sults show that reducing the level of detail, e.g. by applying image abstraction filters, actually increases the recognition and retrieval performance.

7.2 Future Work

During this PhD study many interesting further research directions were identified and are summarized in the following.

A main limitation of almost all state-of-the-art and our superresolution approaches is the restriction to planar and static scenes. This is due to the limitation of the image registration to 8-DOF homographies. These are less generic, but therefore much more accurate if the assumption of planar, static scenes holds. More powerful and generic registration methods have been presented and applied to image fusion [18, 67]. In future research these registration approaches should be made more robust to allow their application to reconstruction-based superresolution.

Our current *Minimal-HDR* algorithm processes each video frame individually. This requires costly computations of the labels by solving a Markov-Random-Field using Graph-Cut. As these solutions are computed independently from frame to frame, temporal flickering might be visible. Therefore, in future work it would be interesting to explicitly enforce temporal consistency. This could be achieved by propagating the labels from frame to frame providing an initialization for the Graph-Cut solution.

The current implementation of our *tiled phase correlation* uses uniform layouts for the tiles as no assumption about the viewing angle of the camera can be made. However, in some applications, e.g. forward-looking aerial cameras, it would be interesting to investigate non-uniform layouts which use larger tiles for homogeneous areas like the sky and a finer layout for image parts related to the ground. Such application specific layouts could significantly improve the registration accuracy. Furthermore, a pre-filtering of tiles based on the structure of the image content could help to filter out unreliable motion offsets.

In our current motion detection framework we focused on reducing inevitable registration errors. The filtered motion maps were then processed by a standard bounding box extraction step. In future work it would be interesting to combine this framework with a statistical tracking to further filter out implausible objects. This could be achieved by analyzing their geo-referenced motion.

The evaluation of gradient-based filters applied as a pre-processing step for scene recognition and logo retrieval did not employ an automatic selection process for finding the optimal filter and parameters. In future work it would be interesting to investigate whether machine learning techniques can be used for this selection. Furthermore, including other and notably larger datasets for retrieval as well as considering other applications such as object detection would help to better understand the impact of image filtering as pre-processing on subsequent computer vision tasks.

A Appendix

A.1 Dataset B

The *dataset B* was used in section 2.4.4 to evaluate the accuracy of the tiled phase correlation presented in section 2.4. The same dataset was also used in section 6.6 to evaluate the performance of the moving object detection presented in chapter 6. In the following we depict multiple sample frames and describe the camera motion in the captions of the figures for each sequence to provide a better understanding of the dataset.

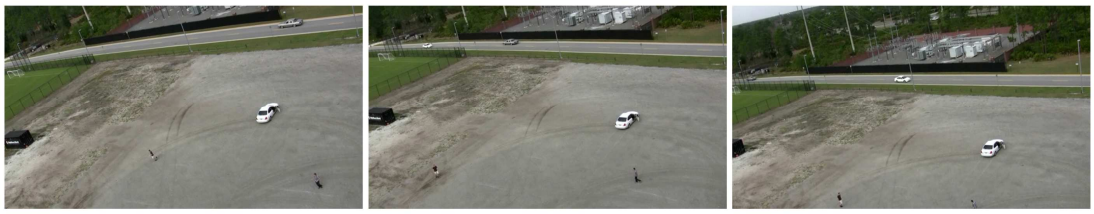


Fig. A.1: Sample frames from sequence 1 of dataset B (camera: translation and rotation).

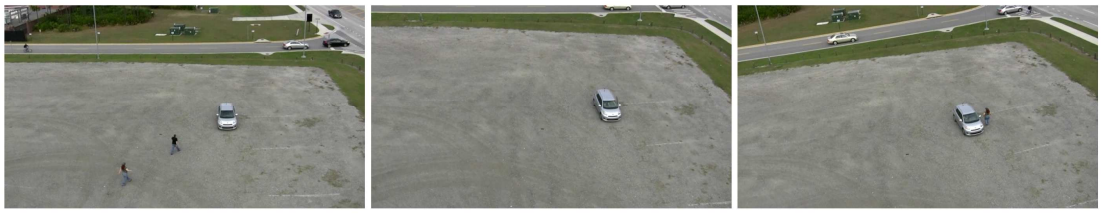


Fig. A.2: Sample frames from sequence 2 of dataset B (camera: translation and rotation).

Fig. A.3: Sample frames from sequence 3 of dataset B (camera: translation and rotation).

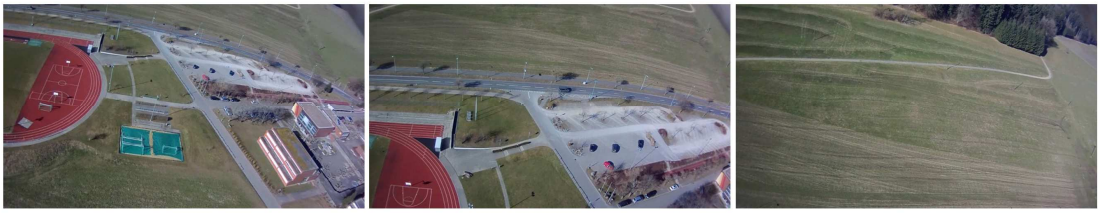


Fig. A.4: Sample frames from sequence 4 of dataset B (camera: forward translation and very little rotation).

Fig. A.5: Sample frames from sequence 5 of dataset B (camera: forward translation and very little rotation).

Fig. A.6: Sample frames from sequence 6 of dataset B (camera: very little translation and rotation).

Fig. A.7: Sample frames from sequence 7 of dataset B (camera: forward translation and little rotation).

A.2 Camera Response Curves

In chapter 4 the camera response curve is required to compute Minimal HDR images as described in section 4.3 or within the HDR-SR framework presented in section 4.4. In the following figures the camera response curves for the cameras used to produce the results shown in sections 4.3 and 4.4.5 are depicted. The curves were computed using the method of Debevec & Malik [38] which is discussed in section 4.2.1.

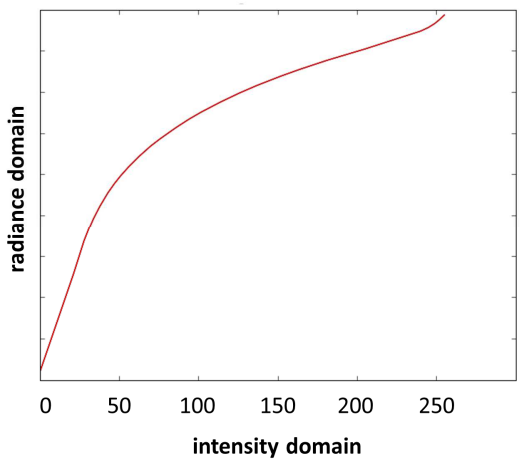


Fig. A.8: Response curve for *Dolphin F-145C*.

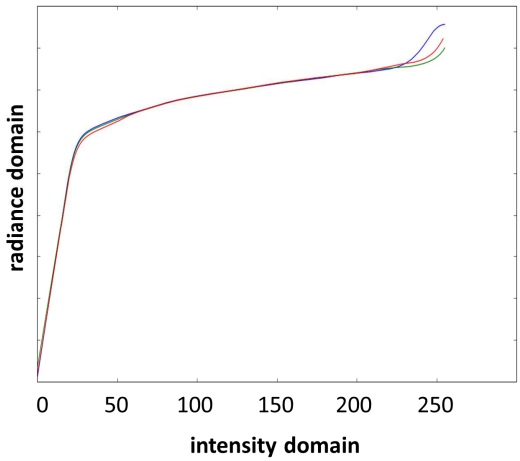


Fig. A.9: Response curve for *Axis 233D*.

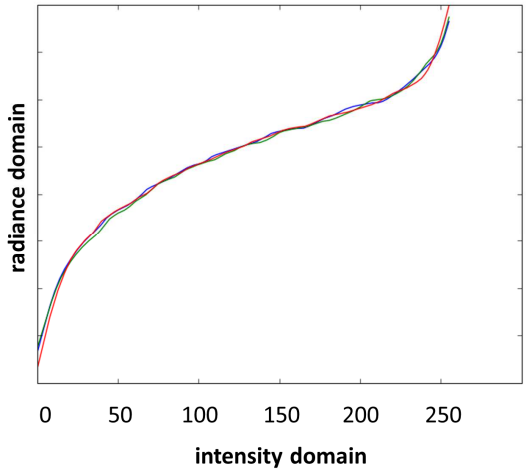


Fig. A.10: Response curve for *Canon IXUS 40*.

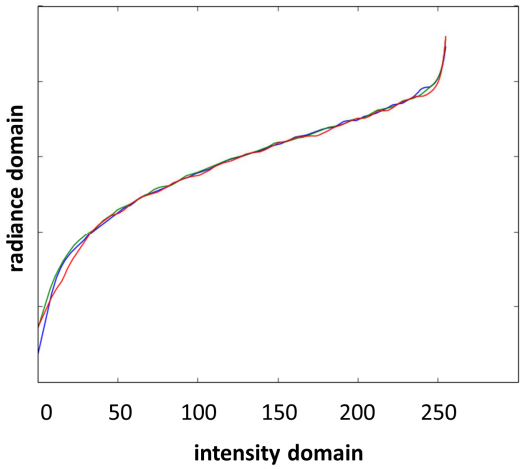


Fig. A.11: Response curve for *Canon IXUS 70*.

A.3 Additional HDR-SR Results

In section 4.4 a method was proposed which generates high-resolution and high-dynamic-range images. Further result images comparing the low-resolution input images and the enhanced images are depicted in figures A.15, A.16, A.17, A.18 and A.19. In all examples the improvements in terms of both, resolution and dynamic-range, are clearly visible. These images were recorded with the *Canon Ixus 70*.

Fig. A.12: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

Fig. A.13: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

Fig. A.14: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

Fig. A.15: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

Fig. A.16: Comparison of low-resolution image captured with auto exposure setting (bicubic interpolated, left) and proposed enhancement (right).

A.4 Logo Dataset

In chapter 5 advanced filter methods were applied to images of a logo dataset. With these altered images the performance of logo retrieval engine was evaluated in section 5.4.1. To obtain a better understanding of the dataset used, we depict a sample image for each of the 30 logos in figure A.20.

Fig. A.17: Sample image for each of the 30 logos contained in the dataset.

A.5 Additional Filter Results

In the following, further results of the motion detection approach, presented in chapter 6, are shown. In Fig. A.21 various intermediate steps, while processing sequence 1, are depicted. In the top left corner a sample frame from the video is shown. Red arrows indicate the true object motions in the scene. The top right image depicts the background image, in which no moving object is visible. Subtracting the video frames from this background image generates motion hypothesis, which are shown in the right center image. For comparison the frame-wise difference image is shown in the left center. The advantage of background subtraction is clearly visible, as the motion hypothesis are solid blobs and not just contours. Registration errors also generate false motion hypothesis, e.g. the white stationary car on the right. These can be eliminated by our filtering scheme as shown on the bottom left. From these final motion hypothesis, all truly moving objects can be extracted, whose bounding boxes are visualized on the bottom right. In figures A.22, A.23, A.24, A.25 and A.26 results for other sequences of *dataset B* are shown, from which similar observations can be made.

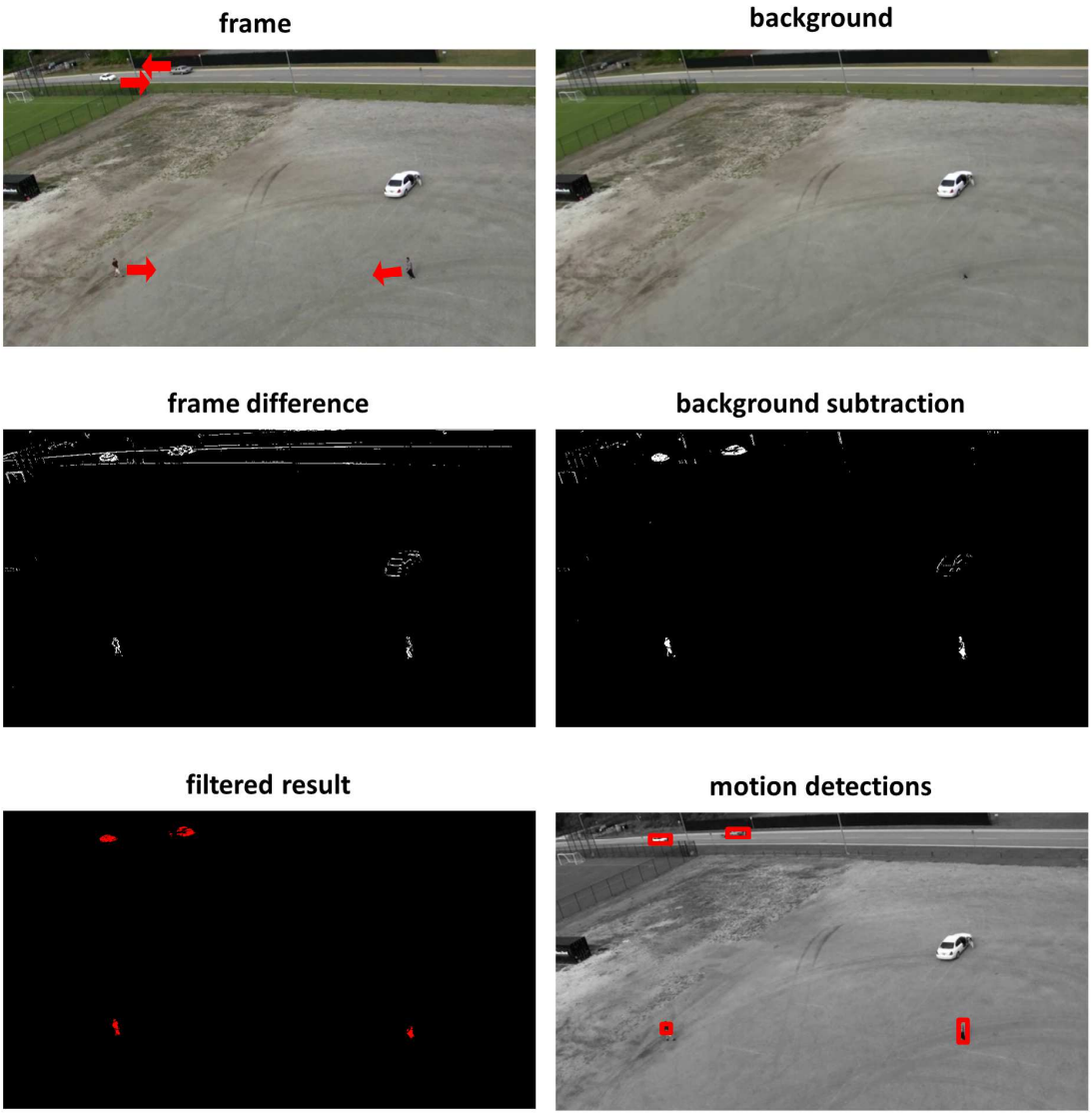


Fig. A.18: Results of intermediate steps for detecting moving objects in sequence 01.

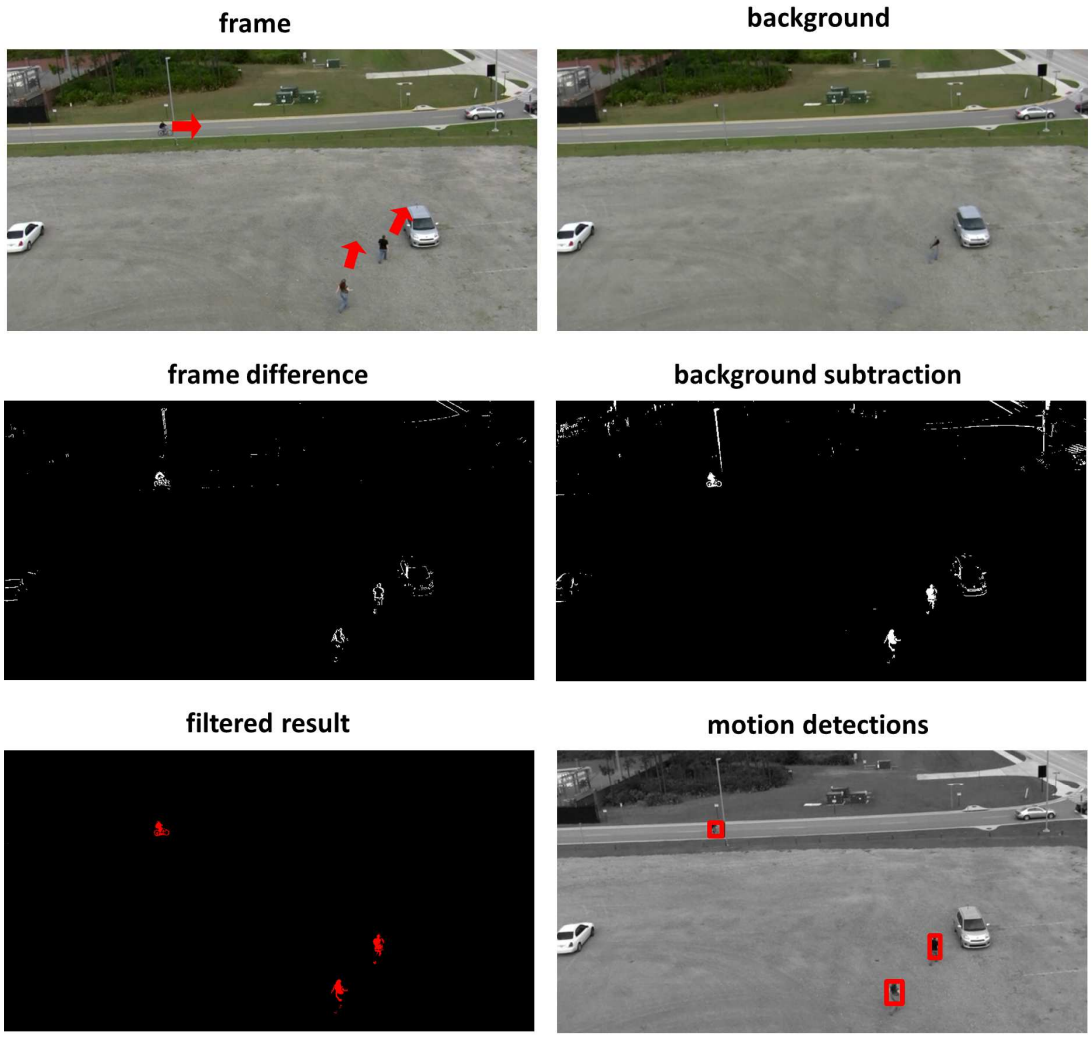


Fig. A.19: Results of intermediate steps for detecting moving objects in sequence 02.

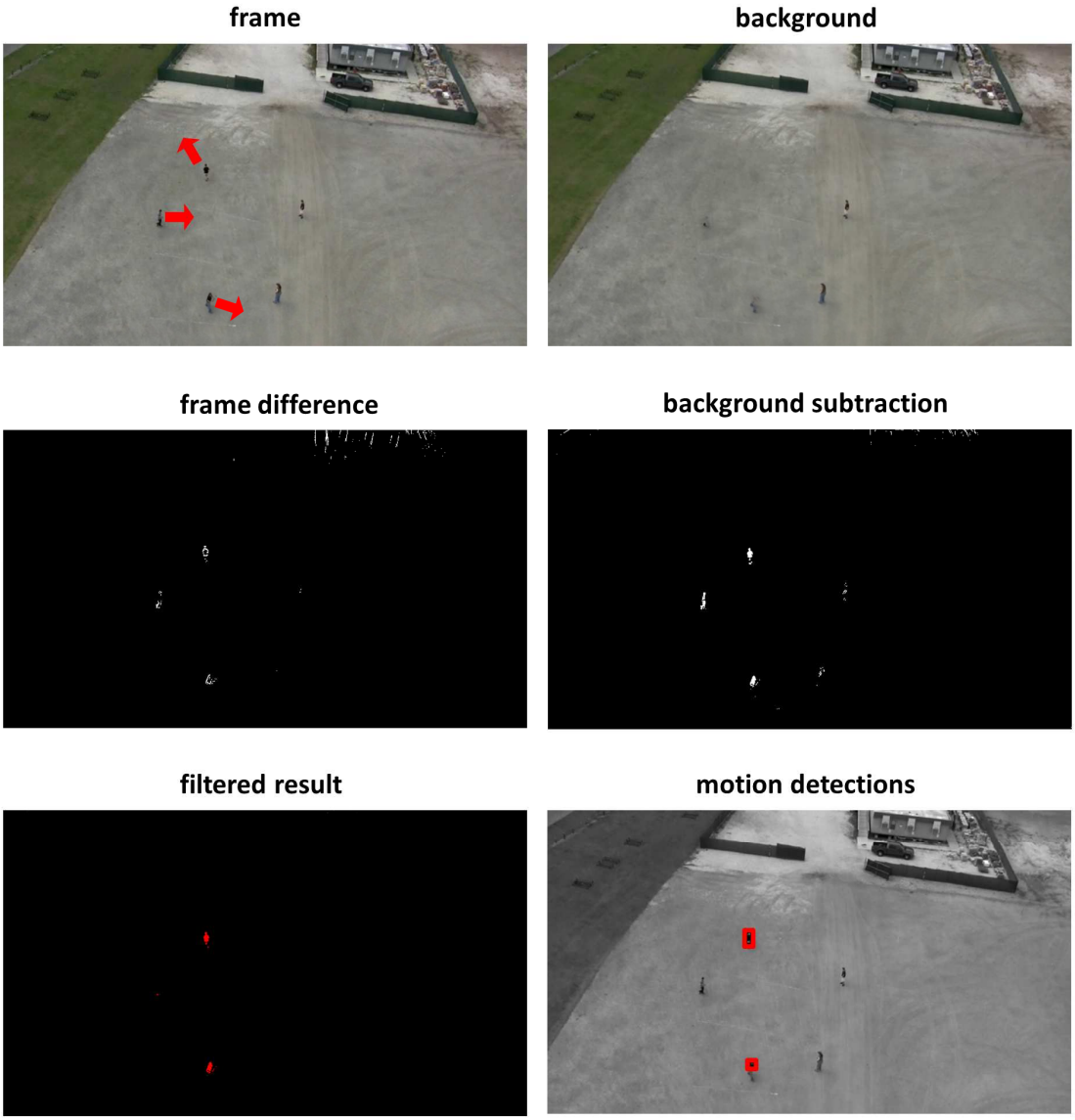


Fig. A.20: Results of intermediate steps for detecting moving objects in sequence 03.

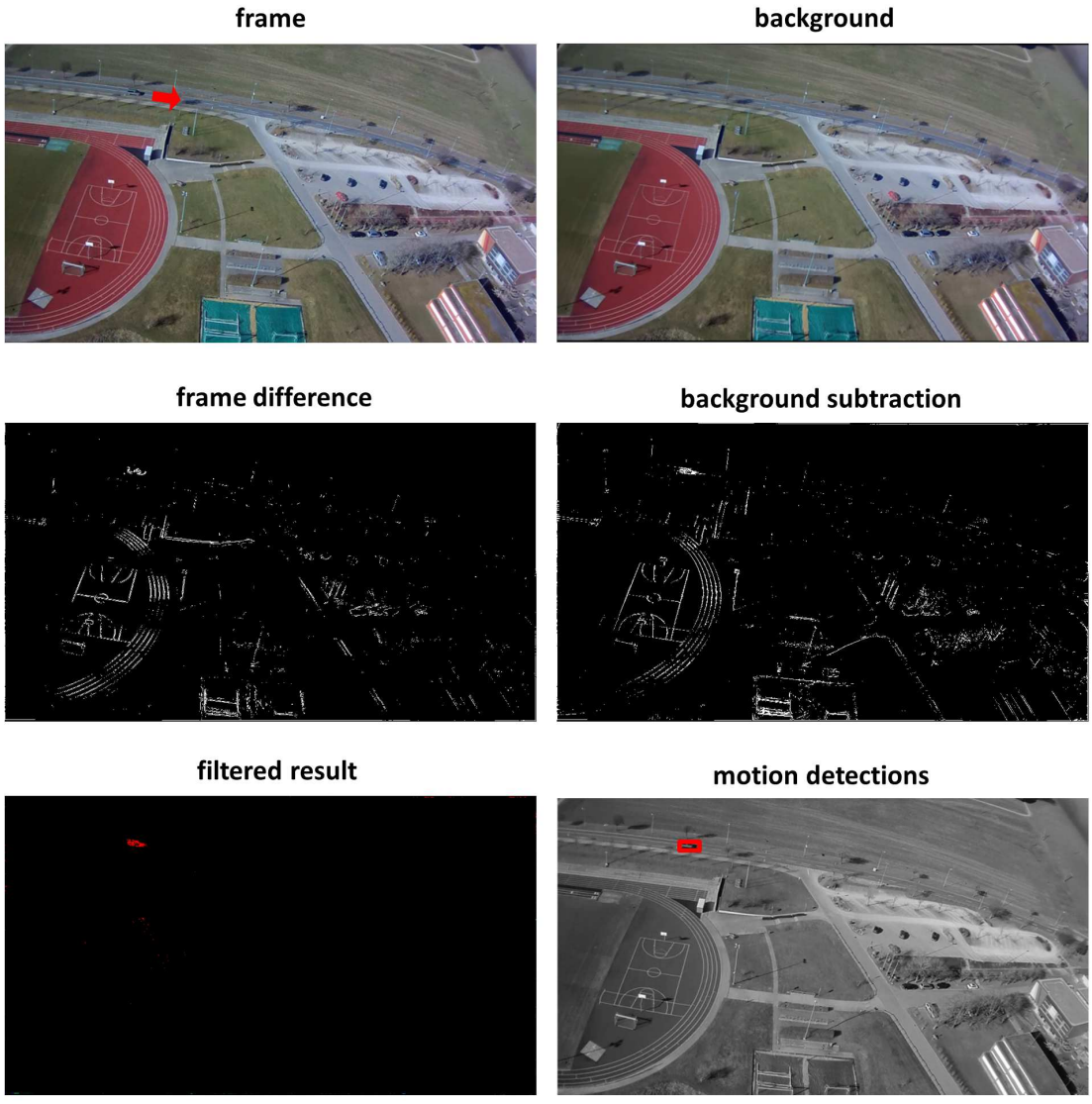


Fig. A.21: Results of intermediate steps for detecting moving objects in sequence 04.

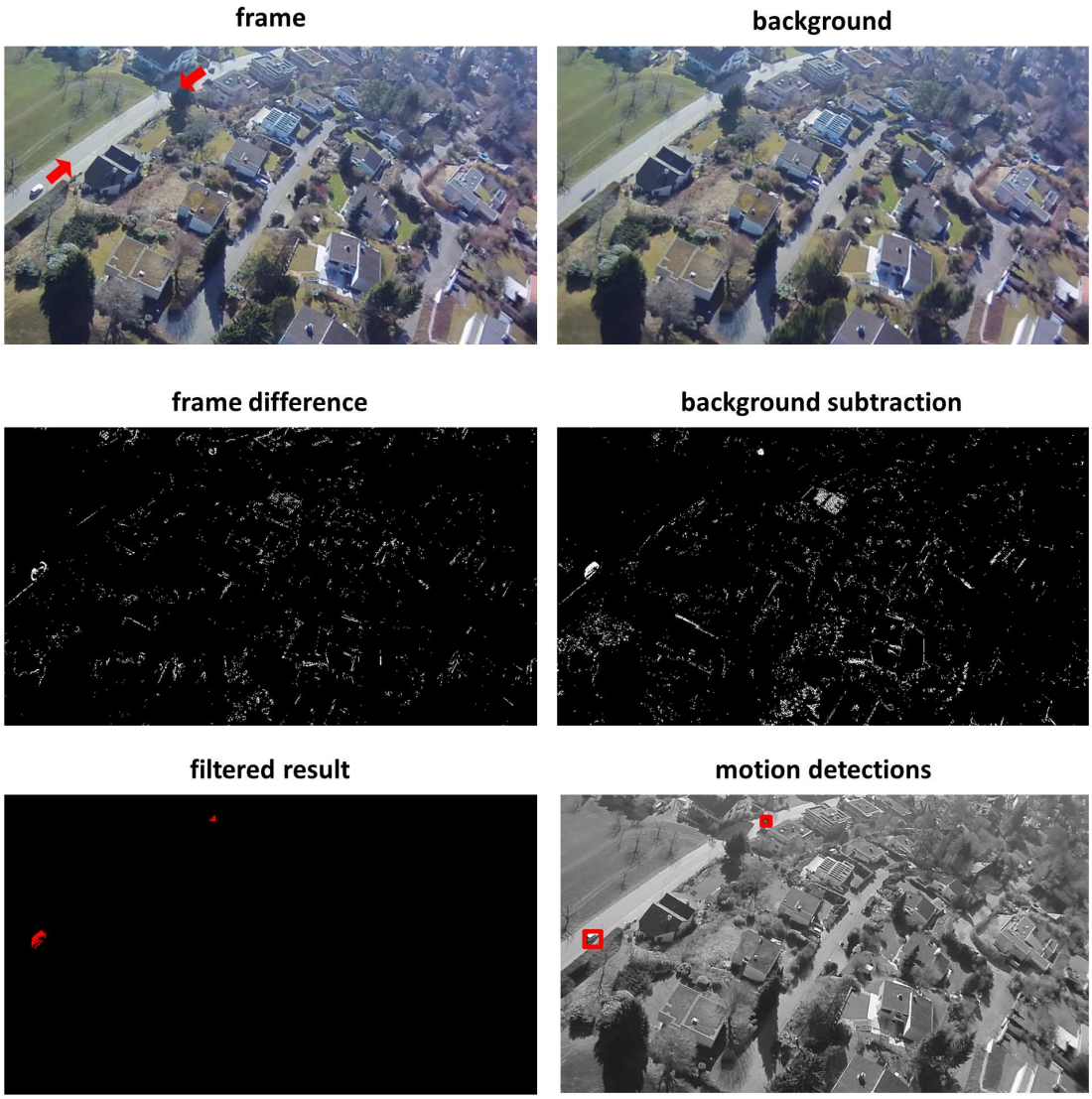


Fig. A.22: Results of intermediate steps for detecting moving objects in sequence 05.

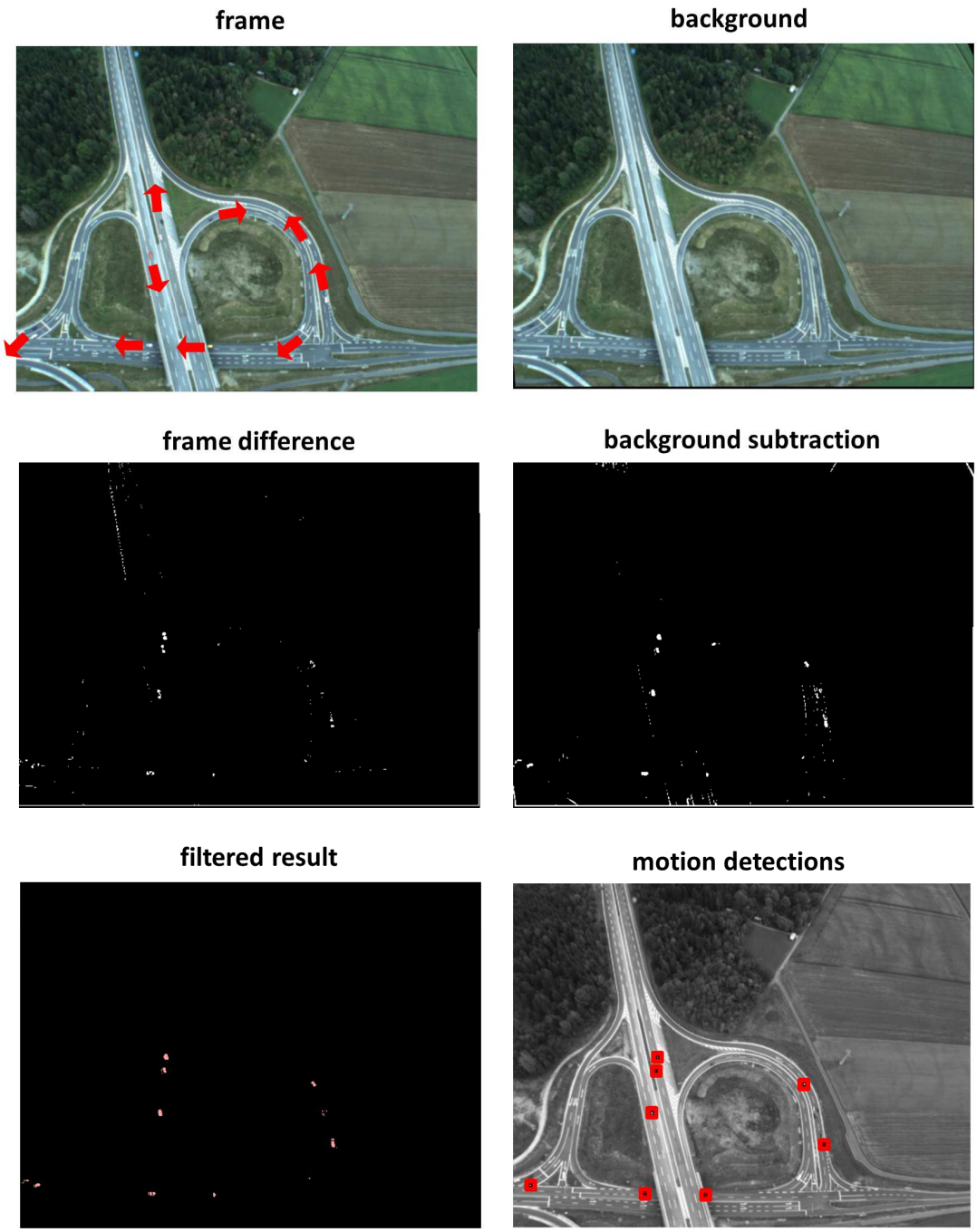


Fig. A.23: Results of intermediate steps for detecting moving objects in sequence 07.

Bibliography

[1] *Digital Image Processing (3rd Edition).* Prentice-Hall, Inc., 2006.

[2] S. Ali, V. Reilly, and M. Shah. Motion and appearance contexts for tracking and re-acquiring targets in aerial videos. *Computer Vision and Pattern Recognition*, 2007.

[3] S. Ali and M. Shah. Cocoa - tracking in aerial imagery. *International Conference on Computer Vision*, 2005.

[4] V. Argyriou. *Advanced Motion Estimation Algorithms in the Frequency Domain for Digital Video Applications.* PhD thesis, University of Surrey, 2006.

[5] Computer Vision Lab at University of Florida. Ucf-lockheed-martin uav dataset. http://server.cs.ucf.edu/~vision/aerial/index.html, 2009.

[6] AXIS Communications, Video Motion Detection Software. http://www.axis.com/products/video/about_networkvideo/vmd.htm. 2011.

[7] S. Baker and T. Kanade. Super-resolution optical flow. Technical report, Carnegie Mellon University, 1999.

[8] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[9] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part 1. Technical Report CMU-RI-TR-02-16, Robotics Institute, Pittsburgh, PA, July 2002.

[10] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 2004.

[11] A. Bartoli and A. Zisserman. Direct estimation of non-rigid registrations. *British Machine Vision Conference*, 2004.

[12] A.E. Bartoli. Groupwise geometric and photometric direct image registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 2008.

[14] I. Begin and F.P. Ferrie. Blind super-resolution using a learning-based approach. *International Conference on Pattern Recognition*, 2004.

[15] C. Benedek, T. Sziranyi, Z. Kato, and J. Zerubia. Detection of object motion regions in aerial image pairs with a multi-layer markovian model. *IEEE Transactions on Image Processing*, 2009.

[16] S. Benhimane and E. Malis. Real-time image-based tracking of planes using efficient second-order minimization. *IEEE/RSJ International Conference on Intelligent Robots Systems*, 2004.

[17] R. Berthilsson. Affine correlation. *International Conference on Pattern Recognition*, 1998.

[18] P. Bhat, C. L. Zitnick, N. Snavely, A. Agarwala, M. Agrawala, B. Curless, M. Cohen, and S. B. Kang. Using photographs to enhance videos of a static scene. *Eurographics Symposium on Rendering*, 2007.

[19] S. Bhattacharya, H. Idrees, I. Saleemi, S. Ali, and M. Shah. *Machine Vision Beyond Visible Spectrum*, chapter Moving Object Detection and Tracking in Forward Looking Infra-Red Aerial imagery, pages 221–252. Spring, 2011.

[20] C. M. Bishop, A. Blake, and B. Marthi. Super-resolution enhancement of video. *Artificial Intelligence and Statistics*, 2003.

[21] S. Borman. *Topics in Multiframe Superresolution Restoration.* PhD thesis, University of Notre Dame, Notre Dame, IN, May 2004.

[22] N.K. Bose, S. Lertrattanapanich, and M.B. Chappalli. Superresolution with second generation wavelets. *Signal Processing: Image Communication*, 2004.

[23] L.G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 1992.

[24] H. Cai, K. Mikolajczyk, and J. Matas. Learning linear discriminant projections for dimensionality reduction of image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[25] M. Calonder, V. Lepetit, P. Fua, K. Konolige, J. Bowman, and P. Mihelich. Compact signatures for high-speed interest point description and matching. *International Conference on Computer Vision*, 2009.

[26] F. Candocia and J.C. Príncipe. Super-resolution of images based on local correlations. *IEEE Transactions on Neural Networks*, 1999.

[27] D. Capel. Super-resolution enhancement of text image sequences. *International Conference on Pattern Recognition*, 2000.

[28] D. P. Capel. *Image Mosaicing and Super-resolution.* PhD thesis, University of Oxford, 2001.

[29] A. Chariot and R. Keriven. Gpu-boosted online image matching. *International Conference on Pattern Recognition*, 2008.

[30] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and Henson R. Super-resolved surface reconstruction from multiple images. Technical Report Technical Report FIA-94-12, NASA, Ames Research Center, 1994.

[31] E. Choi, J. Choi, and M.G. Kang. Super-resolution approach to overcome physical limitations of imaging sensors: An overview. *International Journal of Imaging Systems and Technology*, 2004.

[32] J. Choi, M. K. Park, and M. G. Kang. High dynamic range image reconstruction with spatial resolution enhancement. *The Computer Journal*, 2007.

[33] N. Cornelis and L. Van Gool. Fast scale invariant feature detection and matching on programmable graphics hardware. *Computer Vision Pattern Recognition Workshop*, 2008.

[34] NVIDIA Corporation. CUDA CUFFT library, 2007.

[35] G. Cristóbal, E. Gil, F. Šroubek, J. Flusser, C. Miravet, and F.B. Rodríguez. Superresolution imaging: a survey of current techniques. *Advanced Signal Processing Algorithms, Architectures, and Implementations, Proceedings of SPIE*, 2008.

[36] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition*, 2005.

[37] J. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 2007.

[38] P.E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. *Special Interest Group of Graphics*, 1997.

[39] G. Dedeoglu, T. Kanade, and J. August. High-zoom video hallucination by exploiting spatio-temporal regularities. *Computer Vision and Pattern Recognition*, 2004.

[40] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun. Structure from motion without correspondence. *Computer Vision and Pattern Recognition*, 2000.

[41] S. Dey, V. Reilly, I. Saleemi, and M. Shah. Detection of independently moving objects in non-planar scenes via multi-frame monocular epipolar constraint. *European Conference on Computer Vision*, 2012.

[42] K. Donaldson and G. K. Myers. Bayesian super-resolution of text in video with a text-specific bimodal prior. *Computer Vision and Pattern Recognition*, 2005.

[43] A. Eden, M. Uyttendaele, and R. Szeliski. Seamless image stitching of scenes with large motions and exposure differences. *Computer Vision and Pattern Recognition*, 2006.

[44] A. Efros and T. Leung. Texture synthesis by non-parametric sampling. *International Conference on Computer Vision*, 1999.

[45] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE Transactions on Image Processing*, 1997.

[46] M. Elad and Y. Hel-Or. A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur. *IEEE Transactions on Image Processing*, 2001.

[47] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 2010.

[48] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. *Special Interest Group of Graphics*, 2008.

[49] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Advances and challenges in super-resolution. *International Journal of Imaging Systems and Technology*, 2004.

[50] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and Robust Multiframe Super Resolution. *IEEE Transactions on Image Processing*, 2004.

[51] R. Fattal, D. Lischinski, and M. Werman. Gradient domain high dynamic range compression. *Special Interest Group of Graphics*, 2002.

[52] D. Fedorov, B. Sumengen, and B. S. Manjunath. Multi-focus imaging using local focus estimation and mosaicking. *International Conference on Image Processing*, 2006.

[53] J. A. Ferwerda. Elements of early vision for computer graphics. *IEEE Computer Graphics and Applications*, 2001.

[54] IMAX Films. http://www.imax.com/. 2011.

[55] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *ACM Communications*, 1981.

[56] L. Fletcher, L. Petersson, N. Barnes, D. J. Austin, and A. Zelinsky. A sign reading driver assistance system using eye gaze. *International Conference on Robotics and Automation*, 2005.

[57] R. Fransens, C. Strecha, and L. van Gool. A probabilistic approach to optical flow based super-resolution. *Computer Vision Pattern Recognition Workshop*, 2004.

[58] W.T. Freeman, W.T. Freeman, T.R. Jones, and E.C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 2002.

[59] M. Frigo and S.G. Johnson. The design and implementation of FFTW3. *IEEE Special issue on "Program Generation, Optimization, and Platform Adaptation"*, 2005.

[60] M. Garland, S. Le Grand, J. Nickolls, J. Anderson, J. Hardwick, S. Morton, E. Phillips, Yao Zhang, and V. Volkov. Parallel computing experiences with cuda. *Micro, IEEE*, 2008.

[61] M. Gevrekci. *Super Resolution and Dynamic Range Enhancement of Image Sequences*. PhD thesis, Louisiana State University and Agricultural and Mechanical College, Department of Electrical and Computer Engineering, 2009.

[62] M. Gevrekci and B. K. Gunturk. Superresolution under photometric diversity of images. *Eurasip Journal on Advances in Signal Processing*, 2007.

[63] R. Gross and V. Brajovic. An image preprocessing algorithm for illumination invariant face recognition. *Audio- and Video-Based Biometric Person Authentication*, 2003.

[64] M.D. Grossberg and S.K. Nayar. Determining the Camera Response from Images: What is Knowable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003.

[65] F. Guichard and L. Rudin. Velocity estimation from images sequences and application to super-resolution. *International Conference on Image Processing*, 1999.

[66] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 2003.

[67] A. Gupta, P. Bhat, M. Dontcheva, M. Cohen, B. Curless, and O. Deussen. Enhancing and experiencing spacetime resolution with videos and stills. *International Conference on Computational Photography*, 2009.

[68] M. Hammami, S.K. Jarraya, and H. Ben-Abdallah. A comparative study of proposed moving object detection methods. *Journal of Next Generation Information Technology*, 2011.

[69] R. C. Hardie, K. J. Barnard, and E. E. Armstrong. Joint map registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 1997.

[70] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[71] J. Hays and A. A. Efros. Scene completion using millions of photographs. *Special Interest Group of Graphics*, 2007.

[72] T. Heseltine, N. E. Pears, and J. Austin. Evaluation of image pre-processing techniques for eigenface based face recognition. *ICIG*, 2002.

[73] M.C. Hong, M.G. Kang, and A.K. Katsaggelos. An iterative weighted regularized algorithm for improving the resolution of video sequences. *International Conference on Image Processing*, 1997.

[74] A. Horváth. Aaphoto. *http://log69.com/aaphoto_en.html*, 2011.

[75] B. Q. Huang, Y. B. Zhang, and M. T. Kechadi. Preprocessing techniques for online handwriting recognition. *Proceedings of the Seventh International Conference on Intelligent Systems Design and Applications*, 2007.

[76] M. Irani, P. Anandan, and S. Hsu. Mosaic based representations of video sequences and their applications. *International Conference on Computer Vision*, 1995.

[77] M. Irani and S. Peleg. Super resolution from image sequences. *International Conference on Pattern Recognition*, 1990.

[78] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP*, 1991.

[79] E. Ito, S. Saga, T. Okatani, and K. Deguchi. Gpu-based high-speed and high-precision visual tracking. *SICE Annual Conference*, 2010.

[80] R. Jain and D. Doermann. Logo retrieval in document images. *Document Analysis Systems, IAPR International Workshop on*, 2012.

[81] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. *European Conference on Computer Vision*, 2008.

[82] C.V. Jiji, M.V. Joshi, and S. Chaudhuri. Single-frame image super-resolution using learned wavelet coefficients. *International Journal of Imaging Systems and Technology*, 2004.

[83] H. Jin, P. Favaro, and S. Soatto. Real-time feature tracking and outlier rejection with changes in illumination. *International Conference on Computer Vision*, 2001.

[84] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-Backward Error: Automatic Detection of Tracking Failures. *International Conference on Pattern Recognition*, 2010.

[85] Y. Kalantidis, L.G. Pueyo, and M. Trevisiol. Scalable triangulation-based logo recognition. *International Conference on Multimedia Retrieval*, 2011.

[86] J. Kang, I. Cohen, and G. Medioni. Continuous tracking within and across camera streams. *Computer Vision and Pattern Recognition*, 2003.

[87] J. Kang, I. Cohen, and C. Yuan. Detection and tracking of moving objects from a moving platform in presence of strong parallax. *International Conference on Computer Vision*, 2005.

[88] J. Kang, K. Gajera, I. Cohen, and G. Medioni. Detection and tracking of moving objects from overlapping eo and ir sensors. *Computer Vision and Pattern Recognition Workshop*, 2004.

[89] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High dynamic range video. *Special Interest Group of Graphics*, 2003.

[90] D. Keren, S. Peleg, and R. Brada. Image sequence enhancement using subpixel displacements. *Computer Vision and Pattern Recognition*, 1988.

[91] S.J. Kim, D. Gallup, J.M. Frahm, A. Akbarzadeh, Q. Yang, R. Yang, D. Nister, and M. Pollefeys. Gain adaptive real-time stereo streaming. *International Conference on Computer Vision Systems*, 2007.

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

[93] C. D. Kuglin and D. C. Hines. The phase correlation image alignment method. *IEEE International Conference of Cybernetic Society*, 1975.

[94] M.Prema Kumar, P.H.S.Tejo Murthy, and P.Rajesh Kumar. Article: Performance evaluation of different image filtering algorithms using image quality assessment. *International Journal of Computer Applications*, 2011.

[95] S.H. Lai. Robust image matching under partial occlusion and spatially varying illumination change. *Computer Vision and Image Understanding*, 1999.

[96] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scenes and categories. *Computer Vision and Pattern Recognition*, 2006.

[97] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen. Evaluation of tone mapping operators using a high dynamic range display. *Special Interest Group of Graphics*, 2005.

[98] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[99] Y. Li, L. Sharan, and E. H. Adelson. Compressing and companding high dynamic range images with subband architectures. *Special Interest Group of Graphics*, 2005.

[100] R. Lin, X. Cao, Y. Xu, C. Wei, and H. Qiao. Airborne moving vehicle detection for urban traffic surveillance. *International IEEE Conference on Intelligent Transportation Systems*, 2008.

[101] Z. Lin and S. Heung-Yeung. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[102] Z. Lin, J.H., X. Tang, and C. Tang. Limits of learning-based superresolution algorithms. *International Journal of Computer Vision*, 2008.

[103] C. Liu, H.-Y.g Shum, and C.-S. Zhang. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. *Computer Vision and Pattern Recognition*, 2001.

[104] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 1989.

[105] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

[106] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. *Proceedings of Imaging Understanding Workshop*, 1981.

[107] H. Q. Luong, B. Goossens, A. Pizurica, and W. Philip. Consistent joint photometric and geometric image regristration. *International Conference on Image Processing*, 2010.

[108] E. Malis. Improving vision-based control using efficient second-order minimization techniques. *International Conference on Robotics and Automation*, 2004.

[109] S. Mann. Compositing multiple pictures of the same scene. *Proceedings of the 46th Annual IS&T Conference*, 1993.

[110] S. Mann and R. Mann. Quantigraphic imaging: Estimating the camera response and exposures from differently exposed images. *Computer Vision and Pattern Recognition*, 2001.

[111] S. Mann and R. W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. *Proceedings of IS&T*, 1995.

[112] D. W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 1963.

[113] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[114] Li Min, Li Shi Hua, Wang Fu, Le Xiang, Hong Jin, and Jiang Lian Jun. Super-resolution based on improved sparse coding. *International Conference on Computer Application and System Modeling*, 2010.

[115] T. Mitsunaga and S.K. Nayar. Radiometric Self Calibration. *Computer Vision and Pattern Recognition*, 1999.

[116] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *In VISAPP International Conference on Computer Vision Theory and Applications*, 2009.

[117] H. Nagahara, T. Matsunobu, Y. I., M. Yachida, and T. Suzuki. High-resolution video generation using morphing. *International Conference on Pattern Recognition*, 2006.

[118] M.K. Ng, H. Shen, E.Y. Lam, and L. Zhang. A total variation regularization based super-resolution reconstruction algorithm for digital video. *EURASIP Journal on Advances in Signal Processing*, 2007.

[119] N. Nguyen and P. Milanfar. An efficient wavelet-based algorithm for image super-resolution. *International Conference on Image Processing*, 2000.

[120] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[121] Pan-STARRS Project. http://pan-starrs.ifa.hawaii.edu/public/home.html. 2010.

[122] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 2003.

[123] D.H. Parks and S.S. Fels. Evaluation of background subtraction algorithms with post-processing. *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2008.

[124] A. J. Patti, M. I. Sezan, and A. M. Tekalp. Super resolution video reconstruction with arbitrary sampling lattices and nonzero aperture time. *IEEE Transactions on Image Processing*, 1997.

[125] S. Peleg, D. Keren, and L. Schweitzer. Improving image resolution using subpixel motion. *Pattern Recogn. Lett.*, 1987.

[126] S. Pelletier, S. P. Spackman, and J. R. Cooperstock. High-resolution video synthesis from mixed-resolution video based on the estimate-and-correct methods. *Workshop on the Applications of Computer Vision - MOTION*, 2005.

[127] T.Q. Pham. *Spatiotonal adaptivity in superresolution of undersampled image sequences.* PhD thesis, Faculty of Applied Sciences, Delft University of Technology, 2006.

[128] J. Philbin. *Scalable Object Retrieval in Very Large Image Collections.* PhD thesis, University of Oxford, 2010.

[129] M Piccardi. Background subtraction techniques: a review. *IEEE International Conference on Systems Man and Cybernetics*, 2004.

[130] L. C. Pickup. *Machine Learning in Multi-frame Image Super-resolution.* PhD thesis, University of Oxford, February 2008.

[131] L. C. Pickup, S. J. Roberts, and A. Zisserman. A sampled texture prior for image super-resolution. *Advances in Neural Information Processing Systems*, 2003.

[132] L. C. Pickup, S. J. Roberts, and A. Zisserman. Optimizing and learning for super-resolution. *British Machine Vision Conference*, 2006.

[133] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 2005.

[134] B. S. Reddy and B. N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 1996.

[135] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[136] M. A. Robertson, S. Borman, and R.L. Stevenson. Estimation-theoretic approach to dynamic range enhancement using multiple exposures. *Journal of Electronic Imaging*, 2003.

[137] N. Rosenblum. *A World History of Photography*. Abbeville Press, 2007.

[138] E. Rosten, R. Porter, and T. Drummond. Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[139] L. Rudin. Image frame fusion by velocity estimation using region merging, 2002.

[140] P. Sand. *Long-Range Video Motion Estimation using Point Trajectories*. PhD thesis, Massachusetts Institute of Technology, 2006.

[141] P. Sand and S.J. Teller. Particle video: Long-range motion estimation using point trajectories. *International Journal of Computer Vision*, 2008.

[142] H. S. Sawhney, Y. Guo, K. Hanna, R. Kumar, S. Adkins, and S. Zhou. Hybrid stereo camera: an ibr approach for synthesis of very high resolution stereoscopic

image sequences. *Conference on Computer graphics and interactive techniques*, 2001.

[143] F. Schmidt. Kit ais data set. http://www.ipf.kit.edu/downloads_data_set_AIS_vehicle_tracking.php, 2012.

[144] R. R. Schultz and R. L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 1996.

[145] S. Se, H. Ng, P. Jasiobedzki, and T. Moyung. Vision based modeling and localization for planetary exploration rovers. *International Astronautical Congress*, 2004.

[146] M.I. Sezan. An overview of convex projections theory and its application to image recovery problems. *Ultramicroscopy*, 1992.

[147] E. Shechtman, E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[148] J. Shi and C. Tomasi. Good features to track. *Computer Vision and Pattern Recognition*, 1994.

[149] S. N. Sinha, J. Frahm, M. Pollefeys, and Y. Genc. Gpu-based video feature tracking and matching. Technical report, In Workshop on Edge Computing Using New Commodity Architectures, 2006.

[150] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. *International Conference on Computer Vision*, 2003.

[151] Greg Slabaugh, Bruce Culbertson, Tom Malzbender, and Ron Schafer. A survey of methods for volumetric scene reconstruction from photographs. *Eurographics Conference On Volume Graphics*, 2001.

[152] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. *Special Interest Group of Graphics*, 2006.

[153] F. Sroubek, G. Cristóbal, and J. Flusser. A unified approach to superresolution and multichannel blind deconvolution. *IEEE Transactions on Image Processing*, 2007.

[154] H. Stark and P. Oskoui. High resolution image recovery from image-plane arrays, using convex projections. *Journal of Optical Society*, 1989.

[155] A. J. Storkey. Dynamic structure super-resolution. *Advances in Neural Information Processing Systems*, 2002.

[156] J. Sun, J. Sun, N.-N. Zheng, H. Tao, and H.-Y. Shum. Image hallucination with primal sketch priors. *Computer Vision and Pattern Recognition*, 2003.

[157] R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, 2010.

[158] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields. *European Conference on Computer Vision*, 2006.

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

[160] A.M. Tekalp, M.K. Ozkan, and M.I. Sezan. High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1992.

[161] A. Temizel and T. Vlachos. Wavelet domain image resolution enhancement. *Vision, Image and Signal Processing*, 2006.

[162] P. Thévenaz, U. E. Ruttimann, and M. Unser. A pyramid approach to subpixel registration based on intensity. *IEEE Transactions on Image Processing*, 1998.

[163] M. E. Tipping and C. M. Bishop. Bayesian image super-resolution. *Advances in Neural Information Processing Systems*, 2003.

[164] B. C. Tom and A. K. Katsaggelos. Reconstruction of a high-resolution image by simultaneous registration, restoration, and interpolation of low-resolution images. *International Conference on Image Processing*, 1995.

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

[166] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[167] R. Y. Tsai and T. S. Huang. Multiframe image restoration and registration. *Advances in Computer Vision and Image Processsing*, 1984.

[168] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 2008.

[169] UCF. Ucf aerial action dataset. http://crcv.ucf.edu/data/UCF_Aerial_Action.php, 2008.

[170] H. Ur and D. Gross. Improved resolution from subpixel shifted pictures. *Graphical Models and Image Processing*, 1992.

[171] P. Vandewalle. *Super-Resolution from Unregistered Aliased Images*. PhD thesis, Ecole Polytechnique Federale De Lausanne, 2006.

[172] P. Viola and III Wells, W.M. Alignment by maximization of mutual information. *International Conference on Computer Vision*, 1995.

[173] F. Šroubek. *Image Fusion via Multichannel Blind Deconvolution*. PhD thesis, MFF, Charles University, 2003.

[174] F. Šroubek, G. Cristóbal, and J. Flusser. A unified approach to superresolution and multichannel blind deconvolution. *IEEE Transactions on Image Processing*, 2007.

[175] F. Šroubek and . Flusser. Multichannel blind deconvolution of spatially misaligned images. *IEEE Transactions on Image Processing*, 2005.

[176] Q. Wang, X. Tang, and H. Shum. Patch based blind image super resolution. *International Conference on Computer Vision*, 2005.

[177] G. Ward. Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. *journal of graphics, gpu, and game tools*, 2003.

[178] K. Watanabe, Y. Iwai, H. Nagahara, M. Yachida, and T. Suzuki. Video synthesis with high spatio-temporal resolution using spectral fusion. *Multimedia Content Representation, Classification and Security*, 2006.

[179] N. A. Woods, N. P. Galatsanos, and A. K. Katsaggelos. Em-based simultaneous registration, restoration, and interpolation of super-resolved images. *International Conference on Image Processing*, 2003.

[180] J. Xiao, H. Cheng, H. S. Sawhney, and F. Han. Vehicle detection and tracking in wide field-of-view aerial video. *Computer Vision and Pattern Recognition*, 2010.

[181] J. Xiao, C. Yang, F. Han, and H. Cheng. Vehicle and person tracking in aerial videos. *Multimodal Technologies for Perception of Humans*, 2007.

[182] F. Yan, J. Kittler, K. Mikolajczyk, and A. Tahir. Non-sparse multiple kernel fisher discriminant analysis. *Journal of Machine Learning Research*, 2012.

[183] F. Yan, K. Mikolajczyk, M. Barnard, H. Cai, and J. Kittler. Lp norm multiple kernel fisher discriminant analysis for object and image categorisation. *Computer Vision and Pattern Recognition*, 2010.

[184] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. *Internationl Conference On Computer Vision*, 2009.

[185] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 2010.

[186] G. Ye. *Image registration and super-resolution mosaicing.* PhD thesis, University of New South Wales - Australian Defence Force Academy. School of Information Technology and Electrical Engineering, 2005.

[187] Q. Yu and G. Medioni. Motion pattern interpretation and detection for tracking moving vehicles in airborne video. *Computer Vision and Pattern Recognition*, 2009.

[188] Q. Yu and Gerard Medioni. A gpu-based implementation of motion detection from a moving platform. *Computer Vision and Pattern Recognition Workshop*, 2008.

[189] C. Yuan, G. G. Medioni, J. Kang, and I. Cohen. Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[190] C. Zach, D. Gallup, and J.-M. Frahm. Fast gain-adaptive klt tracking on the gpu. *Computer Vision and Pattern Recognition Workshop On GPUs*, 2008.

[191] B. Zeisl, P.F. Georgel, F. Schweiger, E. Steinbach, and N. Navab. Estimation of location uncertainty for scale invariant feature points. *British Machine Vision Conference*, 2009.

[192] W.Y. Zhao and H. S. Sawhney. Is super-resolution with optical flow feasible? *European Conference on Computer Vision*, 2002.

[193] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 2003.

[194] A. Zomet, A. Rav-acha, and S. Peleg. Robust super-resolution. *Workshop on the Applications of Computer Vision*, 2001.