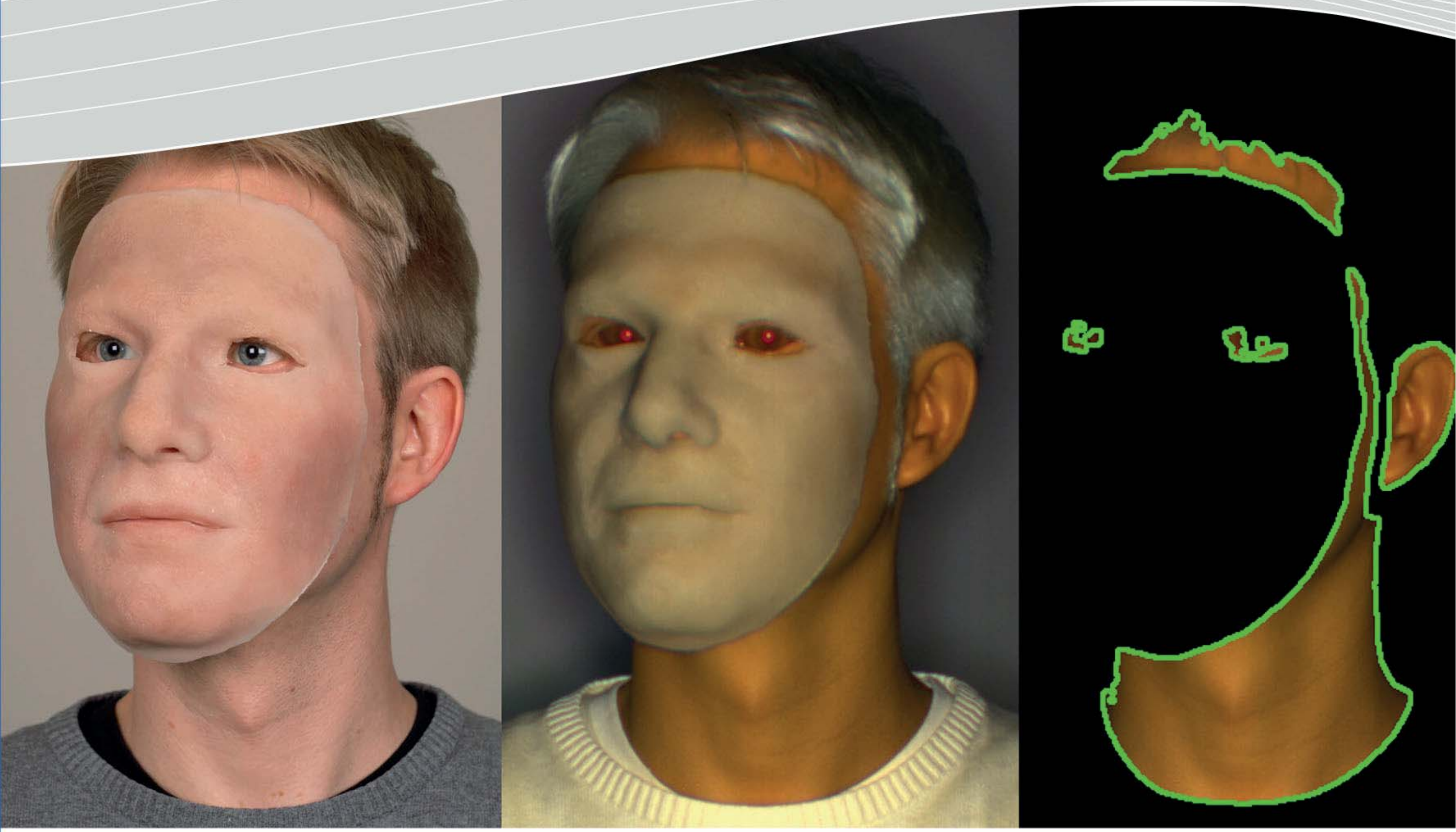


Holger Steiner

Active Multispectral SWIR Imaging for Reliable Skin Detection and Face Verification

Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Active Multispectral SWIR Imaging for
Reliable Skin Detection and Face Verification

Active Multispectral SWIR Imaging for Reliable Skin Detection and Face Verification

DISSERTATION

zur Erlangung des Grades eines Doktors
der Ingenieurwissenschaften

vorgelegt von

Holger Steiner (M.Sc.)

geb. am 07. September 1982 in Bonn

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät

der Universität Siegen

Siegen 2016

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet
über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen: Cuvillier, 2017
Zugl.: Siegen, Univ., Diss., 2016

© CUVILLIER VERLAG, Göttingen 2017
Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2017
Gedruckt auf umweltfreundlichem, säurefreiem Papier aus nachhaltiger Forstwirtschaft.

ISBN 978-3-7369-9450-8
eISBN 978-3-7369-8450-9

Gutachter der Dissertation:

1. Prof. Dr. Andreas Kolb

2. Prof. Dr. Volker Blanz

Tag der mündlichen Prüfung: 02. November 2016

Danksagung

Die vorliegende Dissertation entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter des Institutes für Sicherheitsforschung (ISF) der Hochschule Bonn-Rhein-Sieg. Daher möchte ich zunächst dem Institutsleiter Herrn Prof. Dr.-Ing. Norbert Jung für die Ermöglichung dieser Arbeit und die gute Betreuung danken. Besonderer Dank gilt weiterhin Herrn Prof. Dr. Andreas Kolb und Herrn Prof. Dr. Volker Blanz von der Universität Siegen, die mich durch teils tatkräftige Mitarbeit an Veröffentlichungen, sowie viele Gespräche mit Tipps und konstruktiver Kritik unterstützt haben.

Des Weiteren möchte ich meinen derzeitigen und ehemaligen Kollegen für ihre Hilfe danken, insbesondere den Herren Sebastian Sporrer und Oliver Schwaneberg, die mir stets mit Anregungen, Rat und Tat zur Seite standen, sowie allen Probanden, die sich für die nötigen Messreihen im Rahmen dieser Arbeit zur Verfügung gestellt haben. Nicht vergessen möchte ich auch meine Familie und alle Freunde, die mich bei der Fertigstellung dieser Arbeit unterstützt haben.

Teile dieser Arbeit sind im Rahmen des Forschungsprojektes *Fälschungserkennung in der Gesichtsbiometrie (FeGeb)* entstanden, das durch das Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Förderprogramms „FHprofUnt - Forschung an Fachhochschulen mit Unternehmen“ gefördert wurde (FKZ: 03FH044PX3). In diesem Zusammenhang möchte ich mich auch bei den Projektpartnern bedanken, insbesondere bei Herrn Ralph Breithaupt vom Bundesamt für Sicherheit in der Informationstechnik (BSI) für seine tatkräftige Unterstützung bei der Validierung des vorgestellten Systems. Zu guter Letzt gilt mein Dank der Deutschen Forschungsgemeinschaft für die Förderung des DFG Graduiertenkolleg 1564 „Imaging New Modalities“ an der Universität Siegen.

Zusammenfassung

Die Erkennung menschlicher Haut in Bildern ist für viele Anwendungsgebiete von Vorteil. Insbesondere die biometrische Gesichtserkennung, die zunehmend häufig z.B. bei der automatisierten Grenz- oder Zugangskontrolle Verwendung findet, kann davon profitieren. Im sichtbaren Lichtspektrum allein ist die Unterscheidung echter Haut von anderen Materialien in Anbetracht verschiedener Hauttypen und wechselnder Lichtbedingungen jedoch häufig schwierig. Daher sind Täuschungsangriffe mit Verkleidungen oder Masken immer noch ein großes Problem für den derzeitigen Stand der Technik.

Diese Dissertation beschreibt einen neuen Ansatz zur Hauterkennung, der auf den charakteristischen spektralen Remissionseigenschaften von Haut im Nahinfrarotspekrum basiert, und stellt eine modalitätsübergreifende Methode zur Erweiterung bestehender Lösungen vor, mit der die Echtheit von Gesichtern sichergestellt wird. Weiterhin beschreibt sie ein Referenzdesign für ein aktives multispektrales Kamerasystem und dessen Implementierung, sowie eine umfassende Validierung des Konzepts.

Das System erfasst multispektrale Bilder mit vier Wellenlängenbändern in einer Zeit von $T = 50\,\mathrm{ms}$. Mit Hilfe eines Machine-Learning-basierten Klassifikators erzielt es eine bisher unerreichte Genauigkeit bei der Hauterkennung und unterscheidet selbst hautähnliche Materialien zuverlässig von echter Haut. In Kombination mit einer kommerziellen Gesichtserkennungssoftware wehrt das System erfolgreich alle untersuchten Täuschungsangriffe ab.

Abstract

The detection of human skin in images is a very desirable feature for applications such as biometric face recognition, which is becoming more frequently used for, e.g., automated border or access control. However, distinguishing real skin from other materials based on imagery captured in the visual spectrum alone and in spite of varying skin types and lighting conditions can be difficult and unreliable. Therefore, spoofing attacks with facial disguises or masks are still a serious problem for state of the art face recognition algorithms.

This dissertation presents a novel approach for reliable skin detection based on spectral remission properties in the short-wave infrared (SWIR) spectrum and proposes a cross-modal method that enhances existing solutions for face verification to ensure the authenticity of a face even in the presence of partial disguises or masks. Furthermore, it presents a reference design and the necessary building blocks for an active multispectral camera system that implements this approach, as well as an in-depth evaluation.

The system acquires four-band multispectral images within $T = 50\,\text{ms}$. Using a machine-learning-based classifier, it achieves unprecedented skin detection accuracy, even in the presence of skin-like materials used for spoofing attacks. Paired with a commercial face recognition software, the system successfully rejected all evaluated attempts to counterfeit a foreign face.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The reliable detection of human skin in images is a very desirable feature for a variety of applications, especially in the fields of safety and security: on the one hand, a reliable and skin type independent detection and tracking of persons and their hands around potentially dangerous machinery such as robot workplaces, for example, can help to prevent accidents. On the other hand, the capability of distinguishing authentic human skin from other materials can also be used to detect so-called spoofing attacks on face recognition systems. Face recognition is an important tool for many biometric systems and a very active research topic [1]. The human face has advantages over other biometric traits, as it can easily be captured in a non-intrusive way from a distance [2]. Consequently, biometric face recognition systems are becoming more frequently used, for example, at airports in the form of automated border control systems, for access control systems at critical infrastructure or even for user log-on and authentication at computers or smartphones. However, despite the significant progress in the field, face recognition still has serious problems in real-world scenarios when dealing with changing illumination conditions, poses and facial expressions, as well as facial disguises or spoofs, such as masks [3].

Detecting human skin using solely monochrome or color imagery captured in the visual (VIS) spectrum, *i.e.* from approx. 380 nm to 750 nm [4], is problematic, as variations in skin types and illumination conditions can make it very hard to distinguish skin from other materials. Infrared imaging in the spectral range from 700 nm to 2 400 nm, has shown to provide more reliable results [5]. The existing approaches that make use of the short-wavelength infrared (SWIR)[1] spectral range can be classified into four groups: multispectral image acquisition using multiple cameras with band pass filters [5, 6], hyperspectral imagers [7], single cameras using filter wheels with band pass filters for sequential multispectral image acquisition [8] and, more recently, single cameras with Bayer-like band pass filter patterns applied directly on the sensor [9]. All of these systems are passive, *i.e.*, filter-based and without active illumination, and thus require sufficient daylight or external lighting.

This dissertation presents and validates the concept of an active multispectral SWIR camera system that is specifically optimized for skin detection and face verification based on spectral signatures of object surfaces. A spectral signature is a specific combination of remission intensities in distinct, narrow wavebands that is used for the classification of the object's surface material. The active illumination ensures defined and constant lighting conditions within a typical indoor working range while avoiding any shadowing caused by unknown illumination directions.

1.1 Application Examples and Requirements

Although the research and system concept presented in this dissertation is focused on the field of anti-spoofing for biometric face recognition, it is not restricted to this field alone and does not imply any application specific assumptions. In this section, examples of application scenarios are introduced that benefit from a reliable skin detection method and have been addressed in two research projects conducted at the Bonn-Rhein-Sieg University of Applied Sciences (BRSU) in the recent years: *spoof detection at biometric face recognition systems (FeGeb)* and *safe person detection in working areas of industrial robots (SPAI)*.

[1] In the literature, the infrared spectrum below 1.4 µm is commonly referred to as the near infrared band (NIR, or IR-A), while the infrared spectrum above 1.4 µm and up to 3 µm is referred to as the short wave infrared band (SWIR, or IR-B). The spectral signatures discussed in this work are arranged within the wavelength range of 0.9 µm up to 1.7 µm, which covers parts of both the near infrared and short-wavelength infrared. However, most researchers as well as camera manufacturers use only the term SWIR when describing this wavelength range in order to distinguish their research area or products from those that reach only up to 1 µm. This work will adopt this simplification.

1.1.1 Anti-Spoofing for Face Recognition

Biometric face recognition (FR) has been and still is an active research topic within the past decades [10]. Under controlled conditions, current state-of-the-art face recognition algorithms can achieve even better results than human recognition. However, in unconstrained environments, automated face recognition still faces problems handling varying illumination, facial expressions or poses [3]. To overcome the problem of changing illumination conditions, the use of active infrared imagery has been proposed in recent years. Frontal illumination of faces with near infrared (NIR) radiation that is invisible to the human eye helps to reduce lighting problems significantly without distracting or blinding the subjects [10]. However, especially determining whether a recognized face is authentic or "fake", *i.e.*, a printed picture or a facial disguise, is still an open issue of face recognition systems [1,3].

There are several reasons for attacking a face recognition system using so-called spoofs, such as to counterfeit the face of an authorized person at access control points or to disguise the own identity when entering a sports stadium although being banned [11]. Spoofing attacks range from printed photos over recorded video displayed, for example, on a mobile device, to facial disguises and masks, which might cover the face partially or completely. The impact of such attacks on face recognition has been researched in several studies, for example in the context of the research project TABULA RASA [12].

By using a face recognition system that is capable of distinguishing authentic skin from spoofs reliably, most spoofing attacks can be detected and rejected. In this thesis, the following two applications of face recognition systems are analyzed:

Automated Border Crossing Systems, so called *eGates*, have been introduced in recent years and are becoming more frequently used, for example at airports [13]. These systems consist of an electronic passport reader and a biometric face recognition system, which captures the face of a person and compares its biometric features to those found in the image read from the *ePassport*. If the features match, the person is allowed to pass. Figure 1.1 on the following page shows an example of an eGate system.

Access Control Systems are another common application for face recognition systems. Only users whose facial features are registered on a *whitelist* are granted access by such a system. A simple example is the *face unlock* feature of Android smartphones [14]. More advanced solutions are commercially available on the market. Besides user log-on or granting physical access to high security areas, they can also be used to protect critical infrastructure from unwanted

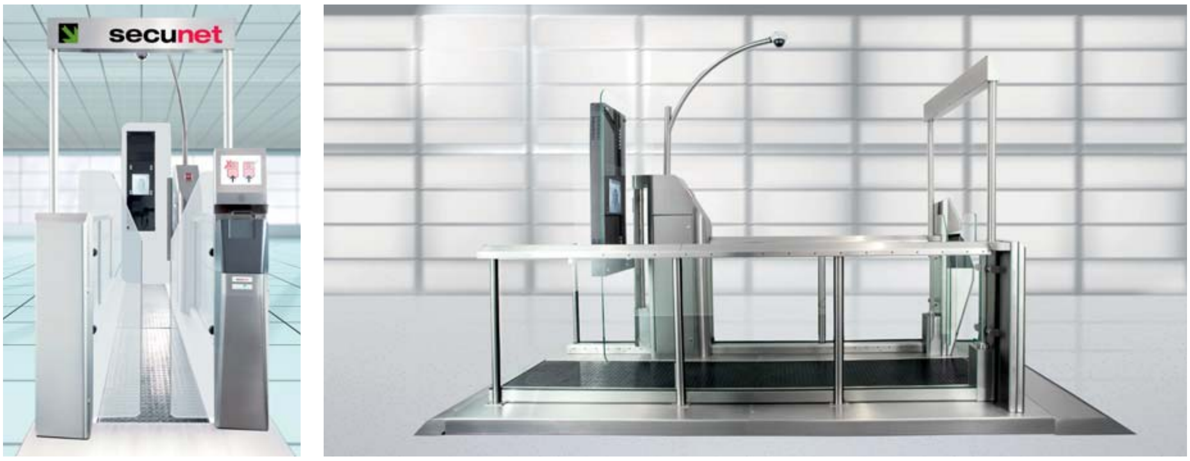


Figure 1.1: Example of an eGate system.
Images: secunet Security Networks AG

individuals. For this purpose, the system may use a *blacklist* containing facial features of persons who are not allowed to enter. A potential application for this *blacklisting* method can be found at sports stadiums: operators often keep registers of people who are not allowed to enter the stadium, *e.g.*, because they have been banned for violent behavior. Automating the identification of these individuals using face recognition at the security check may increase the chance of successfully keeping them from entering the stadium.

Independent of using a white- or a blacklist, both applications describe so-called cooperative user scenarios: users of such face recognition systems can be expected to cooperate with the recognition process by turning their heads towards the camera or by removing any head wear, because they are only granted access if their face has been captured successfully. Without assuming a specific application, the following rather generic requirements on a suitable camera system for anti-spoofing have been formulated in the context of this work:

1. **Reliable material classification.** To detect potential spoofing attacks, all skin and non-skin surfaces must be reliably distinguished and only authentic faces must be accepted by the face recognition system, independent of a users skin type, gender or age. Any material that is falsely classified as skin is a potential security threat.

2. **Detailed image of the facial region.** The face of a user must be captured with sufficiently high spatial resolution in order to extract the biometric features.

3. **Method to combine skin and face detection.** Skin detection and face recognition modules must be combined in order to reliably reject spoofing attacks and to avoid opening up new possibilities to attack the system.

1.1.2 Other Applications

Contactless detection of persons and their limbs is also a desirable feature for many safety applications. At manually-fed machines such as bench saws or presses, for example, potentially dangerous moving parts are difficult to shield off from the reach of the user during normal operation. As productive working requires the user to be near the machine at all time, these machines are very prone to accidents [15]. A similar problem exists at robot workplaces: fast moving parts or equipped tools of industrial robots, for example, pose a safety risk for any humans within the robots' working range. Therefore, robot workplaces are often caged in completely and the robot is stopped while there are people within the cage, making it impossible for humans to work together with the robot in a so-called *joint-action scenario* [16]. To avoid this issue, sensor-based safety technology for industrial robots has been researched since the early 1980s [17] and is still essential for the further development of human-robot collaboration today [18]. Both applications can greatly benefit from the imaging technology proposed in this work.

State-of-the-art safeguarding equipment such as *vision-based protective devices* uses a technique known as *muting* to allow workpieces or moving parts of robots to enter dangerous areas while all other objects, *e.g.*, human limbs, will cause an emergency stop [19]. This technique requires detailed model knowledge about the application and thus restricts joint-action scenarios for humans and robots. By distinguishing human limbs from workpieces through material classification, muting techniques can be implemented in a smarter and much more flexible way. This approach is currently being investigated at the BRSU in the context of the research project *SPAI*. In Section 8.2, findings and results of this research project are summarized and an outlook to future work on such application scenarios is given.

1.2 Contributions

This dissertation presents a concept of an image-based skin detection and face verification system. Some parts of this work have already been presented in scientific publications: the basic idea was first presented at the Imaging and Applied Optics congress [20]. A more detailed description and first evaluation has been published in the Journal of Sensors [21]. A paper presented at the Conference on Biometrics [22] focuses on the detection of spoofing attacks, while a paper presented at the SIAS conference by Sporrer *et al*. [23] proposes a similar camera system for applications in the safety domain. Another paper currently in preparation [24] deals with the problem of motion compensation for multispectral imaging systems that capture spectral information (time-) sequentially.

The contributions of this work are:

- A conceptual reference design and building blocks for an active multispectral SWIR camera system based on field sequential waveband capturing (FSWC).
- A first analysis of approaches to motion compensation for multispectral FSWC-based imaging systems. The major challenge for these approaches is the intensity consistency assumption made by most motion detection techniques, which is in general not fulfilled by waveband-sequential multispectral imagery.
- A robust method for skin classification based on spectral signatures of material surfaces. It extends the work of Schwaneberg [25] to imaging sensors and uses both fast thresholding and more precise machine learning based classifiers in a hierarchical approach.
- A novel and robust cross-modal approach to detect spoofing attacks even in the presence of (partial) disguises and masks that enhances existing solutions based on the visual (VIS) spectrum. It ensures the authenticity of a face captured with a multispectral SWIR camera and verified against a known face given by a VIS image in a cooperative user scenario.
- A practical system design, setup and implementation of an active multispectral camera system optimized for skin detection with a focus on face recognition. The system acquires four-band multispectral image cubes in the SWIR range in real-time with optimized illumination homogeneity.
- An in-depth evaluation of the imaging system with respect to imaging quality, environmental influences and motion compensation, as well as skin detection and anti-spoofing performance. For this evaluation, a set of databases has been created using both an RGB camera and the presented multispectral camera system. The motion compensation performance is evaluated on a database of video sequences showing different test scenarios. Skin detection accuracy is evaluated on another database that contains spectroscopic measurements of skin taken from several selected locations on faces and limbs, as well as portrait

pictures of more than 150 participants of an extensive study. In addition, a third database contains images of spoofing attacks with a focus on masks and 3-dimensional facial disguises that are used for the evaluation of the anti-spoofing performance.

All created databases are available to the research community on the website of the Institute for Safety and Security Research (ISF) at the BRSU: `https://isf.h-brs.de`.

1.3 Outline

The contents of this dissertation are divided into eight chapters. *Chapter 2* describes the fundamentals and techniques related to multispectral SWIR imaging, skin detection and face recognition, as well as the terminology and notation used within this work. *Chapter 3* introduces design goals and the reference design for the skin detecting camera system and presents prior work in the related research fields.

In *Chapter 4*, approaches to motion compensation for field-sequential multispectral imaging systems are discussed. The proposed approach to skin detection on pixel-level based on the spectral signature of different material surfaces is described in *Chapter 5*, which also presents two methods to combine skin detection with face recognition in order to detect spoofing attacks.

Based on these methods, *Chapter 6* describes the system design, setup and implementation details of the *SkinCam* system, which implements the reference design proposed in Chapter 3. Furthermore, an analysis of the eye safety of the active illumination module, as well as an approach to depth estimation based on focus shifts in the different wavebands are presented here.

Chapter 7 presents an evaluation of the *SkinCam* imaging system and the proposed methods for motion compensation, as well as pixel-level skin and image-level spoof detection performance.

Finally, *Chapter 8* summarizes the approaches and findings presented in this dissertation and discusses aspects of possible future work and the use of the imaging system for other applications.

Chapter 2

Fundamentals

This chapter introduces the terminology and notation used in this dissertation. Furthermore, it gives a general overview of fundamentals and techniques in the fields that are relevant in the context of this work.

2.1 Terminology and Notation

2.1.1 Mathematical Notation

In this dissertation, pixel positions are denoted by their coordinates on the image plane given in braces, *i.e.* (x, y). Vectors are marked with an arrow and single elements of a vector are accessed by indices in square brackets: $\vec{s}[n]$ refers to the n-th element of vector $\vec{s}$. Estimations are marked with a hat, while precise or ground truth data is expressed without marks, *i.e.* $\hat{d} \approx d$. Similarly, interpolation results are marked with a tilde, *e.g.* $\tilde{C}_i$ is the result from interpolating between C_{i-1} and C_{i+1}. A change or difference of a variable is denoted by preceding it with a delta, *i.e.* Δd.

All additional notation will be described at first use.

2.1.2 Multispectral and Hyperspectral Imaging

Multispectral and hyperspectral imaging systems are capable of capturing high-density spectral information of a scene or object surface and thus offer several advantages over conventional single- or three-channel cameras. They are used for a variety of applications, such as remote sensing, astronomy, agriculture, medicine or food quality control [26], as well as high quality color image reproduction and conservation of art [27]. Multi- or hyperspectral imaging is not restricted to the visual (VIS) spectrum alone, but might also extend to the infrared spectral range [28].

The datasets acquired by these imaging systems usually consist of three dimensions: besides the two spatial dimensions, there is an additional spectral dimension. They are often referred to as *multi- or hyperspectral image cubes* [28–30], with every pixel (x, y) having a corresponding spectrum denoted as vector $\vec{s}(x, y)$ instead of a single (scalar) intensity value. A single "slice" of the image cube at a given *waveband* yields a monochrome image that represents the intensity of the scene captured in this waveband only. These slices are called *waveband images* or simply *channels* in this work. A waveband is defined by its *peak wavelength* λ_p and its *spectral bandwidth*, or *full width at half maximum (FWHM)* $\Delta\lambda_{0.5}$, respectively, which is measured between those points on the sensor's sensitivity curve at which the spectrum reaches half of its maximum amplitude [31]. In digital systems, the spectrum $\vec{s}(x, y)$ is represented in the form of a vector with n elements, where n is the number of wavebands. In the literature and in the context of this work, $\vec{s}(x, y)$ is denoted as the *spectral signature* of pixel (x, y) [32]. An example of a multispectral image cube is shown in Figure 2.1 on the next page, while Figure 2.2 illustrates the extraction of a low-density spectral signature out of a remission spectrum that has been captured by a single pixel of a corresponding imaging system.

The difference between multispectral and hyperspectral imaging is not clearly defined in the literature. Usually, they are distinguished by the number and width of the wavebands [29], with hyperspectral imaging having a much larger number of wavebands, covering a wide spectral range with high density, while multispectral imagers usually only capture a few selected wavebands [6]. This work focuses on methods that acquire images with a limited number of wavebands in a "staring imager" configuration having a fixed 2-dimensional field of view and that allow to capture scenes including moving objects or persons. Therefore, the term multispectral will be used rather than hyperspectral in the following.

2.1.3 Simultaneous and Field-Sequential Waveband Capturing

Simultaneous acquisition systems capture spatial and spectral information of an image simultaneously. For the acquisition of RGB color images in the VIS spectrum, for example, most modern digital cameras rely on a filter array, such as the Bayer filter mosaic, directly mounted on the surface of an image sensor to detect different wave-

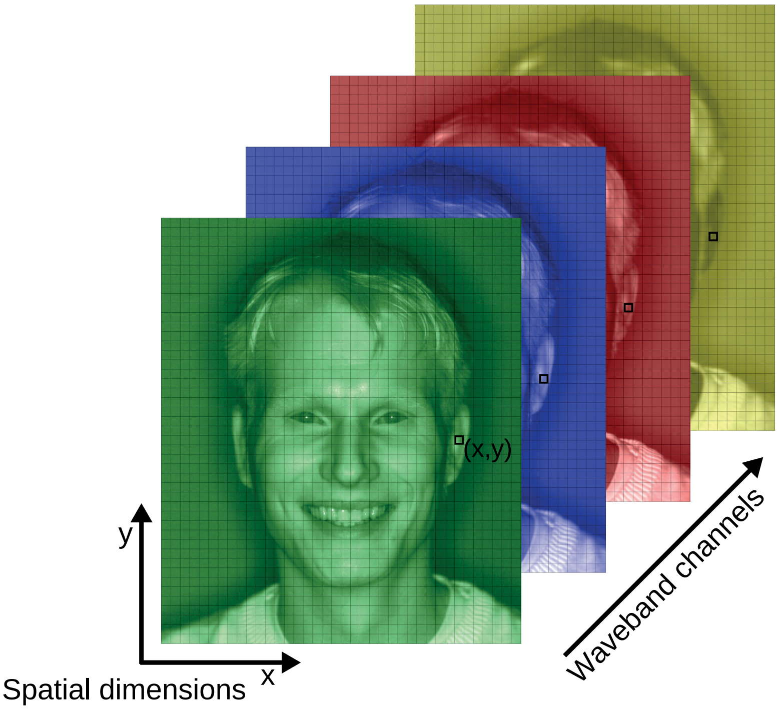


Figure 2.1: Illustration of a multispectral image cube. Waveband channels have been colored similarly to Figure 2.2 for illustration purposes. Based on [30, Fig. 1.1].

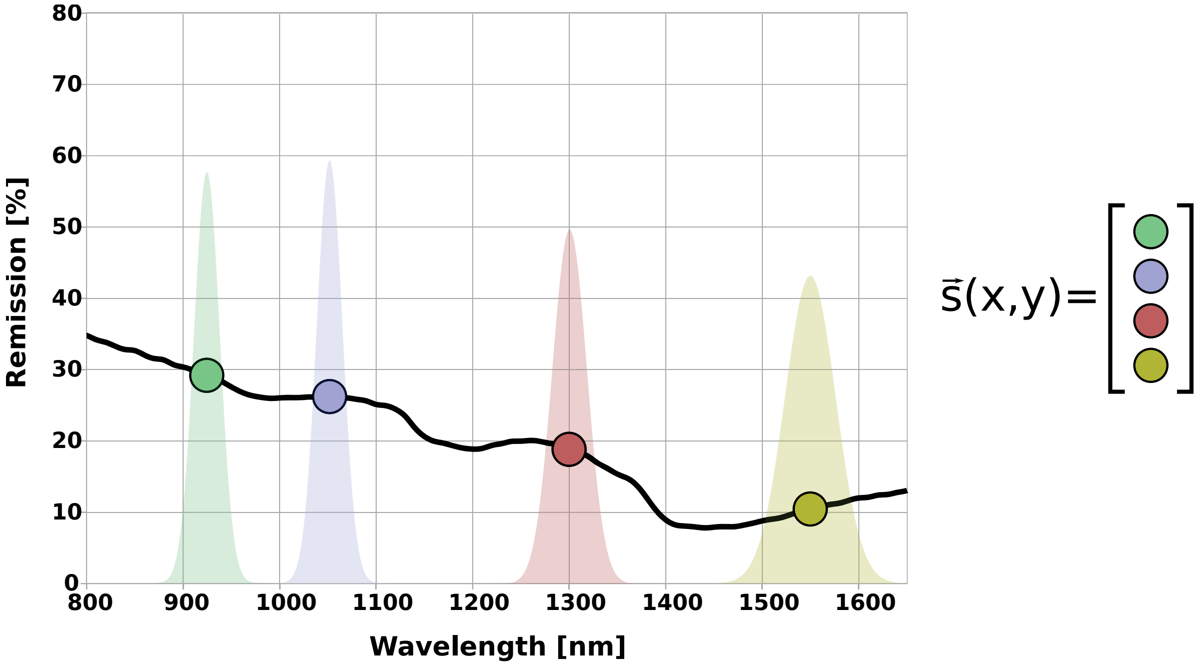


Figure 2.2: Example of a remission spectrum with indicated sensor sensitivity curves and the corresponding 4-dimensional spectral signature of pixel (x,y). Based on [30, Fig. 1.1].

bands with neighboring pixels at the cost of a reduced light gathering capability [33]. The Bayer filter pattern is 50% green, 25% blue and 25% red in order to mimic the spectral sensitivity of the human eye. To achieve a full color image, so called demosaicing algorithms have to be applied on the images captured using a Bayer filter. These algorithms interpolate the missing color information of each pixel from neighboring pixels, leading to a reduced spatial resolution of the final images. A different approach is used by the *Foveon* sensor, which separates the spectral channels using a grid of vertically stacked photodiodes by exploiting the different penetration depth of light in different wavelengths [33]. This way, the highest possible light gathering and spatial resolution is maintained. However, it's spectral sensitivity is comparably low. A third option is the use of *3CCD* cameras, which use dichroic prisms to split light into beams of different wavebands and acquire the different spectral channels with three separate sensors [33]. This ensures high spatial resolution and spectral selectivity, but requires precise spatial adjustment of the mirrors and sensors.

Despite their individual advantages and drawbacks, none of these *simultaneous* acquisition techniques is well suited for multispectral imaging if a "customized" or flexible selection of wavebands is required by a specific application, as complexity and cost will increase drastically with the number of wavebands. Therefore, common general purpose multispectral imaging systems use tunable or interchangeable band pass filters in combination with a single sensor that is sensitive to the full spectrum of interest. They acquire the spectral information of a scene by sequentially capturing images of single wavebands and combining them into one multispectral image cube in a second step. A common implementation of such systems uses bandpass (interference) filters on a rotating filter wheel in front of the camera, which is synchronized to the camera's exposure time [8,27,34]. A large variety of suitable filters with bandwidths of down to $\Delta\lambda_{0.5} \geq 10nm$ are commercially available. An alternative to rotating filter wheels are electronically tunable filters [28]). Compared to filter wheel systems, they offer only slightly better spectral resolution with bandwidths of several nanometers, but allow for more flexible configuration and higher numbers of wavebands. Similarly, the active multispectral camera system presented in this dissertation uses pulsed narrow band illumination instead of passive band pass filters to capture the spectral information of a scene with a single sensor. All of these approaches capture the spectral information of the scene (time-) sequentially. In the field of color imaging, this method is called *field sequential color capturing* [35]. Following this definition, this work will use the term *field-sequential waveband capturing (FSWC)* for this class of imaging systems. For simplicity, image sequences acquired using field sequential waveband capturing (FSWC) methods will further be called *waveband sequential*.

All FSWC imaging systems share one common problem: dynamic scenes with noticeable motion during the time required to capture all spectral channels will lead to motion artifacts, as boundaries and edge details of moving objects will not match between the different channels. Correcting these artifacts requires *dense motion estimation* to determine the direction and amount of motion for each pixel. Motion estimation has a long and successful history in computer vision; an overview is given in Section 2.5. However, existing state-of-the-art motion estimation techniques cannot handle FSWC imagery properly, as it strongly violates the intensity consistency assumption between adjacent channels, which most of these techniques rely upon [36]. Furthermore, FSWC motion compensation needs to be fast in order to be practically relevant. In Chapter 4, the problem of motion compensation for FSWC imagery is addressed in detail.

2.2 Physical Basis of SWIR Skin Detection

Already in 1955, Jacquez *et al*. [37] demonstrated that human skin has very specific remission characteristics in the infrared spectral range: its spectral remission above 1.2 µm is widely independent of the skin type, *i.e.*, the absorption spectrum of melanin, but mainly influenced by the absorption spectrum of water. This has been confirmed repeatedly in more recent research [38,39]. In a study with 330 subjects with different skin types and age, Schwaneberg [25] found a total variation of about factor two between the remission intensities of the darkest and brightest skin sample (average intensity over the full SWIR range), but identified very similar local maxima and minima in the different spectra.

In addition, the spectral remission of most other materials differs strongly from that of skin: Figure 2.3 on the following page shows the remission intensities of different material surfaces, including typical workpieces as well as examples of spoofs (printed and painted materials), compared to remission spectra of human skin in the visual and infrared spectral range up to 1.6 µm. Here, six different skin types, denoted as type 1 (very light colored) to 6 (very dark colored), are distinguished as proposed by Fitzpatrick [40]. RGB and (false color) multispectral short-wavelength infrared (SWIR) portrait images of six persons representing all of these skin types are presented in Figure 2.4 on the next page. As expected from the spectra, the obvious differences of the skin color in the RGB images are almost negligible in the SWIR images.

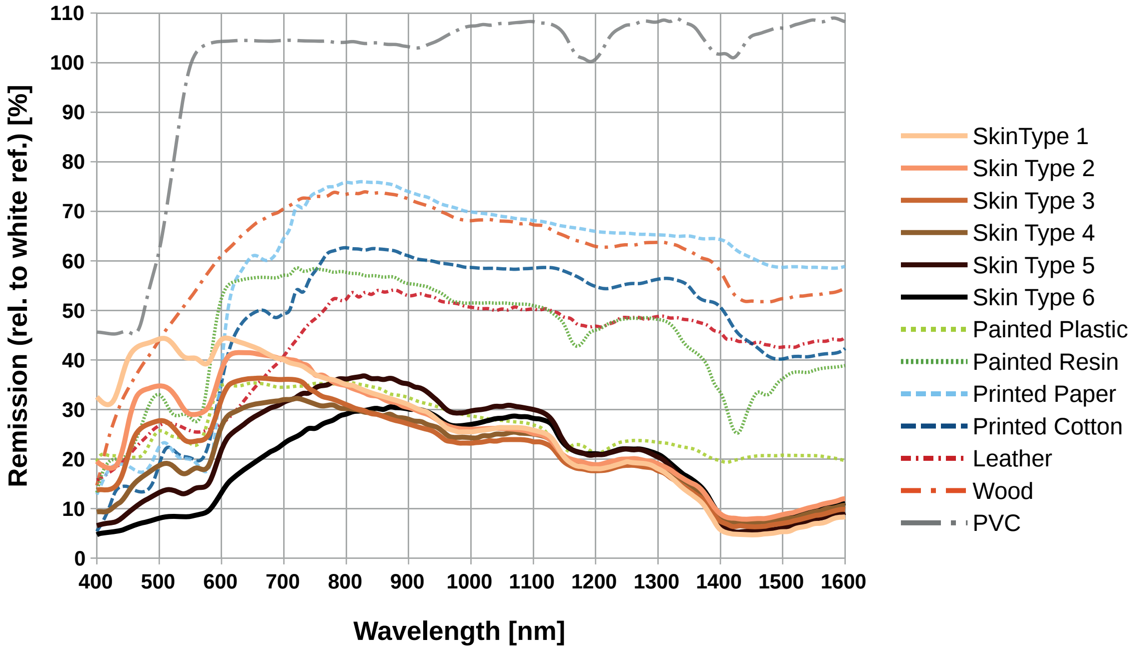


Figure 2.3: Spectral remission intensities of skin and different materials.

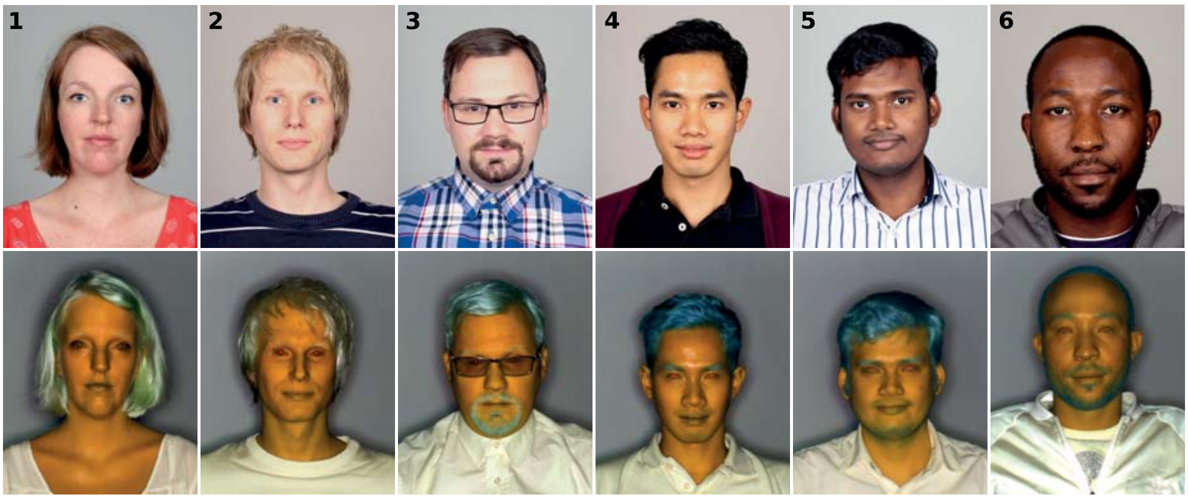


Figure 2.4: Visual spectrum (RGB color) and short wave infrared (false color) portrait images of skin types 1 to 6 according to Fitzpatrick [40].

2.3 SWIR Imaging Technology

Digital imaging sensors consist of an array of semiconductor detectors that is located at the focal plane of the imaging system and, thus, typically called the *focal plane array* [41]. Each detector in the focal plane array represents one pixel element or pixel in the final image. To detect a photon with a semiconductor detector, the photon's energy must be higher than the energy bandgap that separates the semiconductor's conductance band from the valence band in order for it to create an electron-hole pair. As the energy of a photon is determined by it's wavelength [42], the wavelength has to be lower than a specific cutoff wavelength λ_{cutoff}. This cutoff wavelength depends on the energy bandgap in the semiconductor material and can be calculated by

$$\lambda_{\text{cutoff}} = \frac{hc_0}{E_g} \approx \frac{1.24}{E_g}, \tag{2.1}$$

where E_g is the energy gap in electron volts [42], h ist Planck's constant and c_0 the speed of light. As silicon, which is most commonly used in imaging sensors and photodiodes for the VIS spectrum, has a bandgap of $E_g \approx 1.08\,\text{eV}$, its cutoff wavelength is at $\lambda_{\text{cutoff}} \approx 1.15\,\mu\text{m}$ [41]. Thus, silicon-based detectors are not suited to capture the characteristic spectral properties of human skin in the SWIR spectral range. In order to be able to detect photons in higher wavebands, a material with smaller bandgap than silicon has to be used.

A detector's capability to transform incident radiation into electric output is described by its responsivity, which measures the electrical output (in amperes) per incident radiant power (in watts) [42] and depends on the quantum efficiency of the used semiconductor material. The quantum efficiency denotes the ratio of generated electrons to incident photons. As shown in Figure 2.5 on the following page, with respect to its spectral responsivity, indium-gallium-arsenide (InGaAs) is a very well suited semiconductor material for the detection of the SWIR spectral range that is most interesting for skin detection. Due to its lower bandgap of $E_g \approx 0.73\,\text{eV}$ compared to silicon, InGaAs has a higher cutoff wavelength of $\lambda_{\text{cutoff}} \approx 1.7\,\mu\text{m}$ [41].

Besides responsivity, the strength and influence of noise is another relevant characteristic of a semiconductor detector. The most important sources of detector noise are shot noise and thermal noise [42]. Shot noise results from random arrival of photons at the detector [41]. It increases proportionally to the square root of the photo current, dark current and background radiation of the detector. In contrast to this, thermal noise originates in the thermal agitation of the electrons in the semiconductor material [42] and is independent of the incident power. Semiconductor materials with

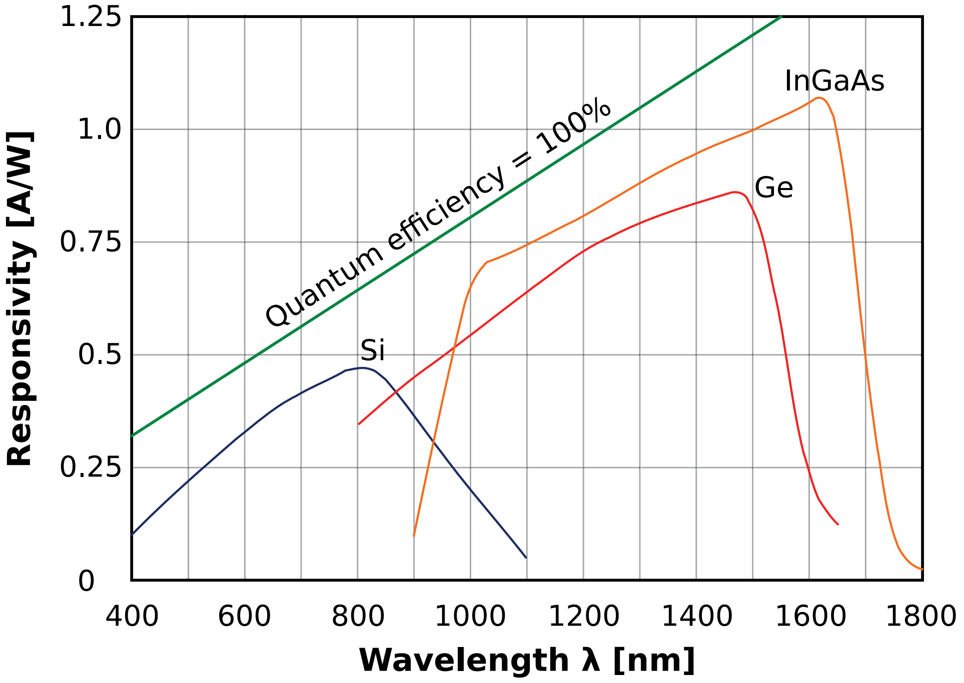


Figure 2.5: Spectral responsivity of photo detectors made of silicon (Si), germanium (Ge) and indium-gallium-arsenide (InGaAs). Adapted from [42, Fig. 1.70].

smaller bandgap are more susceptible to thermal noise than detectors with larger bandgaps [41] and thus require more cooling to achieve similar thermal noise levels to materials with larger bandgaps. Common InGaAs detectors, for example, are operated at 280 K, while silicon detectors are operated at temperatures of around 300 K and thus do not require active cooling at typical room temperatures.

The quality of a detected signal can be expressed by the signal to noise ratio (SNR), which is defined as the ratio of the effective incident power to the effective noise power [42] and typically given in decibel (dB):

$$\mathrm{SNR} = 10\log\frac{P_{\mathrm{sig}}}{P_{\mathrm{noise}}}dB. \tag{2.2}$$

An increase of the signal power P_{sig} by additional incident power ΔP will also lead to an increase of the noise power P_{noise} due to additional shot noise. However, the SNR will get better anyways, as shot noise increases only with the square root of ΔP:

$$\mathrm{SNR}' = 10\log\frac{P_{\mathrm{sig}} + \Delta P}{P_{\mathrm{noise}} + \sqrt{\Delta P}}dB. \tag{2.3}$$

When describing characteristics of semiconductor detectors, common parameters also include the *noise equivalent power* (NEP), which is defined as the signal power which is required to achieve an SNR of 1 and measured in watts, as well as the specific detectivity $D*$ [41], which is the inverse of the NEP normalized to the detector's photosensitive area A and frequency bandwidth BW with

$$D* = \frac{\sqrt{A \cdot BW}}{\mathrm{NEP}}. \tag{2.4}$$

Finally, a parameter often stated by camera manufacturers is the *dynamic range* (DR). Dynamic range of an image sensor is defined as the ratio of the largest signal that the detector can record without saturation Q_{max} to the smallest signal that can still be detected [43]. Similar to the SNR, it is often expressed in dB. Here, the smallest signal is defined as the standard deviation of the readout noise $\sigma_{\mathrm{readout}}$, which is measured under dark conditions. Thus, the dynamic range denotes an upper limit for the achievable SNR of an image sensor and is given by

$$DR = \frac{Q_{\mathrm{max}}}{\sigma_{\mathrm{readout}}}. \tag{2.5}$$

In future SWIR imaging systems, graphene might play an increasingly important role: recent research has shown that this two-dimensional crystalline material allows for very fast and sensitive detection of radiation in a very large spectral range, as graphene does not have a bandgap [44]. Projected NEP and specific detectivity are similar or even better than that of current InGaAs detectors. However, such detectors are not available on the market yet, as further research is needed to enable large-scale production and to improve readout times.

2.4 Chromatic Aberration in Optical Systems

Optical systems are typically designed and described based on "Gaussian optics", which is a simplified mathematical model using paraxial approximation. It assumes ideal conditions and perfect, reversible reproductions of object points to image points [45]. For real optical systems, deviations from this model are inevitable, not only due to flaws in the production of lenses, but also due to the wave nature of light that is not taken into account by the model. These deviations are the cause of different image defects, which are called *aberrations* [45]. Some of these aberrations are independent of the wavelength of the light and thus occur even for monochromatic light, *i.e.*, light of only one small waveband. These include coma or distortion, for example. The manufacturer of an optical system such as a camera lens can try to optimize the design in order to minimize these aberrations to a negligible level. Besides these monochromatic aberrations, there are also *chromatic aberrations*, which are caused by *dispersion* and result in two different effects [45]:

Transversal or lateral chromatic aberrations occur when light of different wavelengths is refracted at different angles due to a wavelength-dependent refractive index of a lens. Therefore, a single lens is not capable of focusing light of different wavebands to the same point on the image plane, unless it comes from an object point located on the optical axis.

Longitudinal or axial chromatic aberrations are caused by a shift in the focal length for light in different wavelengths, as the focal length also depends on the refractive index of the lens. As a result, a lens can only be correctly focused for light of one single wavelength, while other wavelengths will be out of focus, which leads to blur in the final image.

If the wavebands of interest are known in advance, *e.g.* for RGB cameras, manufacturers can address this issues by designing optical systems that use multiple lenses with opposite dispersion characteristics that cancel each other out [45].

2.5 Motion Estimation and Optical Flow

Detecting motion in an image sequence and estimating the motion direction and velocity is a very complex task [46]. The *optical flow* describes the velocity of apparent motion at the image plane as observed by a camera. It can be estimated by finding *displacement vectors* between features in two consecutive images A and B. If vectors can be found for every pixel, the set of displacement vectors is denoted as *displacement vector field* $F_{A \rightarrow B}$. Besides *dense optical flow* calculation, displacement vectors can also be determined by *block matching* methods. By applying the inverted displacement vector field $F_{A \leftarrow B}$ on image B, the apparent motion between the images is compensated. For an FSWC-based camera system used in dynamic environments, this process must be fast enough to allow for real-time compensation of image sequences.

As a fully comprehensive survey of motion estimation and optical flow techniques is beyond the scope of this work, interested readers are referred to the work of Fortun *et al.* [47] to get a deeper insight into optical flow computation methods, and to the survey on block-based methods by Jakubowski and Pastuszak [48]. Here, the description of methods is limited to those used in the context of this work.

The original approaches on the calculation of optical flow (OF) have been proposed by Horn and Schunck [49] as well as Lucas and Kanade [50]. They assumed that an object's intensity is constant between subsequent frames and every change in a pixels brightness must be due to motion. By using the brightness gradients and a constraint on motion smoothness, the flow velocity and direction can be computed. Brox *et al*. [51] extended this assumption by a gradient constancy constraint to deal with slight changes in brightness and an enhanced smoothness assumption. Another approach by Zach *et al*. [52] is based on total variation (TV) regularization using the L^1 norm (TV-L^1) and claims to be very robust against illumination changes and occlusions. Both, Brox *et al*. and Zach *et al*., are available as real-time capable GPU-based implementation. Werlberger *et al*. [53] proposed to replace TV regularization with the Huber norm (Huber-L^1) to further improve the results. They presented a library called FlowLib, which contains GPU accelerated implementations of their algorithm in different variations as well as Lucas-Kanade OF. Although the Middleburry OF ranking [36] lists a number of methods that have been proposed in the meantime and perform better in terms of accuracy, these algorithms can still be counted to the state of the art when processing time is taken into account as well.

In more recent work, Werlberger [54] proposed the use of alternative data terms than pixel intensity to better compensate for violations of the intensity constancy assumption: normalized cross-correlation (NCC), the census transform and consistency of gradients. All of these data terms represent the structure of the image content rather than the color or gray level intensity.

The basic idea of block matching (BM) is to divide one of the images into *macro blocks* with a *block size* of several pixels and to find the best match for each of these blocks within the second image using error functions such as the sum of absolute differences. To keep processing time low, the search range is limited to a maximum displacement (*p-value*). Testing every possible block displacement within this range is called *full search*. As boundaries of moving objects will not necessarily match with the macro blocks, blocking artifacts might occur in the compensated images. To avoid this, different techniques such as overlapping blocks, adaptive block size, multiscale approaches and filtering have been proposed [55]. To further reduce processing time, more efficient search strategies can be applied that try to reduce the number of calculations at the cost of accuracy [56]. Block matching can be implemented with a high degree of parallel computation using GPUs or FPGAs to achieve real-time performance.

2.6 Machine Learning Methods for Data Classification

In order to classify an observation or *instance* of data such as the spectral signatures of material surfaces automatically in two or more categories or *classes*, a *classifier* is required. While classifiers for simple problems can be defined by previous knowledge, for example, when black and white marbles are to be classified by their color, more complex problems require sufficient training data to find a suitable classifier. The process of training a classifier automatically using training data is called *supervised machine learning* [57]. A variety of methods for the design and training of classifiers has been proposed by researchers. In this work, the use of binary decision trees, random forests and support vector machines is evaluated for the problem of classifying spectral signatures into one of the two categories "skin" and "non-skin".

2.6.1 Binary Decision and Model Trees

Binary decision trees can be used to classify instances of input data into exactly one of $n \in \mathbb{N}$ distinct categories by performing a number of binary decisions that are arranged in the form a tree [57]. For every instance, decisions are applied during the traversal on one path along the edges of the tree from its root until a leave is reached, which is associated to one of the specified categories. An example for a decision tree is shown in Figure 2.6 on the next page. Due to the tree structure, the computational costs of traversal depend only on the depth of the tree d. As each decision only consists of one comparison of a single variable with a defined threshold, decision trees can be implemented very efficiently even on microcontroller systems with limited processing power and limited memory.

For the creation and training of a binary decision tree, different learning algorithms have been proposed in the literature, most of which follow a similar approach [57]. This basic approach has been introduced by Quinlan [58]. His ID3 algorithm and its successor, the C4.5 algorithm [59], construct decision trees in a top-down approach starting with the root of the tree and determine statistically how well each single attribute of the data classifies the training examples. If the attribute is numerical (one vector component of a spectral signature, for example), an optimal threshold to separate the training data using only this attribute as good as possible is calculated. This is evaluated using the *information gain*, which measures the separation of the training data with respect to their actual class and is calculated based on the entropy of all classes in the resulting subsets [57]. Finally, the threshold and attribute with the highest information gain is chosen for the first test and decision at the root node. Then,

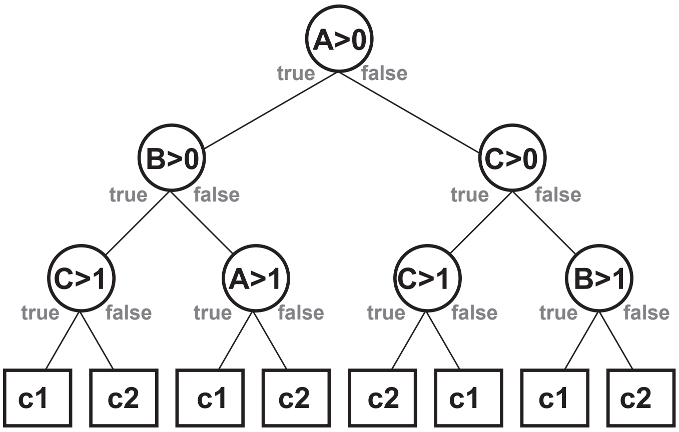


Figure 2.6: Example of a simple binary decision tree with depth $d = 3$ [25, Fig. 2.5].

descendants are created below the root node for both decisions and the procedure is repeated iteratively on the training examples in the remaining subsets.

As decision trees are not capable of learning and predicting continuous data values, Quinlan [60] also introduced the M5 learning algorithm that constructs so-called *model trees* which represent piecewise linear models. Instead of target classes, the leafs of a model tree are either simple output values or linear models (*i.e.* functions of the input value) that are created by combining lower branches using regression techniques.

2.6.2 Random Forests

A major disadvantage of binary decision tree learning is the tendency to *overfitting*: noise and outliers might lead to the growing of branches that specialize on these instances and the ability of the tree to generalize and classify previously unseen data correctly suffers [57]. Therefore, *random forests* have been proposed as an improvement of binary tree learning with advantages especially for large datasets with a high number of attributes [61]. Instead of constructing a single tree, this learning technique constructs a collection of decision trees which vote for the most popular class of a given instance in order to classify it. The individual trees are trained on a random subset of the available training data and also use only a random subset of the set of attributes to avoid specializing on specific training data or a limited number of very strong features [61]. Using a large number of trees, this method reduces the generalization error introduced by overfitting the classifier to the training data significantly.

2.6.3 Support Vector Machines

A *support vector machine (SVM)* is a supervised learning model based on the idea of using linear hyperplanes to separate instances of input data into one of two classes by mapping their feature vector non-linearly into a high dimensional feature space, thus allowing to separate even non-linearly separable data [62]. Although originally designed by Cortes and Vapnik for two-class classification problems only, the technique has since been extended and can be applied to multi-class problems (by subdivision into multiple two-class problems) and regression problems as well [63].

The original idea of using hyperplanes for the separation of data was introduced by Rosenblatt in 1958 [64], who proposed the *perceptron* as a theoretical brain model capable of learning, recognition and classification of data. The perceptron algorithm is an iterative procedure that finds a function

$$f(\vec{\mathbf{x}}) = \begin{cases} 1 & \text{if } \langle \vec{\mathbf{w}}, \vec{\mathbf{x}} \rangle + b > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

which classifies a new data instance using its attribute vector $\vec{\mathbf{x}} \in X \times \{-1,1\}$, with X being a dot product space, as either a positive or a negative instance. Here, $\vec{\mathbf{w}}$ is a vector of *weights* with the same dimension as $\vec{\mathbf{x}}$, $\langle \vec{\mathbf{w}}, \vec{\mathbf{x}} \rangle$ denotes the dot product of $\vec{\mathbf{x}}$ and $\vec{\mathbf{w}}$, and b is a *bias* or offset. For each training sample, the output of the function $y = f(\vec{\mathbf{x}})$ is calculated using the current weights and compared to the actual (desired) result d. Using a specified learning rate α, $0 \leq \alpha \leq 1$, all weights are updated with $\hat{w}_i = w_i + \alpha(d - y)x_i$ for all attributes. This procedure is repeated until the error is below a specified threshold. If the problem is linearly separable, the algorithm will converge to a solution that separates all training samples correctly [63].

Vapnik improved this approach by presenting a method that finds an *optimal* hyperplane, which he defined as a hyperplane that separates two classes with maximal margins on either side, thus ensuring a better generalization ability and robustness to previously unseen data samples [62]. It can be seen from Equation (2.6) that all possible hyperplanes are defined by

$$\langle \vec{\mathbf{w}}, \vec{\mathbf{x}} \rangle + b = 0, \;\; \vec{\mathbf{w}} \in X, \; \vec{\mathbf{x}} \in X, \; b \in \mathbb{R}, \tag{2.7}$$

where $\vec{\mathbf{w}}$ is orthogonal to the hyperplane as described in [63]. To find the decision function $y = f(\vec{\mathbf{x}})$ with optimal margins, the *objective function* $\tau(\vec{\mathbf{w}}) = \frac{1}{2}||\vec{\mathbf{w}}||^2$ has to be minimized while satisfying the *inequality constraint* $y_i(\langle \vec{\mathbf{x}}_i, \vec{\mathbf{w}} \rangle) \geq 1 \; \forall i = 1, \ldots, m$. From this *primal optimization problem*, the so-called *dual problem* can be derived, which is easier to solve and can be shown to have the same solution [63], by introducing *Lagrange multipliers* $\alpha_i \geq 0$ and expressing $\vec{w}$ as a linear combination of training vectors: $\vec{w} = \sum_{i=1}^{m} \alpha_i y_i \vec{\mathbf{x}}_i$. Then, the dual problem is given by

$$\max_{\alpha \in \mathbb{R}^m} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j \langle \vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j \rangle, \tag{2.8}$$

$$\text{with} \quad \alpha_i \geq 0 \ \forall \ i = 1, \ldots, m \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{2.9}$$

and leads to the following decision function [63]:

$$f(\vec{\mathbf{x}}) = \operatorname{sgn}(\langle \vec{\mathbf{w}}, \vec{\mathbf{x}} \rangle + b) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \langle \vec{\mathbf{x}}, \vec{\mathbf{x}}_i \rangle + b \right), \tag{2.10}$$

$$\text{with} \quad \operatorname{sgn}(z) \begin{cases} -1 & \text{if } z < 0, \\ 0 & \text{if } z = 0, \\ 1 & \text{if } z > 0. \end{cases}$$

All vectors with $\alpha_i > 0$ are denoted as *support vectors* and are located directly on the margin. All other vectors can be discarded for the calculation, as they will not have an influence on the hyperplane. An example of a resulting optimal hyperplane in a two dimensional space is illustrated in Figure 2.7 on the next page.

The described hyperplane can only be used to classify linearly separable datasets. Therefore, Cortes and Vapnik [62] proposed to map non-linearly separable input data from the input space X into a *feature space* $\mathcal{H}$ of higher dimensionality, in which the data is linearly separable, by using a transformation function $\Phi : \vec{\mathbf{x}}_i \in X \rightarrow \tilde{\mathbf{x}}_\mathbf{i} \in \mathcal{H}$. To reduce the computational costs, the so-called *kernel trick* is applied by using a positively defined kernel k as substitution for the Euclidean dot product [63] with

$$\langle \Phi(\vec{\mathbf{x}}), \Phi(\vec{\mathbf{x}}_i) \rangle = k(\vec{\mathbf{x}}, \vec{\mathbf{x}}_i), \tag{2.11}$$

which leads to the final decision function

$$f(\vec{\mathbf{x}}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i k(\vec{\mathbf{x}}, \vec{\mathbf{x}}_i) + b \right). \tag{2.12}$$

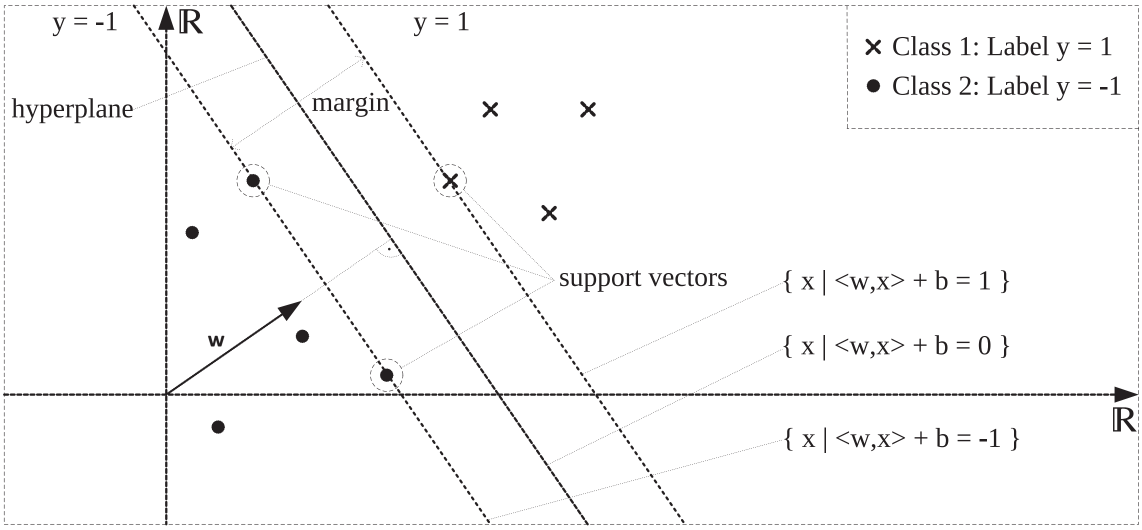


Figure 2.7: Example of linear separation of two classes using a hyperplane with optimized margins [25, Fig. 2.3], compare to [63, Fig. 1.5].

In order to be more robust to noise, errors or outliers in the training data, Cortes and Vapnik [62] additionally proposed a derivative of this method with *soft margins*, which introduces slack variables $\xi_i \geq 0$ to allow violations of the optimization constraints. Each violation leads to an increase of ξ_i. Thus its sum, scaled by a positive constant C, is added to the optimization problem in order to find a balance between the acceptance of errors and the size of the margins:

$$\min_{\vec{\mathbf{w}} \in X, \xi^m \in \mathbb{R}} \quad \frac{1}{2}||\vec{\mathbf{w}}||^2 + \frac{C}{m}\sum_{i=1}^{m} \xi_i, \qquad (2.13)$$
$$\text{with} \quad y_i(\langle \vec{\mathbf{x}}_i, w \rangle) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \; \forall \; i = 1, ..., m.$$

This derivative is denoted as *C-SVM* and will be used in the context of this work.

2.6.4 Evaluation of Classification Performance

To evaluate the performance of classifiers, a common notation has been established in the literature: correct classifications of the target class (here, “skin”) are called true positives (TPs), while correct classifications of the non-target class (here, ”non-skin”) are called true negatives (TNs) [65]. Incorrect classifications are called false positives (FPs) if instances of the non-target class are falsely mapped to the target class (*i.e.* ”non-skin” samples are classified as “skin”), or false negatives (FNs) if instances of the actual target class are falsely mapped to the non-target class (*i.e.* “skin” samples are classified as “non-skin”). This data is often presented in a *confusion matrix* and used to calculate the performance metrics *accuracy* and *precision* [66]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2.14}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.15}$$

The ratio of true positive outcomes compared to the total number of instances is denoted as true positive rate (TPR). Similarly, the share of the other possible outcomes are denoted as true negative rate (TNR), false positive rate (FPR) and false negative rate (FNR), respectively. TPR and TNR are often also denoted as *sensitivity* and *specificity* [65].

2.7 Face Recognition

2.7.1 Categorization and Processing Flow

Face recognition is used for a variety of applications. Most of these applications require that users of the face recognition systems must be known to the system, *i.e.*, their biometric features must have been added to the system’s database. This process is called *enrollment*. The different applications can be roughly categorized by the used operating modes of the face recognition system and their usage scenarios [67]. The two operating modes are face verification and face identification:

Face verification compares a query face image with a known face image of a person whose identity is being claimed in a one-to-one matching process. This person has to be *enrolled* in the face database of the face recognition system in advance.

Face identification compares a query face image with multiple face images in a face database of previously *enrolled* persons in a one-to-many matching process. Here, no specific identity is being claimed. Instead, the system attempts to find the best match and, thus, identify the person shown on the query image.

The user scenarios are defined by a user’s cooperation with the process [67]:

Cooperative user scenarios denote applications in which the user is willing to cooperate with the system in order to gain access or to be granted a requested privilege. Examples for these applications are access control systems or eGate systems for automated passport control; see Section 1.1.1. Here, a user can be expected to present his face to the recognition system in a favorable way, for example by removing glasses or headwear and in frontal pose.

Noncooperative user scenarios differ from cooperative user scenarios by the fact that a user does not know that his face is being captured or does not want to be recognized, which renders this scenario much more difficult to handle. Typical application scenarios are found in the field of surveillance.

In this dissertation, the focus with respect to face recognition is on face identification and verification applications in *cooperative* user scenarios.

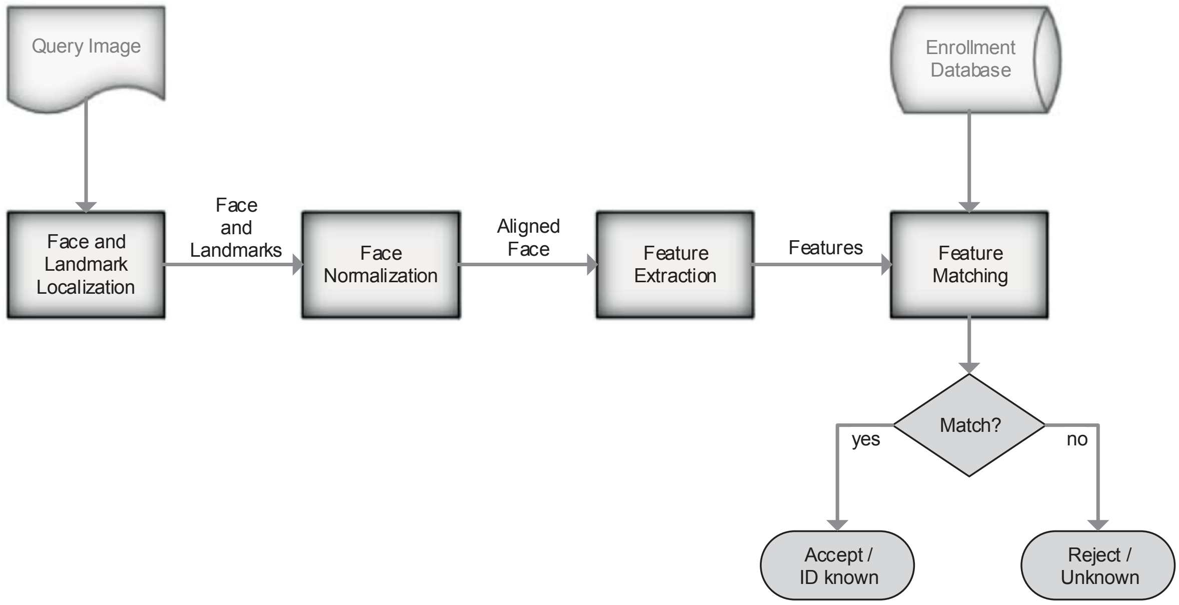


Figure 2.8: Processing flow of a face recognition system. Adapted from [67, Fig. 1.2].

The face recognition process typically consists of four building blocks: face detection and landmark localization, normalization, feature extraction and, finally, the actual face recognition or matching [67]; see Figure 2.8. In *face detection*, the location and scale of a face in the query image is estimated and the facial area is segmented from the background. Then, the location of the *facial landmarks* such as eyes, nose and mouth is extracted by a *landmarking* algorithm. Their exact location is required even for recognition algorithms that don't rely on geometrical constraints [68].

In the *face normalization* step, the image of the face is normalized with respect to both illumination and geometrical alignment [67]. The latter is achieved by applying affine transformations on the face image based on the locations of the detected landmarks and cropping it afterwards to match a defined standard frame. The following *feature extraction* step uses this aligned image to detect specific features, the type of which strongly depends on the type of classification or matching technique used for recognition [68]. Finally, the *feature matcher* compares the features extracted from the query image to the features of a specific identity (verification) or all known identities (identification) that are stored in the enrollment database. If a match is found with a confidence value that is above a predefined threshold, the query image is accepted or the ID of the identified person is displayed.

2.7.2 Overview of Techniques and Approaches

Especially the matching of facial features in the last step of the processing flow is the biggest challenge of face recognition systems [67]. A variety of different methods to address this problem has been proposed since the development of the first automated face recognition system in 1973. The different approaches can be classified into three categories: structural feature-based methods using geometry and/or appearance of facial features, holistic template matching methods, as well as hybrid methods, which combine both approaches in one classifier [68]. The geometry-based approaches were used especially in the early development phase of face recognition systems [67]. They have advantages with respect to data reduction and robustness to varying illumination conditions and poses, but do not use the information contained in the appearance or texture of a face. In contrast to this, the holistic methods are strongly relying on the overall appearance and thus require a large amount of training data to train effective classifiers, but achieve significantly higher accuracy in face recognition.

Principle component analysis (PCA) and linear discriminant analysis (LDA) are two prominent examples of holistic methods. PCA tries to find the most relevant features in a set of images and uses these as basis vectors in the face image space, which are called *Eigenfaces*, to represent each individual face as a linear combination of them [69]. LDA is closely related to PCA, but attempts to model the differences between faces rather than the similarities. It is used in the approach called *Fisherfaces*, which was found to perform much better than PCA for face recognition [69]. For more detailed information on the history of face recognition methods, please refer to the comprehensive survey by Zhao *et al*. [68], which contains methods that were proposed for face recognition until 2003.

More recently, filter-based methods for the extraction of local (appearance) features, such as the responses from applying Gabor filters on the face images or the calculation of local binary patterns (LBPs), have been proposed [67, 70]. Especially LBPs were found to be very successful. LBP features are constructed by dividing the image into several regions and thresholding all pixel values in this region with the value of the pixel in the center. The result of all pixels is considered as a binary number and used as image descriptor [70]. Lately, Taigman *et al*. [71] presented a deep neural network trained on face recognition. A neural network is very similar to the perceptron (see Section 2.6.3), but consists of multiple layers that are connected with each other. The authors reported a performance that is significantly better than the state of the art.

2.7.3 Face Matching Between SWIR and VIS Images

In the context of this work, face images captured in the SWIR spectral range will be used for face recognition. If these images could be matched to an already existing enrollment database created using VIS images, the applicability and acceptance of SWIR-based face recognition systems could be significantly increased. As infrared imaging with active illumination that is invisible to the human eye has advantages for any face recognition approach [72], this problem has been addressed by several researchers in the recent years. Although most of this prior work is based on the spectral range below 1 µm, the findings are still comparable to the images captured by the system proposed in this dissertation.

In 2007, Yi *et al*. [72] proposed a correlation-based learning approach to match near infrared (NIR) to VIS images and achieved a TPR of 93.1% with a FPR of 0.1%. Three years later, Klare and Jain [73] combined different feature descriptors with an LDA-based classifier and achieved a TPR of more than 94% with an FPR of 1%. In 2011, Goswami *et al*. [74] presented a database and an evaluation protocol for the evaluation of cross-spectral face recognition methods and evaluated a number of combined methods on this dataset. They concluded that, given proper preprocessing, good results can be achieved with LBP-based classifiers. In 2014, Zhu *et al*. [75] presented a model called Transductive Heterogeneous Face Matching (THFM) that attempts to solve the cross-spectral matching problem. This approach allegedly outperforms all state-of-the-art approaches on relevant benchmarks and achieved a TPR of 99.66% with an FPR of 1%.

Furthermore, it was found that several commercial of the shelf face recognition systems, including FaceVACS from Cognitec Systems GmbH, are also capable of solving this cross-spectral matching problem with good performance [73]. In conclusion, the use of SWIR images for query with a VIS-based database appears to be feasible.

Chapter 3

Concept of a Skin-Detecting SWIR Camera System

This chapter defines design goals for a skin detecting camera system and presents prior and related work in the relevant research fields: skin detection in general, multispectral imaging, motion compensation, as well as face verification and anti-spoofing. Based on the formulated design goals, a reference design consisting of several building blocks is proposed.

Publications: Design goals for a skin detecting camera system with a focus on face verification and related work on skin detection have partially been published in [21]. Some portions of the related work on multispectral imaging and motion compensation are included in [21, 24] and the related work on facial anti-spoofing has been covered in [22].

3.1 Design Goals

In general, biometric face recognition systems imply strong requirements with respect to robustness and speed of the detection. Here, robustness includes both accurate detection under varying external conditions such as lighting and a reliable detection of skin. Even though this work does not tackle any specific application scenario, the following rather generic design goals can be formulated that allow the realization of various applications:

- The imaging system should be (widely) independent of ambient light. The spectral distribution or any flickering of nearby light sources must not disturb the skin detection process, see Section 6.1.
- The acquisition and processing time should be as short as possible.
- Moving objects must not lead to false classifications, see Section 3.2.2.
- Skin detection must work independent of a subject's skin type, age or gender.
- Classification accuracy must be high enough to detect and reject both two- and three-dimensional spoofing attacks, even if they are made of skin-like material, see Section 5.2.
- Skin detection and face recognition must be matched to ensure that a recognized face is valid, see Section 5.3.
- The operation range should cover practically relevant distances for indoor scenarios. For eGates, for example, typical distances are 1 m to 3 m.

It will be shown that none of the existing approaches can reach all of these goals.

3.2 Related Work

3.2.1 Imaging and Non-Imaging Skin Detection

This section focuses on work that is directly related to the approach proposed in this dissertation, *i.e.* skin detection based on short-wavelength infrared (SWIR) radiation. For skin detection by color in the visual (VIS) spectrum, please refer to the comprehensive survey paper by Kakumanu *et al.* [76].

The advantages of the SWIR spectral range over the VIS spectrum for skin detection have been confirmed in the literature repeatedly, for example in recent work by Mendenhall *et al.* [77], who state that state-of-the-art skin detection methods based on the VIS spectrum achieve true positive rates (TPRs) of $\approx 90\%$ and false positive rates (FPRs) of at least 2% and up to 15%. Using the SWIR spectrum, much better results can be achieved. Prior and related work on this topic can be divided into two categories: imaging and non-imaging approaches.

(Non-imaging) point sensors

As early as 1985, Hacskaylo [78] patented an "automatic human body detector" based on SWIR radiation: three narrow wavebands around 1.22 µm, 1.50 µm and 1.72 µm are analyzed using active (broadband) illumination and three detectors with different narrow band pass filters. The remission intensities acquired by these detectors are normalized and compared in order to detect the spectral signature of skin.

In 2005, Kilgore and Whillock [79] filed a patent describing a skin detecting sensor based on the same principle that can be implemented using either one active (broadband) illumination source combined with two spectrally selective detectors or two active illumination sources emitting different wavebands and one broadband detector. For both implementations, the two filters are used to divide the spectrum of the radiation that is reflected from a surface material into a first waveband of approximately 800 nm to 1 400 nm and a second waveband of approximately 1 400 nm to 2 200 nm. To classify the surface material into "skin" or "non-skin", a weighted difference and threshold is applied on the two remission intensities. The exact same patent was filed again in 2007 by Determan and Wunderlin [80].

Zhang *et al*. [81] described an approach based on light emitting diodes (LEDs) in the two wavebands of about 850 nm and 1 450 nm in combination with a respective photodiode as detector. Compared to [79] and [80], this results in rather narrow wavebands. The authors specifically expected a varying distance between the sensor and the analyzed object and trained a support vector machine (SVM) classifier on multi-distance reflectance data.

Schwaneberg *et al*. [25, 82] developed a similar yet more enhanced point sensor for skin detection with a focus on safety applications. The sensor uses LEDs in four wavebands around 830 nm, 1 060 nm, 1 300 nm and 1 550 nm, which have been selected as result of an extensive study with a total of 330 persons and avoid the absorption band of water vapor, which is located in the range from 1 340 nm to 1 450 nm. Through extensive beam forming, the sensor can be used in operation ranges of 0.1 m to 1.0 m. In contrast to all previous approaches, it is designed to deal with environmental influences such as varying illumination and measures the distance to an object surface with high accuracy in order to correct for distance-dependent distortions in the remission intensities. Using an SVM-based classifier, it achieves very high classification accuracy in all evaluated application scenarios. This sensor system can be regarded as a predecessor of the system that is presented in this dissertation.

(Imaging) camera systems

A disadvantage of point sensors is their limitation to applications that do not require detailed spatial resolution. A single point sensor can neither be used for face recognition, nor for the monitoring of large areas at robot workplaces. To address such applications, imaging-based skin detection systems have been presented in prior work.

Pavlidis *et al*. [5,83,84] demonstrated that the SWIR range has many advantages for skin detection in general and for disguise detection in specific. They proposed a dual band camera system, consisting of two co-registered cameras, with one camera having a spectral sensitivity below 1400 nm (ideally 800 nm to 1400 nm) and the second camera having a spectral sensitivity above 1400 nm (ideally 1400 nm to 2200 nm). This way, both wavebands are captured simultaneously. Their system can work with either sunlight or artificial illumination and uses a fusion algorithm based on weighted differences to detect skin in the acquired images. Depending on the spectral distribution of the illumination source, the weighting factors have to be adapted, as the system is not independent of ambient light. Originally, their system was meant to be used for the automatic detection of vehicle occupants to control the legitimate use of freeway lanes that are reserved for car pools, *i.e.*, cars used by more than one person. The authors conclude that their system achieves very good skin detection, as well as face and disguise detection capabilities compared to systems in the visual spectrum, only limited when it comes to the detection of surgical face alterations, where they see an advantage of systems using the thermal infrared range. In a later publication of the group [85], they presented an extension of the system with a third camera for the visual spectrum and a more advanced face detection approach that included multi-band eye and eyebrow detection. Their system uses beam splitters to allow all cameras to view the scene from the same vantage point in order to avoid problems with image registration.

Chang *et al*. [86] presented a multispectral imaging system for face recognition applications that uses the near infrared (NIR) spectrum up to 1100 nm. Their imaging system consists of a camera with an attached liquid crystal tunable filter and allows to capture 12 wavebands in the range from 660 nm to 1100 nm using field sequential waveband capturing (FSWC). The authors found that their approach increases the face recognition performance in challenging conditions. However, they did not evaluate the influence of varying skin types, which still have an influence on remission intensities in this wavelength range.

At the U.S. Air Force Institute of Technology, Nunez and Mendenhall [6, 7] researched the use of hyperspectral SWIR imagery to detect skin in the context of remote sensing applications. The authors acquired images in 81 narrow spectral bands from 900 nm to 1744 nm with a hyperspectral camera and introduced a detailed reflectance model of human skin based on this data. For real-time and in field use, the authors propose a multi-camera system to acquire images in distinct narrow wavebands simultaneously using different band pass filters on each camera, further described by Peskosky [87]. To avoid problems with image registration, this system uses dichroic mirrors to split up the beam so that all cameras share one single lens and view the scene from the same vantage point, similar to the approach of Pavlidis *et al*. Both this and more recent work of the group [77] proposes a combination of the VIS and SWIR spectral ranges.

Bourlai *et al*. [8] presented a multispectral SWIR image acquisition system with a focus on face recognition that uses a single camera with an attached rotating filter wheel. The filter wheel is equipped with five band pass filters with a full width at half-maximum (FWHM) of 100 nm around the peak wavelengths 1 150 nm, 1 250 nm, 1 350 nm, 1 450 nm and 1 550 nm. By synchronizing the camera's integration time to the filter wheel, the system can capture all five waveband images sequentially using FSWC within 260*ms*, *i.e.*, at a rate of ≈ 3.8 frames per second (FPS).

Bertozzi *et al*. [9] propose a camera with a broadband sensor sensitive to both the VIS and SWIR spectral range from 400 nm to 1700 nm that is equipped with a Bayer-like mosaic filter pattern directly on top of the pixel array to capture different wavebands simultaneously. One clear filter is combined with three high pass filters with cut-off wavelengths of 540 nm, 1 000 nm and 1 350 nm. By subtracting the acquired values of neighboring pixels with different filters, multispectral images with the four wavebands of approx. 400 nm to 600 nm, 600 nm to 1 000 nm, 1 000 nm to 1 300 nm and 1 300 nm to 1 700 nm can be calculated.

Due to the passive, filter-based system design, the spectral distribution of the ambient illumination has a strong influence on the multispectral images acquired by any of these systems. There are also individual disadvantages of all of these systems: both the approaches of Pavlidis *et al*. and Peskosky are based on a very limited number of wavebands, which leads to coarse spectral resolution and a low classification performance when dealing with skin-like material surfaces. The system presented by Bertozzi *et al*. suffers from a reduced spatial resolution due to the mosaic filter pattern and uses comparably wide wavebands, which will also be a disadvantage with respect to classification performance. The system by Chang *et al*. is not independent of varying skin types and has a relatively slow acquisition speed, which also applies to the filter wheel system by Bourlai *et al*. In contrast to this, the approach proposed in this dissertation uses active narrow band illumination instead of filters and combines a comparably high acquisition speed with high spectral resolution. As shown in Chapter 7, it is widely independent from ambient light and allows for a robust detection of human skin.

3.2.2 Motion Compensation

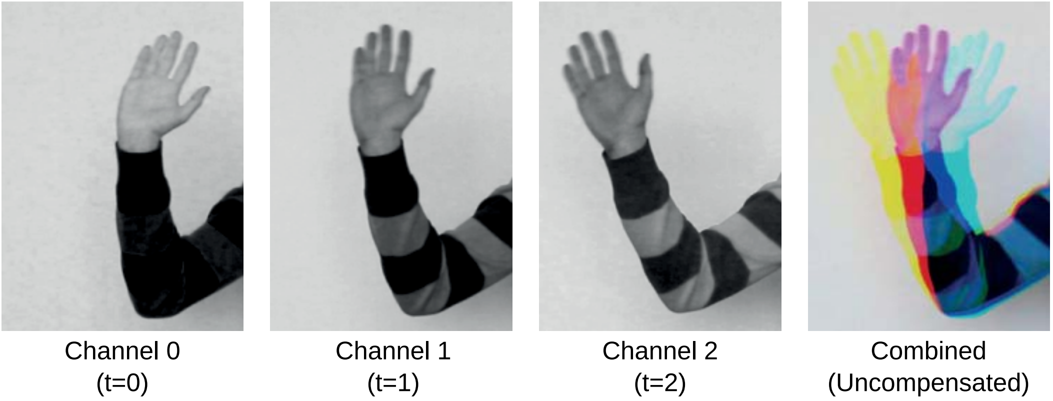


Figure 3.1: A waving hand recorded using field-sequential color capturing, with channels captured at subsequent times t. When the channels are combined in a multispectral image without motion compensation, the color breakup effect occurs.

All of the described FSWC methods require that a scene is static during the acquisition time of a full multispectral image cube [31]. Otherwise, motion artifacts might occur in the form of an effect called *color breakup* in color imaging [35]: boundaries and edge details of moving objects will not match between the different spectral channels and color fringes will appear in the final image. An example for such artifacts is shown in Figure 3.1. Due to this issue, it depends on the specific application and expected amount of motion in the scene whether FSWC-based implementations can be used for multispectral imaging with acceptable results. By compensating the motion artifacts, its applicability can be greatly extended.

To the best of the author's knowledge, there has not been any previous research on motion compensation specifically for FSWC imaging systems. However, a similar problem can be found in time of flight cameras: in order to acquire a full depth image, a sequence of four phase images has to be captured. Object or camera motion during the acquisition of the four phase images leads to motion artifacts in depth estimation that need to be compensated.

Lindner *et al*. [88] proposed a motion compensation approach based on optical flow (OF) calculation using state-of-the-art algorithms for this purpose: OF is calculated between each phase image $P_i,\ i=1,\ldots,3$ and P_0 and every $P_i,\ i=1,\ldots,3$ is resampled accordingly. Lefloch *et al*. [89] improved on this approach by reducing the required number of flow fields $F_{0\leftarrow i}$ for each depth image from three ($F_{0\leftarrow 1}, F_{0\leftarrow 2}, F_{0\leftarrow 3}$) to two ($F_{0\leftarrow 2}, F_{1\leftarrow 3}$), leading to faster processing times. The third flow, which is still required for the compensation, is derived from $F_{0\leftarrow 2}$ and $F_{1\leftarrow 3}$. Using graphics processing units (GPUs) for acceleration, this approach is capable of real time motion compensation with ≈ 25 FPS. An even faster approach for applications with only small motion displacements has been presented by Hoegg *et al*. [90]: in a first step, a binary motion image I_b is calculated to restrict motion estimation to image areas with apparent motion only. Then, motion flow is estimated for all pixels in I_b within a small window for all phase images simultaneously by assuming a linear and constant motion and calculating the error for all possible flow vectors within the motion window. To further reduce processing time, the search space can be reduced by using the mean direction angle of motion in a previous frame as an initial guess for the current frame. Using a window size of five pixels, this method achieves processing times of $\approx 10\,\mathrm{ms}$ and frame rates of up to 100 FPS. Note that all of the aforementioned approaches use the PMD-ToF cameras which offer the option to deliver the full intensity for each phase image. Thus, standard optical flow methods can be applied here.

A different approach on motion artifact correction for ToF cameras that requires significantly less computational effort has been proposed by Schmidt and Jähne [91]. Their method detects the phase images affected by motion artifacts for every single pixel and reconstructs a valid pixel value from the previous frame, assuming that there is at most one event leading to motion artifacts during these two frames. Furthermore, this approach does not require full intensity images, but uses phase images only. A related approach has been presented by Jimenez *et al*. [92] more recently, which requires only one single frame for the detection and correction of motion artifacts by reconstructing valid pixel values either from unaffected previous phase images or from unaffected pixels within the local neighborhood. However, in contrast to the approach by Schmidt and Jähne, the latter can not easily be adopted for imagery acquired using FSWC due to its time of flight specific working principle.

In this dissertation, different approaches to motion compensation for waveband-sequential image sequences will be presented and evaluated. As shown in Chapter 4, even state-of-the-art algorithms need to be modified or extended by appropriate preprocessing in order to achieve good results.

3.2.3 Face Verification and Anti-Spoofing

Based on the two operating modes of face recognition systems, see Section 2.7.1, two scenarios for spoofing attacks on these systems are distinguished in this work:

Counterfeiting of a foreign identity: here, an attacker tries to imitate the identity of a specific person to attack *face verification*. A simple example for counterfeiting attacks are the so-called *presentation attacks* using printed photos of another persons face that are held in front of the attacker's face.

Disguise of the own identity: here, an attacker tries to disguise his own identity to avoid *face identification*, for example, by wearing (partial) masks, fake noses or facial hair, or by applying make-up.

In the past few years, several researchers have addressed the problem of spoofing attacks on face recognition systems. Although some countermeasures for such attacks have been proposed, recent studies clearly point out that especially attacks with facial disguises and masks are still a severe problem for current anti-spoofing techniques [1, 12, 93]. Due to the widespread availability of 3D scanners and printers, the creation of facial masks has become much easier in recent years [93]. These masks can be manufactured using different materials with varying textures and surface properties, such as plastics, resin, silicon, rubber or latex. Applying paint or make-up makes the visual appearance and texture of a mask nearly identical to a real face. In combination with the variations found in human skin color and texture, distinguishing any possible spoof from genuine human skin is a very difficult task using only the VIS spectrum [5].

In the following, the state of the art in the field of face anti-spoofing is presented. The prior work is divided into two categories: approaches that are based on the visual spectrum alone and approaches using different or additional modalities.

Anti-Spoofing in the Visual Spectrum

As most face recognition systems rely on inexpensive monochrome or color cameras for the VIS spectrum, there is a large variety of approaches to anti-spoofing that use image processing based on these images only. Presentation attacks with still images can effectively be detected using motion based approaches: Kollreider *et al*. [94], for example, proposed to analyze and compare the trajectories of face parts by optical flow estimation. A similar approach has been presented by Wang *et al*. [95]: their method tracks facial features to recover the 3D structure of a (real) face. A different method by Anjos and Marcel [96] relies on the detection and correlation of motion patterns between head movements and the scene context. All of these methods require several images or a video sequence as input and can be deceived with a so-called replay attack by displaying a video sequence on a mobile device, for example.

According to Nixon *et al*. [11], different challenge-response methods have been described in the literature and implemented in commercial software, for example requiring the user to blink or smile at the right time or to repeat randomly generated phrases. If such measures are sufficiently random and comprehensive, they can significantly increase the difficulty to deceive a recognition system using presentation or replay attacks at the cost of reduced user comfort and recognition speed.

Another approach to spoof detection is the analysis of the texture of a face or the detection of image artifacts on a presented photo or video spoof. Maattaa *et al*. [97] were among the first researchers who exploit changes in the micro textures of an image during re-capturing. A similar method is used by Yang *et al*. [98], who proposed a component-based face coding approach to detect spoofs based on micro texture differences. The authors demonstrate good performance on three standard databases containing spoofing attacks using printed photos and videos on mobile devices. More recently, Mei *et al*. [99] combined spatial and temporal information in a new descriptor that outperformed the previous state of the art. Specifically for spoofing attacks using mobile devices, an approach by Buciu and Goldenberg [100] relies on the detection of oscillating patterns, while Patel *et al*. [14] described a method to detect moiré pattern aliasing artifacts. Both kinds of image artifacts occur when images or videos shown on a screen are recaptured by the face recognition system.

By combining different approaches, both reliability and applicability can be increased. Komulainen *et al*. [101], for example, proposed a combination of motion and texture analysis, while Yan *et al*. [102] combined facial motion with background consistency and an analysis of image banding effects.

However, none of these approaches sufficiently addressed the problem of detecting (three-dimensional) facial masks. By principle, they can neither be detected by 3D structure or scene context analysis, nor by challenge-response methods or the detection of image artifacts. Distinguishing masks with applied make-up from faces with applied make-up will also be difficult for any texture analysis method. Therefore, the approach proposed in Chapter 5 of this dissertation focuses especially on such masks and partial disguises and can detect them reliably, as shown in Chapter 7.

Anti-Spoofing Using Different Modalities

A variety of approaches to address spoofing attacks using additional modalities have been proposed in prior work. The acquisition of 3D depth images allows to reliably reject presentation and replay attacks with printed photos or mobile devices, as they consist of a flat surface [11]. To further detect attacks using masks, Kose and Dugelay [103] presented an approach that combines the texture analysis of 2D facial images from [97] with 3D depth images. However, their results were still far from reliable.

Dhamecha *et al*. [104] proposed the combination of images acquired in the visible and thermal infrared (TIR) spectra to detect spoofing attacks. They define patches on a detected face, classify each patch as authentic or disguised and use only the authentic patches for recognition. A shortcoming of this approach is that the patches are rather big and that small but possibly important details might be overlooked.

The most promising results so far have been achieved using multispectral SWIR imagery. Pavlidis and Symosek [5] described an approach for face and disguise detection using two co-registered SWIR cameras with a sensitivity range of 800 nm to 1400 nm and 1400 nm to 2200 nm, respectively. Their imaging system is described in Section 3.2.1 on page 30. The described detection method fuses the images using weighted differences and applies a threshold to distinguish skin from other materials.

Zhang *et al*. [81] presented a scanning sensor based on LEDs in the two wavebands of about 850 nm and 1450 nm in combination with a respective photodiode as detector. It serves as addition to a common RGB color camera for face recognition and introduces a spoof detection mechanism by distinguishing real skin from disguises at one single point in front of the camera. Due to this working principle, the system is susceptible to attacks using specifically prepared (partial) spoofs.

Wang *et al*. [105] described a multispectral method using 420*nm* and 800*nm* wavebands that divides the image of a face into blocks and creates feature vectors for each block that are compared to those acquired during enrollment, which has to be done using the same system. Again, the use of a block-wise approach might lead to smaller forged facial features staying undetected.

A patent of Zhang *et al*. [106] describes an anti-spoofing method that can be implemented into mobile devices and is based on the fusion of VIS and NIR images. The decision if a face is authentic is made by applying a threshold on the number of pixels within the facial area that have been classified as skin on the basis of normalized reflectance differences. Unfortunately, a practical implementation and evaluation of this approach has not been presented so far.

Besides the described individual shortcomings of these approaches, a robust solution on matching a reliable spoofing detection to specific facial features has not been introduced by any of them as well. In this dissertation, two different methods are presented for this matching procedure; see Chapter 5. As shown in Chapter 7, both methods achieve very good results.

3.3 Reference Design and Methodology

Considering the design goals formulated in Section 3.1 and the state of the art described in Section 3.2, a reference design for a skin detecting camera system with a focus on face recognition applications is proposed in this section. In Figure 3.2, the building blocks of the reference design and their relationships are shown. Besides the camera system hardware, the reference design provides further building blocks for image processing and image analysis.

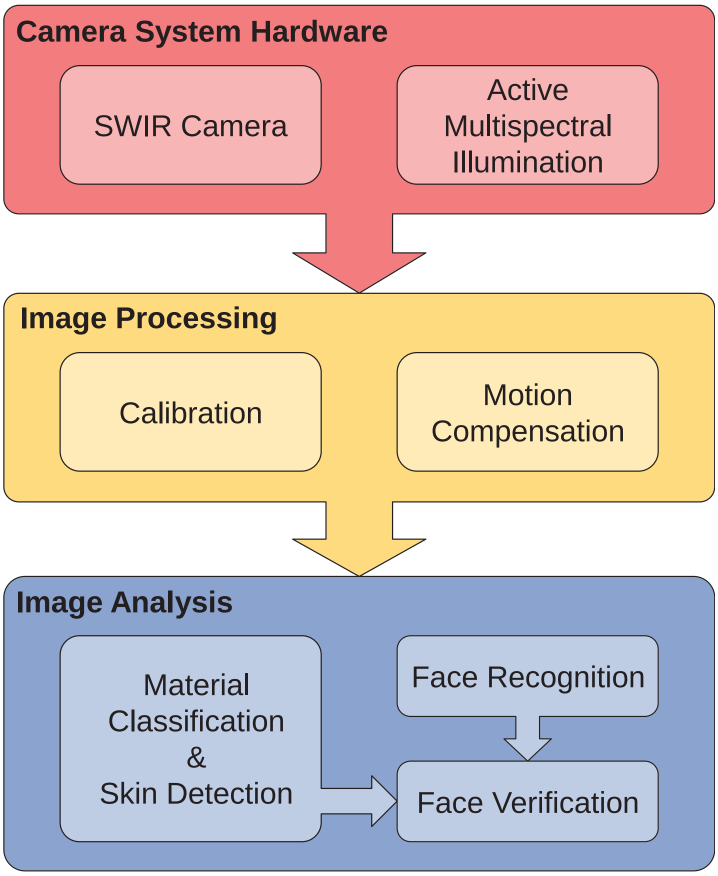


Figure 3.2: Building blocks of the proposed reference design.

Due to the advantages of the SWIR spectral range compared to the VIS spectrum for skin detection and face verification, the reference design is based on a multispectral SWIR camera system. As SWIR cameras are still very expensive, the design contains only one single camera. This avoids the need to align the optical path of multiple cameras or to apply complex image registration methods. Different wavebands are captured (time-) sequentially using field-sequential waveband capturing (FSWC). Instead of passive band pass filters, the design relies on active pulsed LED illumination in different narrow wavebands. This has several advantages:

- Active, frontal illumination helps to avoid shadows and ensures reliable face recognition.
- Influences of ambient light can be widely eliminated.
- No mechanical or moving parts are necessary.
- Wavebands can be switched without noticeable delay to keep acquisition times as short as possible.

Image processing consists of calibration, which comprises the typical intrinsic camera calibration and the calibration of the illumination module, as well as motion compensation. Motion compensation is necessary to remove artifacts that occur at the boundaries and edge details of moving objects when they are captured using FSWC imaging systems. Chapter 4 presents and evaluates different approaches to address this problem.

Material classification and skin detection are the first and most important components of the image analysis block. In order to address face anti-spoofing applications, the image analysis is complemented by face recognition and face verification modules. For these applications, it is preferable to enhance existing VIS-based systems using new modalities rather than building new multimodal SWIR systems from scratch. This way, already existing face image databases can still be used for face verification. Different researchers, *e.g.*, Bourlai *et al.* [4] or Klare and Jain [73], have achieved high verification rates when using SWIR images to verify faces that have been *enrolled* using VIS images with both commercial and scientific state-of-the-art face recognition software; see Section 2.7.3. Due to these promising results, this reference design proposes to use existing face recognition systems. Chapter 5 describes a novel approach for reliable skin detection based on spectral signatures and a method to combine skin detection with face recognition for anti-spoofing.

A specific system setup and practical implementation of this reference design is described in Chapter 6 and evaluated in Chapter 7.

Chapter 4

Motion Compensation for Field-Sequential Imaging Systems

This chapter addresses the problem of *motion artifacts*, which occur if moving objects are captured using an imaging system based on field-sequential waveband capturing (FSWC). It presents approaches to motion compensation for any kind of FSWC-based imaging with a focus on both real-time capability and accurate compensation of complex motion scenarios.

Publications: A first approach on motion compensation for an active multispectral SWIR camera system using interpolation between two successive multispectral image cubes has been presented in [21]. More generic approaches to motion compensation suitable for any field-sequential imaging systems are discussed in [24] (in preparation).

4.1 Introduction

As described in Section 3.2.2, all field sequential waveband capturing (FSWC) methods suffer from motion artifacts if they are used to capture dynamic scenes involving camera motion and/or moving objects [31]. In color imaging, these artifacts are denoted as *color breakup effect* [35]. Figure 3.1 on page 34 illustrates this effect for a simple example sequence. In practice, these motion artifacts limit the applicability of common FSWC-based multispectral imaging systems. To solve this problem, this work proposes a frame interpolation method based on motion estimation and compensation techniques to properly align all edges in every channel image of the multispectral image cube. For video sequences that have been acquired using simultaneous waveband capturing, optical flow methods have proven to be a very effective, but computationally expensive approach for frame interpolation [107]: sufficiently high performance for real-time applications can currently only be achieved by implementations using graphics processing units (GPUs).

A number of very successful and efficient motion estimation techniques have been proposed in recent years. An overview of these techniques is given in Section 2.5. However, it was found that even approaches addressing illumination variations cannot handle FSWC imagery properly; see Section 7.2. To the best of the author's knowledge, until now there has not been any research on motion compensation approaches that are able to handle the specific requirements of FSWC imagery. The major challenge for such an approach is the assumption of consistent intensities of corresponding pixels in subsequent images that is made by most motion detection techniques [36]. This assumption is in general not fulfilled by waveband-sequential multispectral imagery. Furthermore, there are no appropriate databases for comparing motion estimation and compensation techniques on FSWC imagery.

In Section 4.2, two fundamental concepts to FSWC motion compensation are introduced, *i.e.* inter-frame interpolation (see Section 4.2.1) and inter-channel matching (see Section 4.2.2). As inter-channel matching requires a solution to the problem of varying pixel intensities in the different spectral channels, Section 4.3 describes different approaches to address this problem. In Section 4.4, a smoothness constraint to enhance block matching algorithms is introduced, while Section 4.5 describes an adaptation of the artifact reduction from [91] for multispectral FSWC image sequences.

4.2 General Concepts to FSWC Motion Compensation

Consider a full multispectral image cube M_i, with $i \in \mathbb{N}$ being a sequential number, consisting of n channels $C_{i,w}$, which were acquired using FSWC and at sequential times $t_{i,w}$, $w = 0, \ldots, n-1$, with w being the waveband. Furthermore, a discrete and equidistant acquisition time $\Delta t = t_{i,w} - t_{i,w-1}$ is assumed for each channel and a constant acquisition time $T = t_{i,0} - t_{i-1,0} = n\Delta t$ for the full image cube, as illustrated in Figure 4.1.

When optical flow is calculated directly between adjacent waveband images of the image sequence, *i.e.* between $C_{i,w}$ and $C_{i,w+1}$, purely intensity-based optical flow algorithms will produce invalid displacement vectors due to the violation of the intensity consistency assumption. However, if a preceding multispectral image cube M_{i-1} is taken into account and optical flow is calculated for corresponding waveband

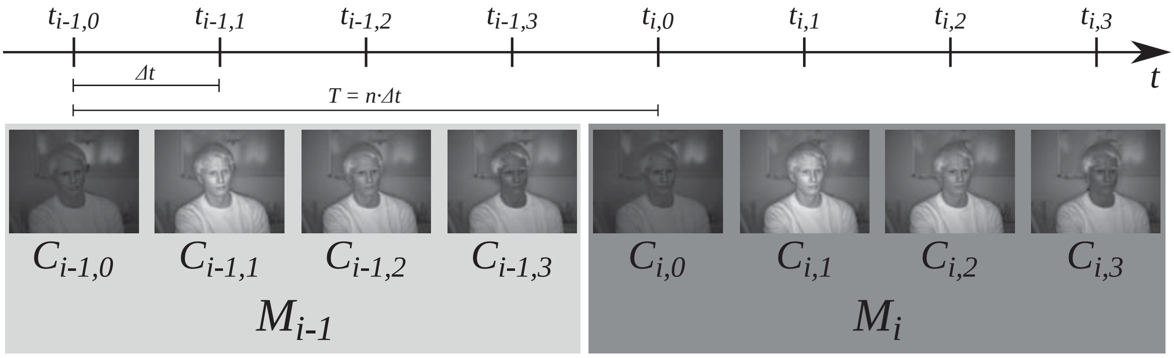


Figure 4.1: Waveband-sequential image sequence of two successive multispectral image cubes M_{i-1} and M_i with $n = 4$ channel, with the first channel being the "dark" reference of an active system. Adapted from [21].

images, *i.e.* between $C_{i-1,w}$ and $C_{i,w}$, $w = 1, \ldots, n-1$, the results are much better. This is demonstrated in Figure 4.2 on the next page.

In conclusion, motion estimation cannot be applied on FSWC image sequences directly, as illumination conditions and intensity values of object surfaces might differ strongly between the waveband images. Especially for active camera systems, the first step in image merging, the subtraction of the (not actively illuminated) "dark" reference image, might cause problems: properly exposed image areas with much detail in the actively illuminated waveband images might be completely dark and without detail in the reference image, as shown in the example in Figure 4.1.

One approach to motion compensation for FSWC image sequences is to avoid violating the intensity constancy assumption by using two consecutive multispectral image cubes and estimating motion only between pairs of corresponding channels, as shown in Figure 4.3 on page 45. Despite the larger displacement between the compared images, state-of-the-art motion estimation techniques will most likely produce accurate displacement vectors based on this method, as shown in the lower row of Figure 4.2 on the following page. Assuming a constant and linear motion between corresponding channels $C_{i-1,w}$ and $C_{i,w}$, every vector $F_{(i-1,w)\to(i,w)}(x,y)$ in the displacement map describing the movement of pixel (x,y) between $C_{i-1,w}$ and its successor $C_{i,w}$ can be regarded as a linear combination of n identical partial vectors describing a pixels movement between $C_{i,w-1}$ and $C_{i,w}$,

$$F_{(i,w-1)\to(i,w)}(x,y) \equiv \frac{1}{n} F_{(i-1,w)\to(i,w)}(x,y). \tag{4.1}$$

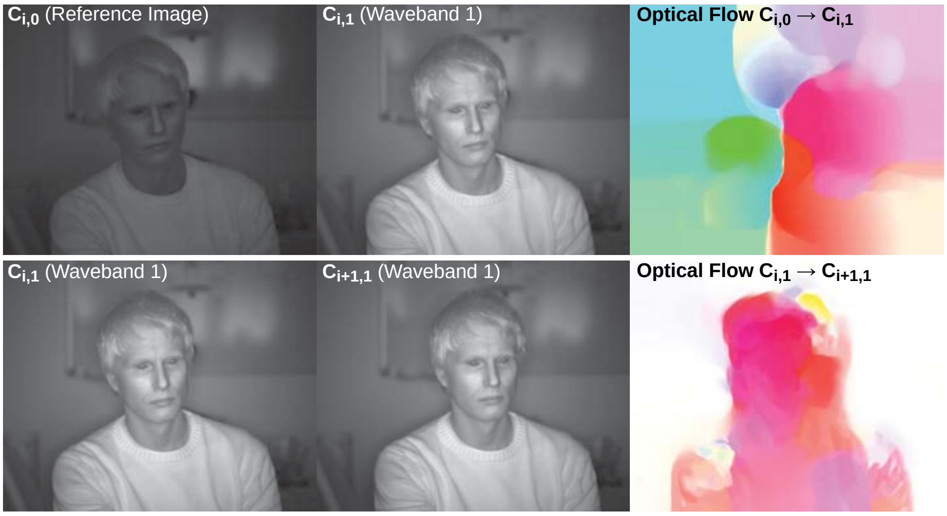

$C_{i,0}$ (Reference Image)
$C_{i,1}$ (Waveband 1)
Optical Flow $C_{i,0} \rightarrow C_{i,1}$
$C_{i,1}$ (Waveband 1)
$C_{i+1,1}$ (Waveband 1)
Optical Flow $C_{i,1} \rightarrow C_{i+1,1}$

Figure 4.2: Effect of violations of the intensity consistency assumption on optical flow calculation. Optical flow has been calculated between $C_{i,1}$ and its reference $C_{i,0}$ (upper row), and between $C_{i,1}$ and its successor $C_{i+1,1}$ (lower row) using TV-L1 OF. Colored pixels represent detected motion. Previously presented in [21].

This sort of interpolation between two multispectral image cubes will be called *inter-frame interpolation (IFI)*. In Section 4.2.1, this approach is described in detail. Furthermore, modifications to this approach that significantly improve the processing time by reducing the number of required optical flow calculations while preserving the compensation accuracy as much as possible are discussed.

As the IFI method assumes constant motion during the acquisition time of both cubes, changes in motion direction and speed will lead to incorrect displacement vectors. Therefore, Section 4.2.2 presents a method for the calculation of displacement vectors from each spectral channel to its closest reference channel, which might be a dedicated reference or simply the first spectral channel of each frame. This method will be denoted as *inter-channel matching (ICM)*. It can be applied in two variants:

1. Motion flow can be estimated directly from each individual channel to the reference. Here, the temporal offset between the compared images and, thus, the displacement of moving objects will be higher for higher waveband numbers w. This variant is denoted as channel to reference (C2R) matching and illustrated in Figure 4.3 on the next page (b).

2. Motion flow can be estimated between each pair of adjacent waveband channels. Here, the temporal offset is always the same and minimal. Thus, the displacement of moving objects will be minimal as well. In order to compensate motion for higher wavebands, the calculated displacement vectors between each pair

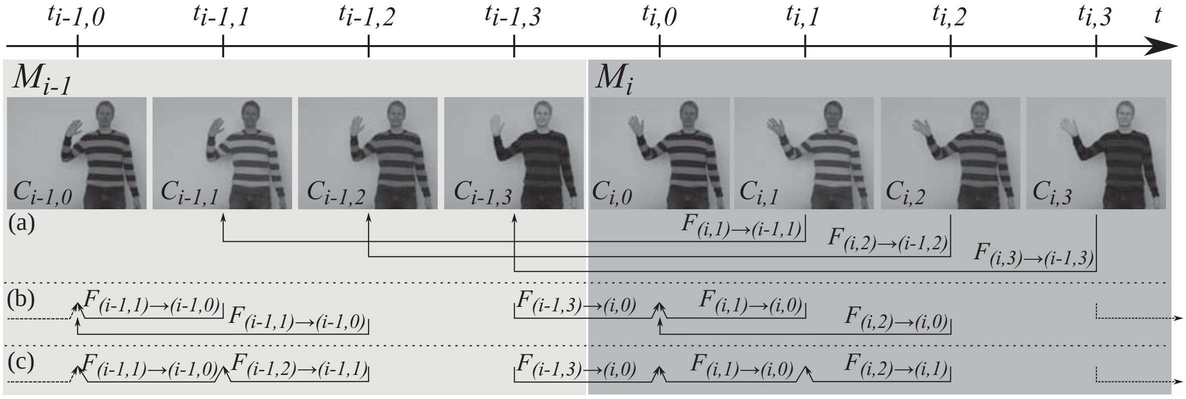

$t_{i-1,0}$
$t_{i-1,1}$
$t_{i-1,2}$
$t_{i-1,3}$
$t_{i,0}$
$t_{i,1}$
$t_{i,2}$
$t_{i,3}$
t
M_{i-1}
M_i
$C_{i-1,0}$
$C_{i-1,1}$
$C_{i-1,2}$
$C_{i-1,3}$
$C_{i,0}$
$C_{i,1}$
$C_{i,2}$
$C_{i,3}$
(a)
$F_{(i,1)\rightarrow(i-1,1)}$
$F_{(i,2)\rightarrow(i-1,2)}$
$F_{(i,3)\rightarrow(i-1,3)}$
(b)
$F_{(i-1,1)\rightarrow(i-1,0)}$
$F_{(i-1,1)\rightarrow(i-1,0)}$
$F_{(i-1,3)\rightarrow(i,0)}$
$F_{(i,1)\rightarrow(i,0)}$
$F_{(i,2)\rightarrow(i,0)}$
(c)
$F_{(i-1,1)\rightarrow(i-1,0)}$
$F_{(i-1,2)\rightarrow(i-1,1)}$
$F_{(i-1,3)\rightarrow(i,0)}$
$F_{(i,1)\rightarrow(i,0)}$
$F_{(i,2)\rightarrow(i,1)}$

Figure 4.3: Flow calculation for two successive multispectral image cubes using either (a) inter-frame interpolation (IFI), (b) inter-channel matching (ICM) from each channel to the reference (C2R), or (c) ICM from channel to channel (C2C).

of channels has to be linearly combined until the reference channel is reached. This variant is denoted as channel to channel (C2C) matching and illustrated in Figure 4.3 (c).

Finally, in Section 4.2.3, a combination of IFI and ICM is proposed.

4.2.1 Inter-Frame Interpolation (IFI)

The IFI approach requires two full consecutive multispectral image cubes M_{i-1} and M_i consisting of n spectral channels $C_{i,w}$, acquired at times $t_{i,w}$, with $w = 0,\dots,n-1$ being the waveband. Without violating the intensity consistency assumption for motion estimation, forwards and backwards displacement vectors can be calculated for each pair of channels ($C_{i-1,w}$, $C_{i,w}$), $w > 0$ using state-of-the-art optical flow or block matching methods:

$$\begin{aligned} \text{forward flow: } & F_{(i-1,w)\rightarrow(i,w)}, w = 1,\dots,n-1, \\ \text{backward flow: } & F_{(i,w)\rightarrow(i-1,w)}, w = 1,\dots,n-1. \end{aligned}$$

Here, the reference channel C_0 is used as the target, therefore it does not need to be compensated. As each optical flow calculation is computationally very expensive, different interpolation methods can be implemented with a focus on either compensation accuracy or processing time. These methods will be described in the following.

Bidirectional, All Channel Optical Flow (IFI-B)

This method calculates both forwards and backwards optical flow fields for each pair of channels $\langle C_{i-1,w},\ C_{i,w}\rangle$ with $w > 0$, as defined above. Then, assuming a constant linear motion between corresponding channels $C_{i-1,w}$ and $C_{i,w}$, both forward and backward flow are applied in order to interpolate a motion corrected spectral image $\tilde{C}_{i,w}, w = 1, \ldots, n-1$, for the reference time $t_{i,0}$ bidirectionally:

$$\tilde{C}_{i,w} = \tfrac{(n-w)}{n} F_{(i-1,w)\to(i,w)}[C_{i-1,w}] \oplus \tfrac{w}{n} F_{(i,w)\to(i-1,w)}[C_{i,w}], \tag{4.2}$$

where $F_{(j,w)\to(k,w)}[C_{j,w}]$ indicates the application of the displacement vector field $F_{(j,w)\to(k,w)}$ to channel image $C_{j,w}$, resulting in all pixel values $p(x,y)$ from $C_{j,w}$ being shifted according to the two-dimensional displacement vectors $\vec{d}(x,y)$ from $F_{(j,w)\to(k,w)}$. In the final "corrected" spectral images $\tilde{C}_{i,w}$, the positions of moving objects will match those in the reference channel $C_{i,0}$, if the motion estimation has been accurate.

The bidirectional interpolation function $\oplus$ calculates the intensity of every pixel in $\tilde{C}_{i,w}$ by averaging the corresponding pixel values in $C_{i-1,w}$ and $C_{i,w}$. In conjunction with the detection of occlusions, this function provides high interpolation accuracy. The main disadvantage of this approach is its extremely high computational complexity, as it requires $2 * (n-1)$ optical flow calculations for each multispectral image cube.

In the following, approaches to reduce the computational complexity of the bidirectional, all channel IFI method (IFI-B) are presented. However, reducing the number of optical flow computations also reduces the compensation accuracy; see Section 7.2.

Unidirectional, All Channel Optical Flow (IFI-U)

This method simplifies the interpolation by using only one optical flow calculation for each pair of channels $C_{i-1,w}$ and $C_{i,w}$, $w = 1, \ldots, n-1$, leading to $(n-1)$ optical flow calculations for the full image cube. It calculates either the forwards or backwards flow depending on the current waveband, to keep the length of the resulting displacement vectors and, thus, the expected error as small as possible:

$$\tilde{C}_{i,w} = \begin{cases} \frac{w}{n} \cdot F_{(i,w)\to(i-1,w)}[C_{i,w}] & \text{if } w \leq \frac{n}{2} \\ \\ \frac{(n-w)}{n} \cdot F_{(i-1,w)\to(i,w)}[C_{i-1,w}] & \text{if } w > \frac{n}{2} \end{cases} \tag{4.3}$$

Unidirectional, Partial Channel Optical Flow (IFI-1 / IFI-2)

The number of required optical flow calculations can be further decreased by interpolating a given flow field to subsequent channels. Assume the backwards flow $F_{(i,u)\rightarrow(i-1,u)}$ for waveband u is known and the motion is constant during both cubes M_{i-1} and M_i. Then, the backward flow $F_{(i,v)\rightarrow(i-1,v)}$ for channel $v > u$ can be interpolated by scaling $F_{(i,u)\rightarrow(i-1,u)}$ appropriately and applying the result on the original flow field:

$$F_{(i,v)\rightarrow(i-1,v)} = \left(-\tfrac{v-u}{n} \cdot F_{(i,u)\rightarrow(i-1,u)}\right)[F_{(i,u)\rightarrow(i-1,u)}]. \tag{4.4}$$

The motion compensated image $\tilde{C}_{i,v}$ is calculated similar to Equation (4.3) on the preceding page, but only in backwards direction:

$$\tilde{C}_{i,w} = \frac{w}{n} \cdot F_{(i,w)\rightarrow(i-1,w)}[C_{i,w}]. \tag{4.5}$$

This way, the number of required optical flow calculations for each frame can be reduced down to one (IFI-1). However, the more flow fields are interpolated from a previous one, the higher the approximation error will be if the assumption of constant motion does not hold true. Neglecting this fact, one calculated flow field could even be extrapolated over several image cubes, further reducing the processing time at the cost of an even higher approximation error. In practice and depending on the amount and nature of expected motion in the scene, it might be a better choice to interpolate only a limited number of flow fields from others.

To achieve a more accurate interpolation, a second known flow field $F_{(i,w)\rightarrow(i-1,w)}$ of a subsequent channel w can be used to interpolate $F_{(i,v)\rightarrow(i-1,v)}, u < v < w$ bidirectionally, thus requiring two optical flow calculations for each frame (IFI-2):

$$\begin{aligned} F_{(i,v)\rightarrow(i-1,v)} = & -\tfrac{v-u}{n} \cdot F_{(i,u)\rightarrow(i-1,u)}[F_{(i,u)\rightarrow(i-1,u)}] \\ & \oplus \tfrac{w-v}{n} \cdot F_{(i,w)\rightarrow(i-1,w)}[F_{(i,w)\rightarrow(i-1,w)}] \end{aligned} \tag{4.6}$$

4.2.2 Inter-Channel Matching (ICM)

If the problem of varying pixel intensities between the different spectral channels of FSWC imagery can be overcome, ICM is a potentially more accurate alternative to IFI, as it allows to compensate dynamic changes of motion speed and direction during the acquisition of a multispectral image cube. It can be applied in two different ways:

Channel to Reference (C2R) Matching

With the C2R method, each channel image $C_{i,w}$ with $w > 0$ is matched to its respective reference channel $C_{i,0}$ directly to calculate the backward optical flow $F_{(i,w)\rightarrow(i,0)}$, which is used to compensate motion in $C_{i,w}$:

$$\tilde{C}_{i,w} = F_{(i,w)\rightarrow(i,0)}[C_{i,w}]. \tag{4.7}$$

This method can be applied to a single multispectral image cube M_i. However, it is obvious that the compensation accuracy can be increased by using a second preceding multispectral frame M_{i-1} and matching channels of higher wavebands $C_{i-1,w}$, with $w > n/2$, to the subsequent reference channel $C_{i,0}$, as object displacement will (in general) be smaller between these two images. This approach is illustrated in Figure 4.3 (b) and leads to the following updated equation:

$$\tilde{C}_{i,w} = \begin{cases} F_{(i,w)\rightarrow(i,0)}[C_{i,w}] & \text{if } w \leq \frac{n}{2} \\ F_{(i-1,w)\rightarrow(i,0)}[C_{i-1,w}] & \text{if } w > \frac{n}{2}. \end{cases} \tag{4.8}$$

Channel to Channel (C2C) Matching

In contrast to C2R, C2C matches each channel image $C_{i,w}$ with $w > 0$ to it's direct predecessor $C_{i,w-1}$ and calculates the backwards optical flow $F_{(i,w)\rightarrow(i,w-1)}$. To compensate motion in $C_{i,w}$, all partial flow fields $F_{(i,w)\rightarrow(i,w-1)}$ are applied to $C_{i,w}$ sequentially:

$$\tilde{C}_{i,w} = F_{(i,1)\rightarrow(i,0)}[\ldots[F_{(i,w)\rightarrow(i,w-1)}[C_{i,w}]]]. \tag{4.9}$$

Similar to the C2R method, C2C can also be further improved by matching subsequent channels $C_{i,w}$ and $C_{i,w-1}$ stepwise either forwards or backwards until the closest reference channel is reached; see Figure 4.3 on page 45 (c). The compensated image $\tilde{C}_{i,w}$ can then be found by sequentially applying the resulting flow vectors either forwards or backwards:

$$\tilde{C}_{i,w} = \begin{cases} F_{(i,1)\rightarrow(i,0)}[\ldots[F_{(i,w)\rightarrow(i,w-1)}[C_{i,w}]]] & \text{if } w \leq \frac{n}{2} \\ F_{(i-1,n-1)\rightarrow(i,0)}[\ldots[F_{(i-1,w)\rightarrow(i-1,w+1)}[C_{i-1,w}]]] & \text{if } w > \frac{n}{2} \end{cases} \tag{4.10}$$

The C2C method keeps the displacement of moving objects as small as possible for each optical flow calculation. However, for multispectral image cubes with less than four wavebands the optimized C2R and C2C methods are identical.

4.2.3 Combining ICM and IFI

The inter-frame interpolation and inter-channel matching methods can also be used in combination. In the first step, an IFI method is applied to calculate the interpolated displacement vector field

$$\tilde{F}_{(i,w)\rightarrow(i,0)} = \frac{w}{n} \cdot F_{(i,w)\rightarrow(i-1,w)}. \tag{4.11}$$

Then, $\tilde{F}$ is used as initial flow for an ICM method, which tries to optimize the displacement vectors to account for inconstant motion during the acquisition of both cubes. For this purpose, a cost factor is introduced which punishes strong deviations from the initial flow vectors.

4.3 Handling Inconsistent Intensities

There are three principle ways to address the intensity inconsistency problem: reducing the intensity differences by applying some kind of normalization (Section 4.3.1), intensity transformation (Section 4.3.2), or by correlation between channels. (Section 4.3.3). Figure 4.4 on page 52 and Figure 4.5 on page 53 illustrate their effect on the channels of an exemplary multispectral image frame.

4.3.1 Intensity Normalization

If pixel intensities are used as data term for the calculation of motion flow, varying intensities between the spectral channels have to be reduced in order for inter-channel matching methods to achieve better results than inter-frame interpolation. The following approaches are evaluated in this work:

Global Linear normalization

An average illumination and optimized image contrast is achieved by mapping the original intensity values $f(x,y)$, having a range of $[f_{min}, f_{max}]$, to new values $g(x,y)$ in the range $[0, g_{max}]$, with $0 \leq f_{min}$ and $f_{max} \leq g_{max}$ [32]; see Figure 4.4b on page 52:

$$g(x,y) = \frac{f(x,y) - f_{min}}{f_{max} - f_{min}} g_{max} \tag{4.12}$$

Local Linear Normalization

In order to compensate for non-uniform illumination within the images, *local normalization* [108] can be applied; see Figure 4.4c on page 52. This approach estimates the local mean value $m_f(x,y)$ and local variance $\sigma_f(x,y)$ within a window around each pixel $f(x,y)$ by using Gaussian filters and computes the normalized intensity value $g(x,y)$ as follows:

$$g(x,y) = \frac{f(x,y) - m_f(x,y)}{\sigma_f(x,y)} \tag{4.13}$$

Histogram equalization

To achieve a better compensation for varying brightness of different surfaces in the scene while emphasizing details, normalization can be extended by a nonlinear operation called *histogram equalization*, which uniformly distributes the intensity values over the available range [32]; see Figure 4.4d on page 52. It normalizes the histogram $H(i)$ of an input image and calculates the cumulative distribution $H'(i)$, which is used to remap the intensity values:

$$H'(i) = \sum_{0 \le j \le i} H(j) \tag{4.14}$$

$$g(x,y) = H'(f(x,y)) \tag{4.15}$$

Contrast Limited Adaptive Histogram Equalization (CLAHE)

Histogram equalization can also be performed in a local (moving) window individually for each pixel. As this operation tends to amplify noise in homogeneous areas, an enhanced algorithm called contrast limited adaptive histogram equalization (CLAHE) has been proposed that introduces a clipping limit for histogram redistribution to avoid this issue [109]; see Figure 4.4e on page 52.

4.3.2 Intensity Transformation

Reducing the intensity inconsistency between non-corresponding channels can also be achieved by converting intensity values into another domain and using the conversion result as data term for motion estimation. The following approaches are evaluated, in this respect, in Section 7.2.

Census Transform

The census transform describes the local spatial structure around a specific pixel of an image. It has been proposed by Zabih and Woodfill [110] as an approach to the correspondence problem for stereo depth and optical flow calculation and calculates a vector of binary values for each pixel $p_{x,y}$ depending on its local neighborhood: if a neighboring pixel has a lower intensity than p, a 1 will be appended to the vector, otherwise a 0. After the transformation, correspondence is calculated by finding the minimum Hamming distance between two vectors. Due to the nature of the census transform, a graphical representation does not make sense here.

Gradients

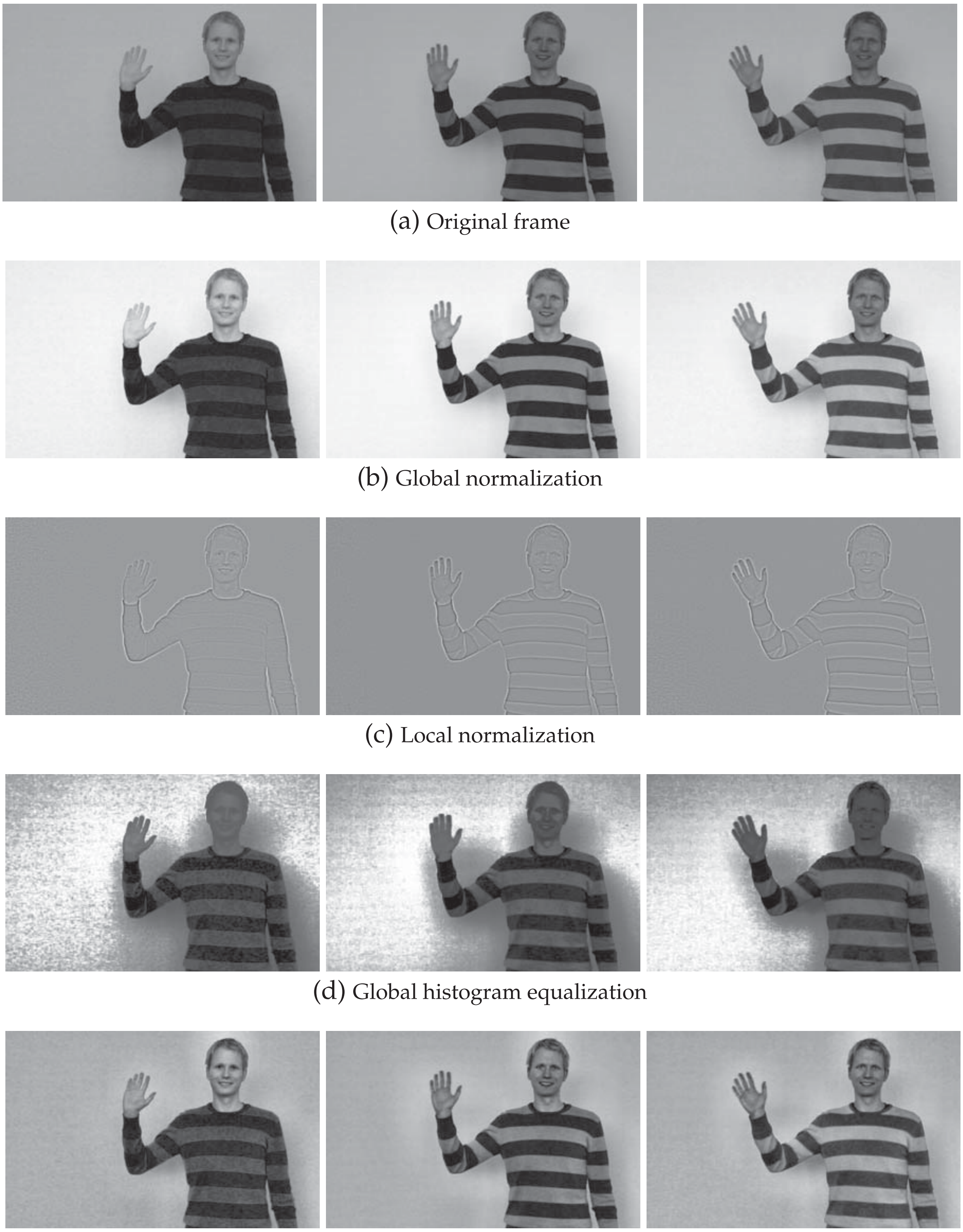
Gradient images can be created by differentiating the original image, *e.g.*, using a Sobel filter in both x and y direction. In gradient images, a pixel's intensity value depends on the change of brightness in it's local neighborhood: edges will appear bright, while constantly colored areas will appear black [46]. If all spectral channels show similar edges and contours, their resulting gradient images will be similar as well, independent of their absolute intensity values; see Figure 4.5a on page 53.

4.3.3 Correlation Based Methods

Instead of normalizing or transforming the channel images, correlation-based approaches can be used to estimate flow fields as well. The following approaches have been investigated in the context of this work:

Cross-Spectral Feature Detection

Feature detection methods are frequently used in multispectral or multimodal image registration [111], for example, in the field of remote sensing. This process involves scaling, shifting or rotating an image to find the best match to another image. While these methods can not be used to estimate dense motion fields between two images directly, they might be used for the registration of blocks in block matching algorithms. To find a match, a sufficient number of distinctive features that are robust to the differences between the spectral channels must be present within each block.



(a) Original frame

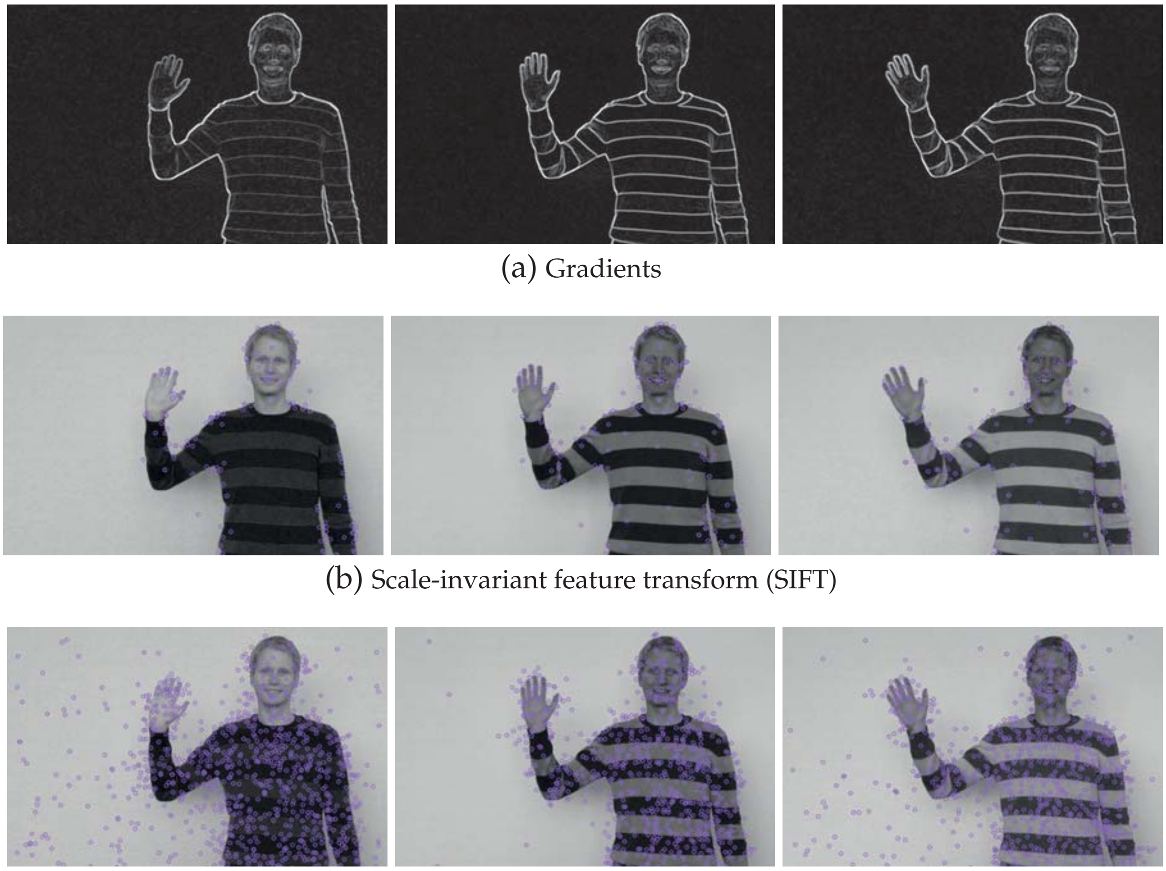
(b) Global normalization

(c) Local normalization

(d) Global histogram equalization

(e) CLAHE

Figure 4.4: Examples of approaches to handle inconsistent intensities between different waveband images by different kinds of normalization.



(a) Gradients

(b) Scale-invariant feature transform (SIFT)

(c) Speeded up robust features (SURF)

Figure 4.5: Examples of approaches to handle inconsistent intensities between different waveband images by intensity transformation and feature detection. Partially adapted from [24].

To investigate the potential of cross-spectral feature detection for FSWC motion compensation, the *Scale Invariant Feature Transform* (SIFT), *Speeded Up Robust Features* (SURF), as well as *Maximally Stable Extremal Region Extractor* (MSER) and *Binary Robust Invariant Scalable Keypoints* (BRISK) have been explored based on implementations from the openCV library[1]. Results for SIFT and SURF are shown in Figure 4.5b and Figure 4.5c. Unfortunately, in almost all of our test examples these approaches could not robustly detect a sufficient number of distinctive features, or the detected features varied strongly between the channels. As this leads to a very low motion field accuracy, this approach was excluded from the evaluation in Section 7.2.

[1] `http://opencv.org/`

Mutual Information (MI)

Mutual information is based on the entropy of an image pair and yields a high value if the information gain of a new image in addition to an existing image is low, *i.e.*, if two images of the same scene are geometrically aligned. It is known to be robust against non-linear intensity relationships and has been proposed for both multispectral and multimodal image registration applications [111, 112]. It can be used as a cost function for block matching, but so far it cannot be linearized for the use in optical flow algorithms.

A block of pixels $B_M(x_1, y_1)$ from image M is matched to a block $B_N(x_2, y_2)$ within image N by finding the spatial transformation vector $\vec{v} = (\Delta x, \Delta y)$ that maximizes the mutual information between the two blocks $\mathrm{MI}(a, b)$. The resulting vector $\vec{v}$ denotes the optical flow of pixel (x,y), $F_{M\to N}(x, y)$:

$$F_{M\to N}(x, y) = \underset{\vec{v}=(\Delta x, \Delta y)}{\arg\max}\ \mathrm{MI}(B_M(x, y), B_N(x + \Delta x, y + \Delta y))$$

Unfortunately, mutual information is computationally far too intensive to be applied in real-time. Calculating MI between two macro blocks with a block size of 15x15 pixels takes about 0.2 ms on the computer used for evaluation; see Section 7.2 on page 100. For an image of 640x480 pixels and a (small) search window of 11x11 pixels (i.e. $p = 10$), calculation of all displacement vectors takes $t = 0.2ms \cdot 11 \cdot 11 \cdot 640 \cdot 480 = 7434.24\,\mathrm{s}$, which is very far from real time performance. Parallel computation using a GPU, however, is limited due to the amount of independent memory needed for the 2D histograms. With an approximate GPU-based calculation method, a speedup of about 25 times can be achieved [113], so the motion estimation for one image pair would still require $\approx 300\,\mathrm{s}$. Therefore, this approach was excluded from the evaluation as well.

Normalized Cross-Correlation (NCC)

In digital image processing, cross-correlation is commonly used as cost function in order to find the position of specific features in an image [46]. NCC additionally normalizes the image which improves the robustness against illumination changes. NCC can be used as an inverse cost function for block matching, as well as a linearized data term in optical flow calculation [54, 114].

4.4 Extended Cost Function (ECF) for Block Matching

Homogeneous image areas provide very sparse information for motion estimation, especially when using strongly normalized or gradient images. While state-of-the-art dense optical flow approaches assume motion to be smooth, block matching is likely to produce incorrect displacement vectors for these areas. To avoid this, an extended cost function for block matching based on the sum of absolute differences (SAD, see [56]) is proposed. The SAD calculates the costs of a block matching operation by summing up the absolute intensity differences of pixels within block size *bs* around the original pixel position (x, y) and the shifted position $(x + d_x, y + d_y)$. In addition, the ECF rewards smooth motion by adding the deviation between the displacement vector $\vec{d}(x, y)$ of pixel (x, y) and those of neighboring pixels $\vec{d}(x+k, y+l)$ within a close neighborhood of size *ns* to the costs, weighted by a factor k_n:

$$\begin{aligned}\mathrm{SAD}_n(\vec{d}(x,y)) = \sum_{j=0}^{bs}\sum_{i=0}^{bs}\Big| & f(x+i,\ y+j)\\ & - g(x+i+d_x(x,y),\ y+j+d_y(x,y))\Big|\\ & + k_n\cdot\sum_{l=0}^{ns}\sum_{k=0}^{ns}\Big(|d_x(x+k,y+l)-d_x(x,y)|+|d_y(x+k,y+l)-d_y(x,y)|\Big)\,,\end{aligned} \tag{4.16}$$

where $d_x(x, y)$ and $d_y(x, y)$ are the components of vector $\vec{d}(x, y)$. Strong features will still be matched correctly, as the neighborhood relations have only minor influence in comparison, while weakly featured regions are stabilized by surrounding features.

4.5 Pixelwise Artifact Correction (PAC)

Based on the approach by Schmidt and Jähne [91], which has been described in Section 3.2.2 on page 34, a computationally very efficient approach to motion compensation can be deduced on the basis of two main assumptions:

- Pixel intensities in the different channels change only slowly over time, whereas rapid changes in intensity are assumed to be based on motion events.
- A pixel will be influenced by (at most) one motion event during the acquisition time of two adjacent frames or image cubes M_{i-1} and M_i.

Thus, given a motion detection threshold δ, a motion event is detected between channels $C_{i,w}$ and $C_{i,w+1}$, if

$$\left|C_{i-1,w} - C_{i,w}\right| < \delta \ \wedge \ \left|C_{i-1,w+1} - C_{i,w+1}\right| > \delta. \tag{4.17}$$

If motion is detected between the last channel of frame M_{i-1} and the first channel of frame M_i, *i.e.*, $w = n-1$, no motion artifacts occur. For $w \neq n-1$, motion artifacts are corrected by simply replacing intensity values for the affected channels $C_{i,v}, w \leq v < n$, by the intensity values of the previous frame $C_{i-1,v}$.

4.6 Summary

This chapter presents and discusses enhancements on existing motion estimation approaches in order to allow their successful application to field-sequential waveband captured (FSWC) multispectral imagery of dynamic scenes. The major challenge for the application of existing methods is the assumption of consistent intensities for corresponding pixels made by most motion estimation approaches, which is in general not fulfilled for adjacent wavebands of FSWC imagery.

While inter-frame interpolation (IFI) methods estimate motion fields between corresponding channels of successive multispectral cubes to avoid intensity inconsistencies, inter-channel matching (ICM) estimates motion fields between neighboring channels within a multispectral cube, which requires a successful handling of inconsistent intensities between the channels but (potentially) benefits from interpolation for shorter time intervals and displacement vectors.

For IFI, an optimal bidirectional interpolation approach requires $2*(n-1)$ motion flow calculations for multispectral images with n wavebands. As motion flow calculations are computationally very expensive, different variants are presented which reduce the number of required motion flow calculations for each frame significantly while preserving compensation accuracy as much as possible. With respect to ICM, different techniques for the handling of intensity inconsistencies based on normalization, intensity transformation and correlation are introduced.

In Section 7.2 of Chapter 7, an in-depth evaluation of the described approaches with focus on both *compensation accuracy* and *real-time capability* is presented.

Chapter 5

Skin Detection and Face Verification

This chapter describes the approach to pixel-level skin detection based on spectral signature in the short-wavelength infrared (SWIR) spectral range and it's application for the detection of spoofing attack or the verification of authentic faces.

Publications: The concept of pixel-wise skin classification and a first evaluation of the classification performance has been published in [21]. The approaches to the combination of skin detection and state-of-the-art face recognition systems have been discussed in [22].

5.1 Introduction

The detection of skin in acquired multispectral short-wavelength infrared imagery is performed by a binary classification method that analyzes the spectral signature of each pixel and decides whether or not a specific pixel shows human skin. The design of this classifier is described in Section 5.2 on the next page.

In addition, two further steps are performed in the context of image analysis. First, a face detection and recognition algorithm searches for faces in the acquired images. Finally, the locations of detected faces are matched against the results of the skin classification in order to verify their authenticity. Section 5.3 presents fundamental approaches to combine the developed skin detection method with state-of-the-art face recognition systems, including already acquired face databases.

5.2 Skin Classification Based on Spectral Signatures

The skin classification method presented in this dissertation is an extension of the work by Schwaneberg [25]. His approach is based on the calculation of quotients between the absolute remission intensity values of the different wavebands in the spectral signatures of material surfaces and uses both simple thresholding operations and machine learning based classifiers. In this work, however, the use of normalized differences instead of quotients is proposed, as they are more robust to offset changes and have a well-defined range between −1 and +1, which makes processing easier.

Following the approach by Schwaneberg, the skin classification method proposed in this work also consists of both a thresholding and a machine learning based classifier. Here, the two algorithms are applied in a hierarchical manner in order to optimize both classification accuracy and runtime performance for real-time imaging applications. The first algorithm performs fast, but rather coarse-grained classification, while the second algorithm allows for more fine-grained classification. Both algorithms perform pixel-per-pixel classification using the spectral signatures $\vec{s}$ of the individual pixels:

$$\vec{s}(x,y) = [g_1, .., g_{n-1}],$$

with each element $g_w, 1 \leq w \leq n-1$ being the grayscale value of pixel (x,y) in spectral channel $C_{i,w}$ of the multispectral image cube M_i, which consists of n channels. As $w = 0$ is the reference channel, it is not contained in the spectral signature.

5.2.1 Thresholding on Multidimensional Normalized Differences

For each pixel (x,y), the thresholding algorithm calculates normalized differences $d(g_a, g_b)$ for all possible combinations of grayscale values g_w within $\vec{s}(x,y)$:

$$d(g_a, g_b) = \left(\frac{g_a - g_b}{g_a + g_b}\right)$$

with $1 \leq a < n-1$ and $a < b < n$. For $n = 5$, for example, this results in a vector of normalized differences $\vec{d}$ with

$$\vec{d}(x,y) = [d(g_1,g_2), d(g_1,g_3), d(g_1,g_4), d(g_2,g_3), d(g_2,g_4), d(g_3,g_4)]$$

for each pixel (x,y). The normalized differences range from $-1 \leq d(g_a, g_b) \leq +1$. In contrast to the values of the spectral signatures, they are independent of the absolute brightness of the analyzed pixel (x, y), which differs not only with the remission properties of the surface material, but also with the measurement distance and angle of incidence. Thus, the vector of normalized differences allows for a robust and fast classification of materials into “skin” and “non-skin” by specifying upper and lower thresholds for each vector component, leading to a multidimensional bounding box. Only instances that are mapped to a point within this box are classified as “skin”. Depending on the training method, this box can be designed either more accepting (*i.e.* by choosing thresholds that include all positive training samples with a large margin) or more rejecting (*i.e.* by using smaller margins).

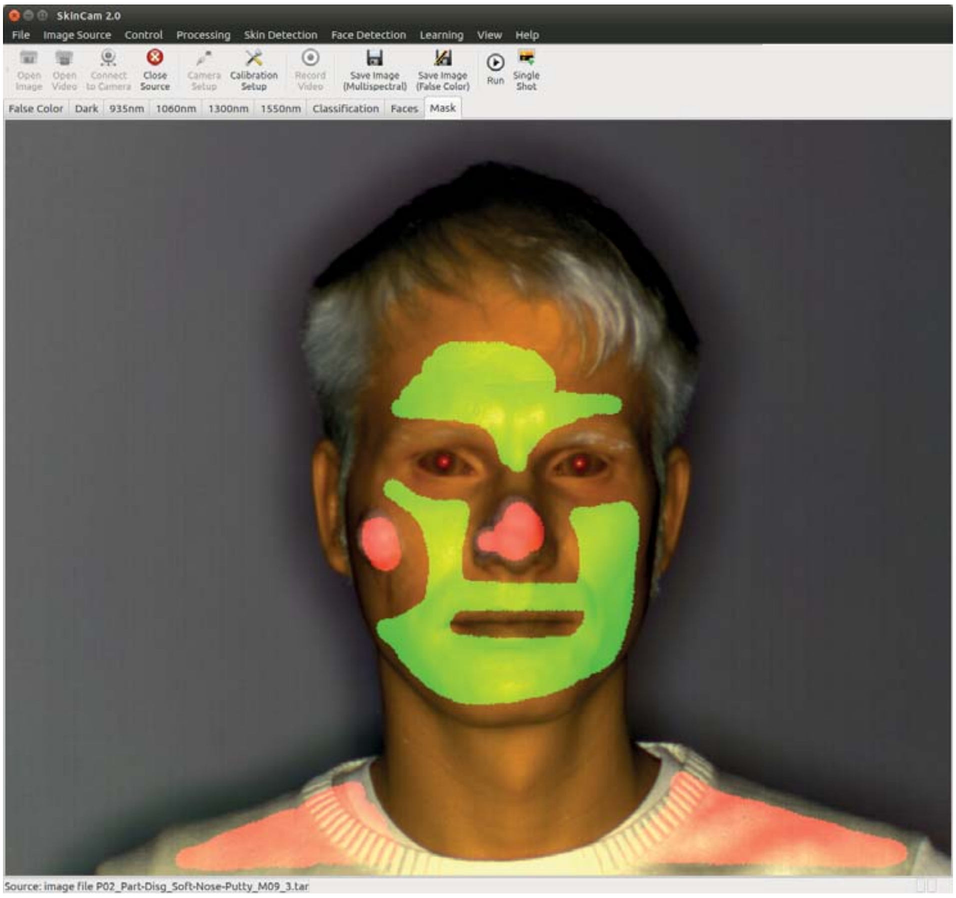
However, this “difference filter” algorithm is not capable of distinguishing actual human skin from materials that are very similar to skin, such as some kinds of silicon that are used for the creation of masks. Due to the simplicity of the mapping method, such samples might be mapped to points very close to authentic skin samples, making it impossible to achieve a linear separation of all “skin” and “non-skin” samples. Therefore, for biometric anti-spoofing applications that require a very accurate classification with a low false negative rate (FNR) and are likely to be attacked with such skin-like material, this difference filter alone is not sufficiently reliable.

To solve this problem, the difference filter is only used in conjunction with a second algorithm. Thus, it is designed to be very accepting by specifying the thresholds in a way that includes all real skin samples as well as skin-like materials. Then, a computationally more expensive fine-grained classification algorithm based on machine learning techniques is applied on this set of “skin or skin-like” spectral signatures.

5.2.2 Classification with Machine Learning Techniques

For the implementation of a fine-grained machine learning based classifier, three different techniques are evaluated in the context of this work: binary decision trees, random forests and support vector machines (SVMs). The basics of these techniques are described in Section 2.6 on page 20.

All classifiers are trained using normalized difference vectors $\vec{d}$, which are calculated based on spectral signatures of skin, skin-like materials and other materials, as described in Section 5.2.1 on the preceding page. In order to find a robust and accurate classifier, the training data must cover a sufficiently high amount of samples from all relevant material surfaces in varying distances and observation angles to avoid overfitting. Therefore, a large variety of multispectral SWIR images is required to extract the necessary training data from. In the context of this work, training images



Source: image file P02_Part-Disg_Soft-Nose-Putty_M09_3.tar

Figure 5.1: Generation of training data for machine-learning-based classifiers from a face with applied partial masks. Spectral signatures from areas highlighted in green are extracted as positive, signatures from areas highlighted in red as negative samples.

are acquired with the active camera system that is presented in Chapter 6. Using a specifically developed software tool, only unquestionable “skin” and “non-skin” areas in these images have been manually segmented to extract the annotated spectral signatures of the respective pixels; see Figure 5.1. Details about the classification performance will be given in Chapter 7. In general, all machine learning based classifiers perform significantly better than the difference filters alone, but have a much higher computational complexity.

Limiting the fine-grained classification to those samples that have been positively classified by the difference filter reduces the overall run time of the skin detection module noticeably in typical use cases. In addition, outliers and “unknown” material samples (samples that were not included in the training data) are less likely to create false positives when using two different classifiers. The result of the classification process is stored in the form of a binary image with

$$p(x,y) = \begin{cases} 1 \text{ if } \vec{s}(x,y) \in S_{\text{skin}} \\ 0 \text{ else,} \end{cases} \tag{5.1}$$

where S_{skin} is the set of spectral signatures $\vec{s}$ that are classified as skin.

5.3 Combining Skin Detection and Face Verification

Classical biometric face recognition is limited in spoof detection, as solely imagery in the visual (VIS) spectrum is used. Methods using alternative modalities are more successful in the authentication of skin and faces as such, but often require to set up new databases for face recognition. The approach presented in this work, therefore, aims at a scheme integrating multispectral SWIR skin authentication into existing face verification systems. Here, it is expected that the face recognition system's database has been created using images captured in the VIS spectrum. As it will be shown in Chapter 7, this approach achieves unprecedented anti-spoofing performance in cooperative user scenarios even in the presence of partial disguises or facial hair.

As described in Section 5.2, multispectral SWIR imaging allows for a reliable classification of material as "skin" or "non-skin" at pixel level. Even material similar to skin, such as silicon that has been specifically designed to model human limbs, can be distinguished from authentic human skin with high accuracy. However, using these classification results for face verification is still a challenging problem, as facial areas showing skin naturally vary strongly across different individuals, while at the same time, spoofs may also address very different regions and amounts of a person's face; see Figure 5.2 on the next page.

As it is not feasible to individually re-engineer any potentially given face recognition system in order to analyze its "facial regions of interest" and to apply skin verification there, this work proposes two fundamentally different methods to integrate SWIR-based skin detection into existing face recognition systems that are widely independent of the actual recognition algorithm:

(A) Masking Out Non-Skin Pixels: For this method, only SWIR images have to be acquired. Thus, it requires the given face recognition system to be able to handle SWIR images as input for face recognition. Here, skin classification is applied on the SWIR images and non-skin regions are masked out prior to face recognition

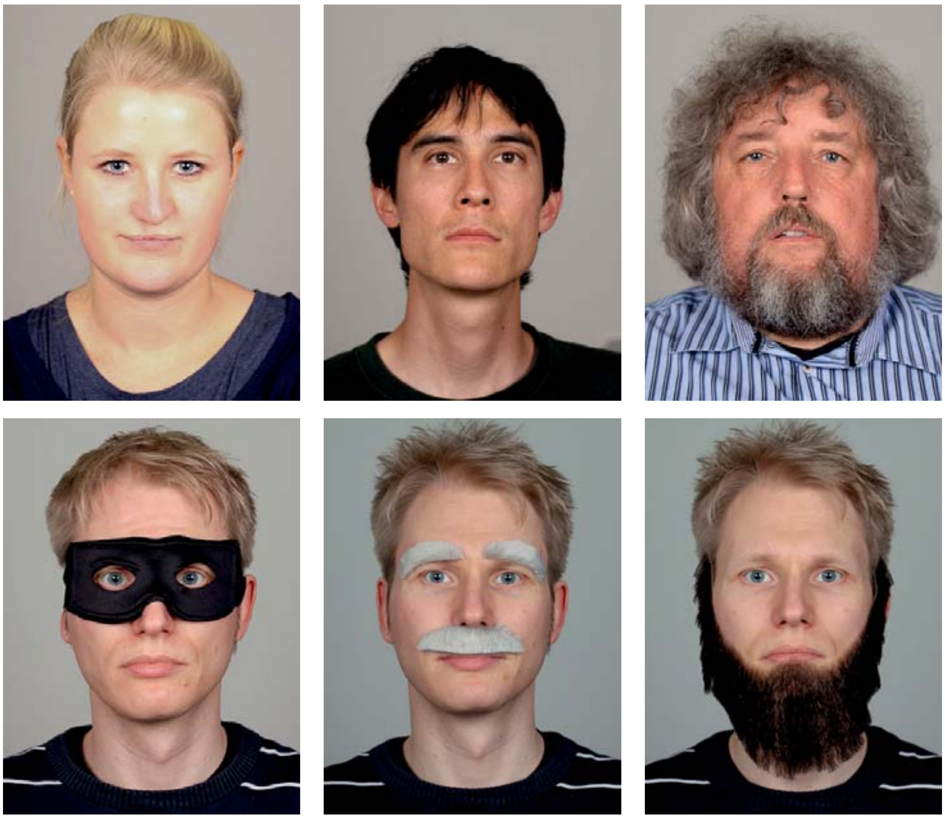


Figure 5.2: Portraits of persons showing different amounts of skin in the facial region (upper row) compared to a face with different partial disguises (lower row). Previously presented in [22].

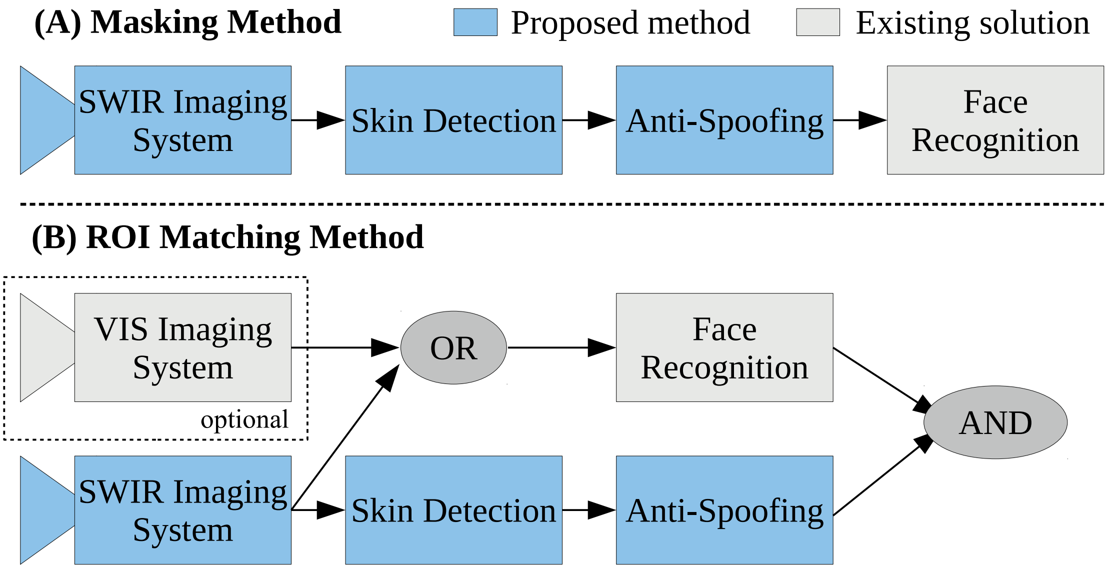


Figure 5.3: Components of the proposed anti-spoofing methods: (A) masking out of non-skin regions; (B) region of interest (ROI) matching. Previously presented in [22].

in a preprocessing step. This ensures that no (possibly forged) non-skin areas are used for the recognition process.

(B) Generic regions of interest (ROIs): This method can be applied to any given face recognition system. In addition to the SWIR image required for anti-spoofing, a VIS image of the face can optionally be acquired and used for face recognition instead of the SWIR image. The two cameras do not need to be co-registered as long as they have a similar field of view and a negligible baseline shift to ensure that both cameras capture the same face. Here, skin classification is applied on the SWIR image and anti-spoofing is performed based on a generic region of interest in a postprocessing step.

Both of these methods can be applied if subjects have been enrolled using either typical VIS images or appropriate SWIR images similar to those that are used for query. They consist of the following components; see Figure 5.3 on the facing page:

A multispectral SWIR image source with at least three well-chosen wavebands in the range of approximately 900 nm to 1 600 nm. Additional wavebands can be used to further increase the reliability and accuracy of the skin classification method. For the concept validation presented in Section 7.5, the camera system described in Chapter 6 with four wavebands around 935 nm, 1 050 nm, 1 300 nm and 1 550 nm is used, but the proposed approach can be applied to other image sources as well.

A face recognition and verification module that is considered as a black box and is potentially implemented as academic state-of-the-art or commercial off the shelf software. For both face detection and recognition, a waveband around 1 050 nm was found to be suited best, as the remission intensity of skin is comparably high in this waveband, with eyes and mouth appearing darker. Especially if subjects have been enrolled using VIS images, this waveband has advantages over higher wavebands when being compared to the reference image.

An accurate machine learning-based skin classifier trained on authentic skin samples, as well as a variety of relevant material samples, which include different types of makeup and materials that might be used for spoofing attacks.

An innovative anti-spoofing module that detects spoofing attacks reliably without rejecting authentic faces due to facial hair or uncritical occlusion of skin. This module has two modes of operation (see above): *masking out non-skin regions* as a preprocessing step to face recognition systems that can work on SWIR imagery as input, or *region of interest matching* as postprocessing of the FR systems verification result.

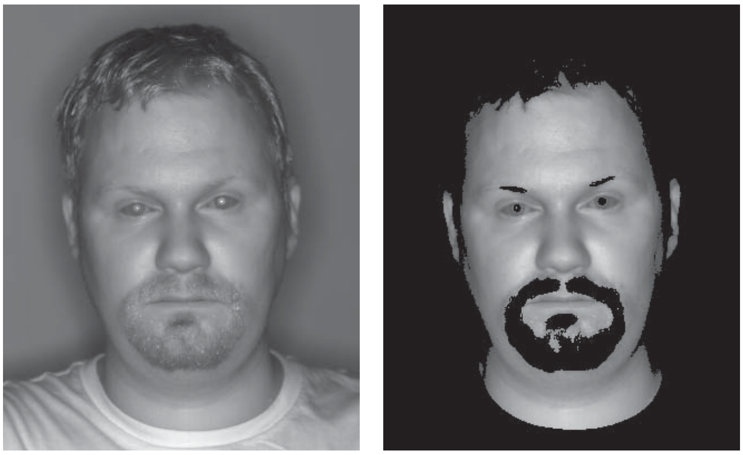


Figure 5.4: SWIR image before (left) and after (right) masking out all non-skin pixels. Previously presented in [22].

5.3.1 Masking Out Non-Skin Pixels

This first method, (A), see Fig. 5.3, to integrate SWIR-based skin detection into existing face recognition systems removes, or masks out, all pixels that have been classified as “non-skin” in the input images as a preprocessing step before the SWIR image is analyzed by the face recognition algorithm. If the subjects have been enrolled using VIS images, this method requires that the face recognition module is capable of matching these images with the SWIR face images acquired for the query.

The basic principle of the masking method is comparable to the approach described by Dhamecha *et al*. [104], which has been described in Section 3.2.3 on page 36. However, the method proposed here is much more fine-grained, as the decision whether or not to use a certain facial area for the face recognition is made for each pixel individually instead of larger patches. This ensures that no forged information will be contained in the image used for face recognition, while all authentic information is maintained. Fig. 5.4 shows a face image before and after masking.

5.3.2 Generic Regions of Interest (ROIs)

The alternative approach, (B), to masking non-skin regions verifies the authenticity of a face in a postprocessing step using a generic region of interest (ROI). This method does not impose specific constraints on the face recognition module as such. Especially, the face recognition system can be fed with either SWIR or VIS query images, whatever the system requires.

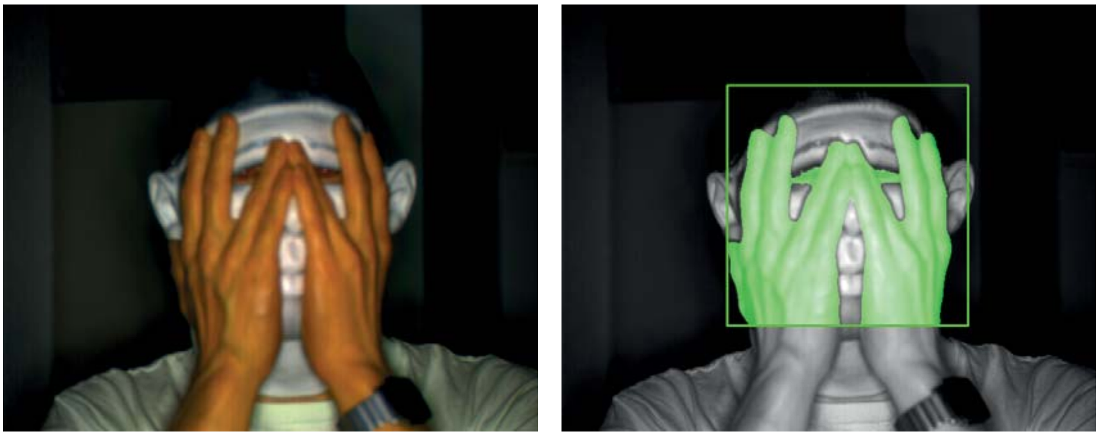


Figure 5.5: Example of an attack that overcomes a simple spoofing detection based on the total amount of skin in the facial area.

As shown in Fig. 5.3, the anti-spoofing module uses the SWIR face image to check the authenticity of a face presented to the system. In postprocessing, this information is combined with the result of the face recognition module. If a face has been verified by both the face recognition (*i.e.*, the captured face image matches the claimed identity) and the anti-spoofing module (*i.e.*, the captured face is authentic), it is accepted by the system.

Template Design

A simple approach to detect a spoofing attack based on the SWIR image would be to measure the total amount of skin in the complete image or facial region. However, this approach is too simple to distinguish actual spoofing attacks with partial disguises from partial occlusions by facial hair, for example, as shown in Figure 5.2 on page 62. Furthermore, this approach potentially opens up new possibilities to attack the face recognition system: an attacker could cover a presented mask partially using his hands, for instance, to generate a sufficiently high amount of authentic skin in the facial area; see Figure 5.5.

Therefore, this work proposes a different approach that restricts the skin verification to regions in the human face that are commonly not occluded by, *e.g.*, facial hair: the central area around the nose and eyes, as well as the mouth. Biometric face verification systems are usually robust against changing hair styles or beards, which leads to the hypothesis that these regions are most significant to be checked for skin authenticity.

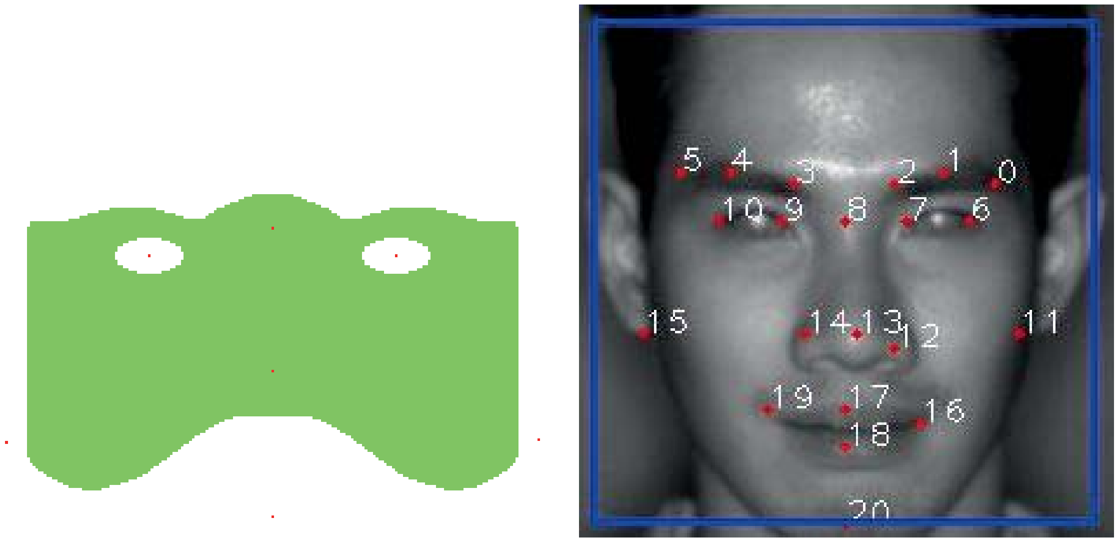
Based on this hypothesis, a generic template of the central facial area has been deduced, which includes only those areas that can be expected to show uncovered skin for every subject. The template is shown in Figure 5.6a on the facing page. The template's shape and dimension has been experimentally optimized using a database of face images, which includes several persons wearing a full beard; see Section 7.5.1.

Template Matching

In order to match the template to a captured image, a feature-based cascade classifier from the openCV library is used to detect faces in the first step. In the second step, the facial landmark detector presented by Uricar *et al*. [115] is applied on the locations of detected faces to locate facial features. Using a previously trained model, this approach is capable of detecting a set of 20 landmarks, as shown in Figure 5.6b on the next page. The algorithm is robust against (moderate) rotation and changes in perspective and allows to estimate the orientation and pose of a face with high accuracy and real-time capable processing times in the order of milliseconds [115].

After extracting the facial features, three relevant orientation points are derived from them. The center positions of both eyes are found by averaging features 6 and 7, as well as 9 and 10, respectively, while the center position of the mouth is calculated as the center of features 16, 17, 18 and 19. These three points have shown to be most stable under motion and changing illumination conditions, as variations in the locations of single features can be compensated to a certain extent. Based on these points, an affine 2D transformation matrix is calculated and applied on the template. Then, its width is adjusted to the width of the face, which is estimated from features 11 and 15. In addition, the features marking the outer edges of the mouth are used to calculate form and position of the lips and the outlined area is added as a second template.

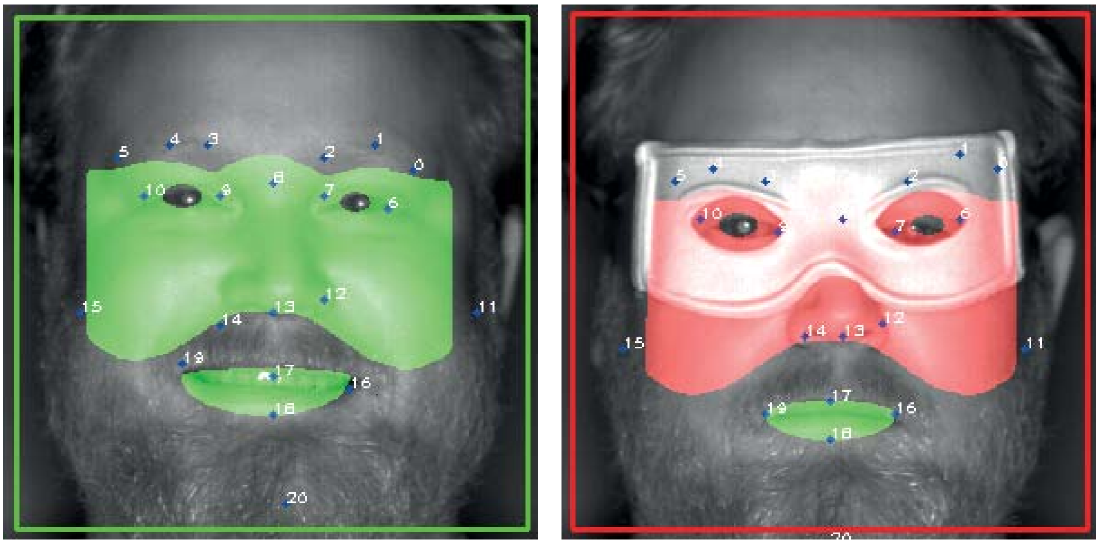
Finally, the amount of authentic skin pixels is calculated for both ROIs. As the template matching process suffers from slight inaccuracies of the landmarking algorithm, the matching is not always perfect. Therefore, the threshold for the verification must be set to a level that tolerates these inaccuracies while being sensitive enough to reject any spoofing attacks. Similarly to the template design, the optimal threshold has been found experimentally based on a database of face images from more than 150 persons; see Section 7.5.1. It has been set to 90% of the pixels in the central face area and to 50% of the pixels in the mouth area. This setting works for all faces in the database and is expected to be sensitive enough to detect spoofing attacks. Figure 5.7 on the facing page shows an example of the successful template matching.



(a) Template of the central face area.

(b) Example of a facial landmarking result.

Figure 5.6: Components of the ROI matching method. Previously presented in [22].



(a) Face without mask

(b) Face with partial mask

Figure 5.7: Results of the ROI matching method (B). Green: successful verification; red: spoof detected. Previously presented in [22].

5.4 Summary

In this chapter, a pixel-level skin detection approach and its application for the task of face verification and anti-spoofing are presented and described. The two-stage skin classification method is based on the analysis of spectral signatures in the SWIR spectral range, which is ideally suited to distinguish human skin from other materials as described in Section 2.2. The spectral signatures $\vec{s}(x, y)$ of each individual pixel are extracted and normalized differences $\vec{d}(x, y)$, which are independent from absolute pixel intensities, are calculated using all combinations of components in $\vec{s}(x, y)$.

In the first stage, a coarse-grained "difference filter" classifier applies upper and lower thresholds on $\vec{d}(x, y)$ and creates a binary output matrix, classifying each pixel as either "skin-like" or "non-skin". The thresholds are defined to include all skin samples from a training dataset. In the second stage, the remaining positive samples are fed into a more fine-grained classifier that is based on machine learning techniques and has been trained with a large amount of both positive and negative samples. This two-stage approach ensures high classification accuracy with minimal computational costs.

In order to combine the per-pixel skin classification with face recognition, an anti-spoofing method with two modes of operation is proposed that enhances existing face recognition solutions and ensures the authenticity of a face, while rejecting both two- and three-dimensional facial masks and (partial) disguises. The two modes are (A) the masking of non-skin pixels as preprocessing step to face recognition systems that can handle SWIR imagery as input and (B) verification of a generic region of interest as postprocessing step after a successful recognition, which can be performed either on SWIR or visual (VIS) spectrum images from a (probably already existing) second camera. For method (B), a suited ROI template has been designed based on a dataset of face images. It covers the central facial area, which is typically free of facial hair, as well as the mouth area. By using a facial landmark detector, the template is matched to a subject's face and the areas covered by the template are checked for authenticity using a threshold of positively classified pixels.

A detailed evaluation of the proposed methods is presented in Chapter 7.

Chapter 6

System Design

In this chapter, a specific system design for a skin detecting camera system called *SkinCam* is presented that implements the reference design proposed in Chapter 3. Besides an overview of the setup, implementation details of hard- and software and an analysis of eye safety for the active illumination module are given. Finally, an approach to estimate depth information from axial chromatic aberrations is described.

Publications: An earlier development stage of this system design with a focus on face verification has been presented in [21]. Here, the system design has been enhanced and is described in greater detail, including variations and possible modifications. Estimating depth from defocus introduced by chromatic aberration has already been covered in work by Velte [116].

6.1 Camera System Setup

This section introduces a system setup for an active multispectral short-wavelength infrared (SWIR) camera system for skin detection which will be denoted as *SkinCam*. This setup is composed of three major building blocks which are illustrated in Figure 6.1 on the next page. These blocks have been derived from the reference design proposed in Section 3.3 on page 39. They will be explained in sequential order in the following sections: Section 6.2 on page 71 describes the implementation and design decisions for the camera system with a focus on the hardware and the software architecture. Section 6.3 on page 82 presents the implemented image processing methods, while Section 6.4 on page 86 focusses on the application level software.

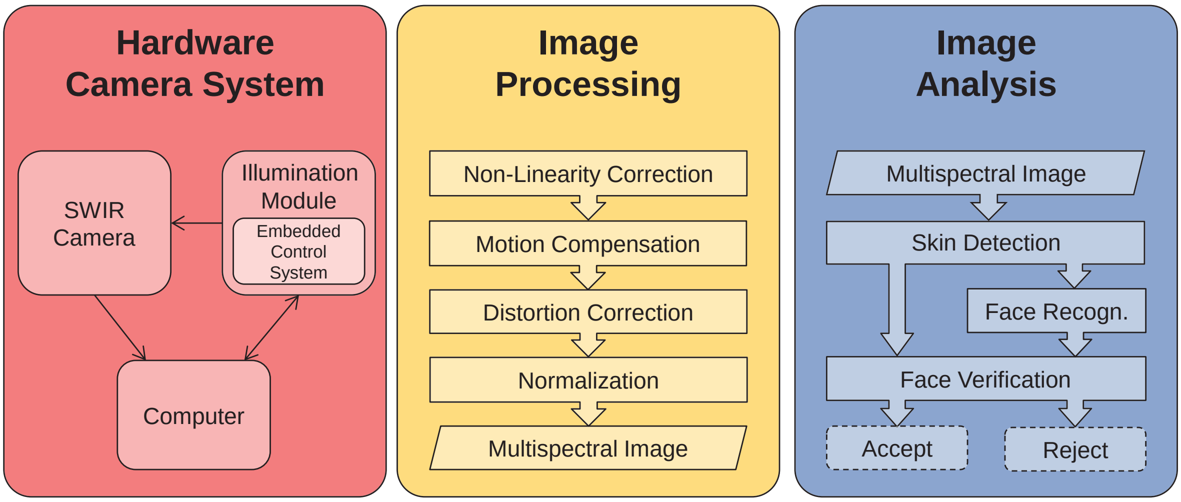


Figure 6.1: Building blocks of the proposed implementation. Prev. shown in [21].

Based on the design goals specified in Section 3.3, this work proposes a system setup consisting of a single SWIR camera that is sensitive to a spectral range of 900 nm to 1 700 nm with an attached illumination module that illuminates the camera's field of view in up to four distinct narrow wavebands within this spectral range (one at a time), as illustrated in Figure 6.2 on the facing page. For the implementation of the illumination module, the use of light emitting diodes (LEDs) is an obvious choice, as they produce rather narrow band illumination without the need for additional band pass filters and can be pulsed with high intensities and variable frequencies. A microcontroller system, which is embedded into the ring light module, triggers short pulses in alternating wavebands and signals the camera to start and stop the exposure of a new image synchronized to the light pulse. The camera transmits the acquired images to a connected computer via Gigabit Ethernet, which in turn is connected to the microcontroller system via USB in order to configure and start the acquisition. In addition, a special software tool has been developed that allows a user to control the image acquisition and to perform all related image processing and analysis tasks with a graphical user interface.

In practice, the illumination module is working as a pulsed light source. The microcontroller system enables its different wavebands one after the other in a fixed order and simultaneously triggers the camera exposure. To remove the influence of ambient light, in each acquisition cycle an additional camera exposure is triggered without the LEDs flashing. This *"dark" reference image* is subtracted from each of the *waveband images* or *channels*, respectively, in preprocessing, so that only light emitted by the illumination module in one single waveband remains. Each set of waveband

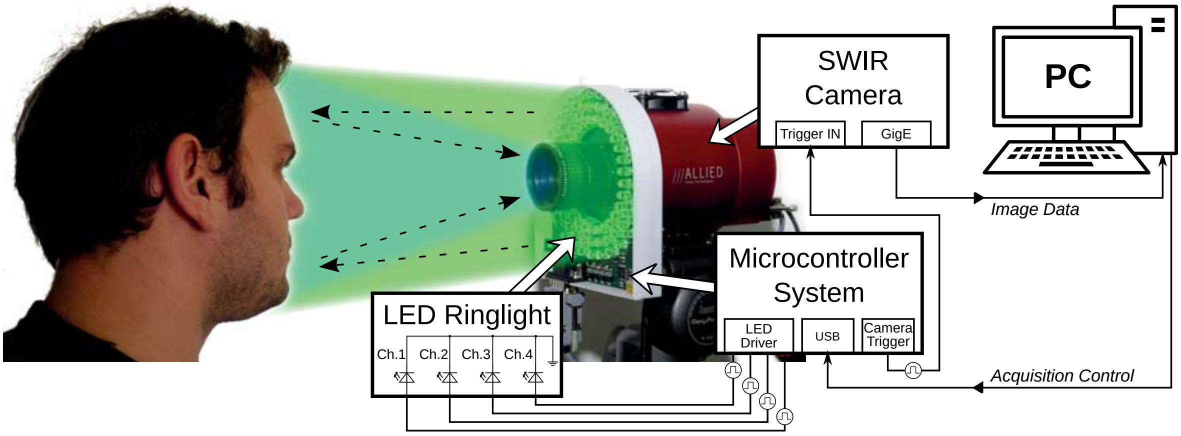


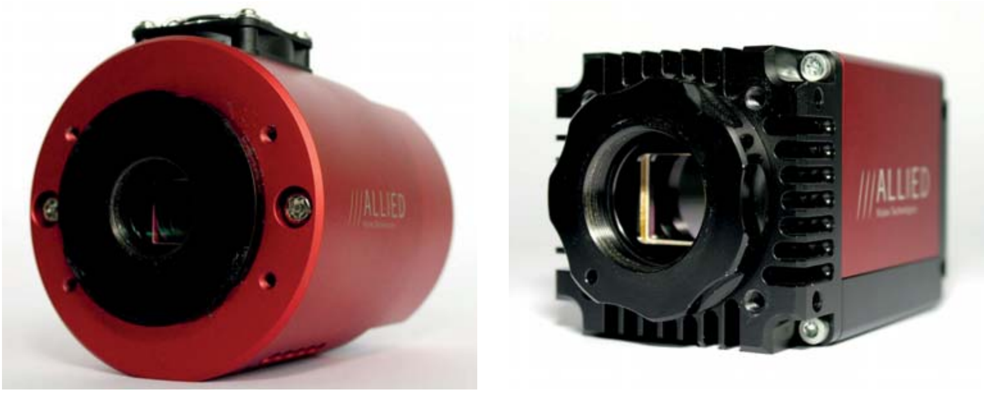
Figure 6.2: Schematic of the camera system setup. Previously presented in [21].

images and its corresponding reference image are combined in a *multispectral image cube*. This method works well for ambient light originating from continuous light sources, such as daylight. Here, all light sources with intensity variations that are either very slow or very fast compared to one full acquisition cycle can be regarded as continuous. However, "flickering" or pulsed light sources, which change their intensity with frequencies similar (but not identical) to the acquisition frequency, might cause distortions of the spectral signatures. In practice, most flickering light sources are incandescent or fluorescent electric lamps, flickering at twice the local power line frequency of 50 Hz or 60 Hz, therefore having periods of 10 ms or 8.3 ms, respectively. By using an exposure time that matches this period or any multiples of it, the flickering can easily be reduced to a negligible level.

6.2 Implementation Details

6.2.1 SWIR Camera

At the time of writing, most commercially available cameras for the SWIR spectrum are based on indium-gallium-arsenide (InGaAs) detectors. As described in Section 2.3 on page 15, InGaAs is a very efficient semiconductor detector material and has a high responsivity up to wavelengths of 1.7 µm. Current InGaAs focal plane arrays have a spatial resolution of up to 640x512 pixels and can be read out with very high frame rates, if necessary.



(a) Goldeye P-032 (b) Goldeye G-032

Figure 6.3: SWIR cameras used in this work (not the same scale).

In the context of this work, two *Goldeye*[1] cameras are used, which integrate identical FPAs and provide industrial standard *GigE* network and external trigger interfaces with optical isolators. In both cameras, the FPAs are cooled to an operating temperature of using a thermoelectric cooling system based on the Peltier effect. The cameras are shown in Figure 6.3. The older Goldeye P-032 achieves a frame rate of 30 frames per second (FPS), while the newer Goldeye G-032 reaches up to 100 FPS and features a much smaller casing and redesigned cooling system. The most important technical specifications are listed in Appendix A, Table A.1 on page 139. For comparison, technical data of the fastest camera that is currently available on the market, the Xenics Cheetah 640 CL, is given as well. The Cheetah allows to capture images at a frame rate of 1730 FPS using multiple *CameraLink* interfaces. For an active camera system with frame rates in this order, motion compensation would not be necessary for application scenarios that involve nothing faster than human movements. Unfortunately, due to its significantly higher price, it could not be used for this work. With respect to the sensitivity of the detector array, all cameras are very similar: the Goldeyes have larger detector cells with 25 μm x 25 μm compared to the Cheetah's 20 μm x 20 μm, which compensates for their slightly lower quantum efficiency of approx. 73% compared to 80% peak efficiency.

In addition to the camera itself, a suited lens is necessary as well. Many lenses for imaging systems are made of materials and with applied coatings that are optimized for wavelengths in the visual spectral range [45], for example in order to remove unwanted reflections. This might lead to unforeseen absorption bands in the captured SWIR spectra or increased aberrations; see Section 2.4 on page 17. To address this

[1] Allied Vision Technologies GmbH, Stadtroda, Germany (http://www.alliedvision.com)

issue, some manufacturers offer lenses that are optimized especially for the SWIR spectral range. For the *SkinCam* system proposed in this dissertation, lenses from EHD Imaging[1] are used. Although these lenses are designed for an image format of one inch, *i.e.*, a diagonal of 16 mm [117], while the Goldeye cameras have a larger image format of approx. 20.5 mm in diagonal, they illuminate the full sensor area with only marginal vignetting at the outer edges. The required focal length for the lenses depends on the particular application.

A suited focal length for face recognition applications can be calculated using the lens Equation (6.1) [117]. By specifying the desired image size I as the full height of the FPA, the object size O as the average height of a human head according to industrial standard DIN-33402-2 [118] and the desired minimum distance between lens and object d_O, the equation yields the optimal focal length f.

$$f = \frac{d_O \cdot I}{O + I}. \tag{6.1}$$

With an object distance of $d_O \geq 1\,\mathrm{m}$, a head height of approx. $O = 0.235\,\mathrm{m}$ and the height of the FPA $I = 12.7\,\mathrm{mm}$, an optimal focal length of $f_{\mathrm{opt}} \approx 51.3\,\mathrm{mm}$ is calculated. Therefore, the *EHD50HC-SWIR* lens with $f = 50\,\mathrm{mm}$ has been selected for the face recognition application scenario, which produces an angle of view of $\alpha \approx 18°$ along the (wider) x-axis when used with the Goldeye cameras.

6.2.2 Microcontroller System

The microcontroller system is embedded into the illumination module and triggers LED pulses and the connected camera simultaneously. The cameras are configured to start the exposure immediately when the trigger input changes to high level and to keep exposing until it returns to low level. To avoid a negative influence of flickering or pulsed light sources in the vicinity of the camera system, an exposure time of 10 ms is used within areas with a local power line frequency of 50 Hz and an exposure time of 8.3 ms within areas using 60 Hz, respectively. As the used Goldeye cameras have a very short readout time, the G-032 can be operated at its maximum frame rate of 100 FPS and still maintain an exposure time close enough to $10ms$ to remove the flickering effect of electric lamps. Figure 6.4 on the following page illustrates the chronological order of the signals given by the microcontroller system within one full acquisition cycle of 50 ms, resulting in an effective multispectral frame rate of 20 FPS when using four distinct wavebands plus reference channel.

[1] EHD imaging GmbH, Damme, Germany (http://www.ehd.de/)

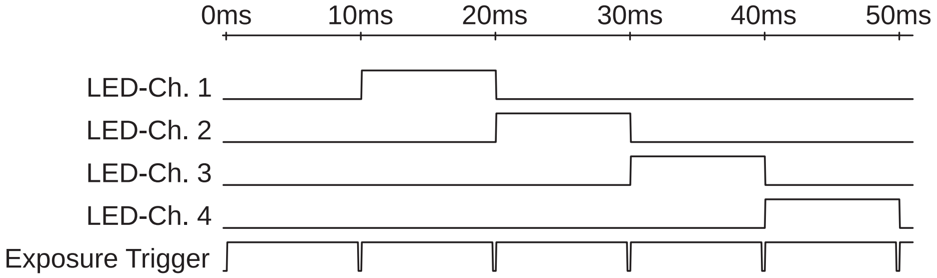


Figure 6.4: Timing diagram of the signals given by the microcontroller system. Previously presented in [21].

The control system is implemented using an Atmel ATmega 168 8-bit controller. An FTDI serial interface to USB converter chip is used to connect the controller with a desktop computer. Both the FTDI and the controller are powered with 5 V supply voltage via USB. Schematics of the circuit design and the layout of the printed circuit board can be found in appendix A.

Figure 6.5 on the next page illustrates the standard program flow of the control system. In a first step, timers and interface ports are initialized and interrupts are configured. One timer is configured according to the specified frame rate and, on overflow, triggers an interrupt service routine (ISR) that sets a flag indicating the start of a new light pulse and exposure phase. A second timer is configured to match the specified integration / exposure time and triggers an ISR that sets a stop flag on overflow. Similarly, if data is received on the serial interface, *i.e.*, from the connected computer, a notification flag will be set by another ISR.

Afterwards, the program performs an endless loop that checks if any of the flags is set and reacts to it appropriately. If a command was received, the command and optional parameters are read and interpreted. The possible commands are *start* or *stop* of the first timer (and, thus, the image acquisition procedure), the change of the *frame rate* or *exposure time*, as well as the *waveband configuration*. If the start flag is set, the next waveband in line is selected, the respective LEDs are turned on and the exposure is started by setting a high level to the camera's trigger input. If the stop flag is set, the LEDs of the current waveband are turned off, the exposure is stopped by setting the camera's trigger input back to low level and finally, the current waveband number is transmitted to the connected computer for synchronization.

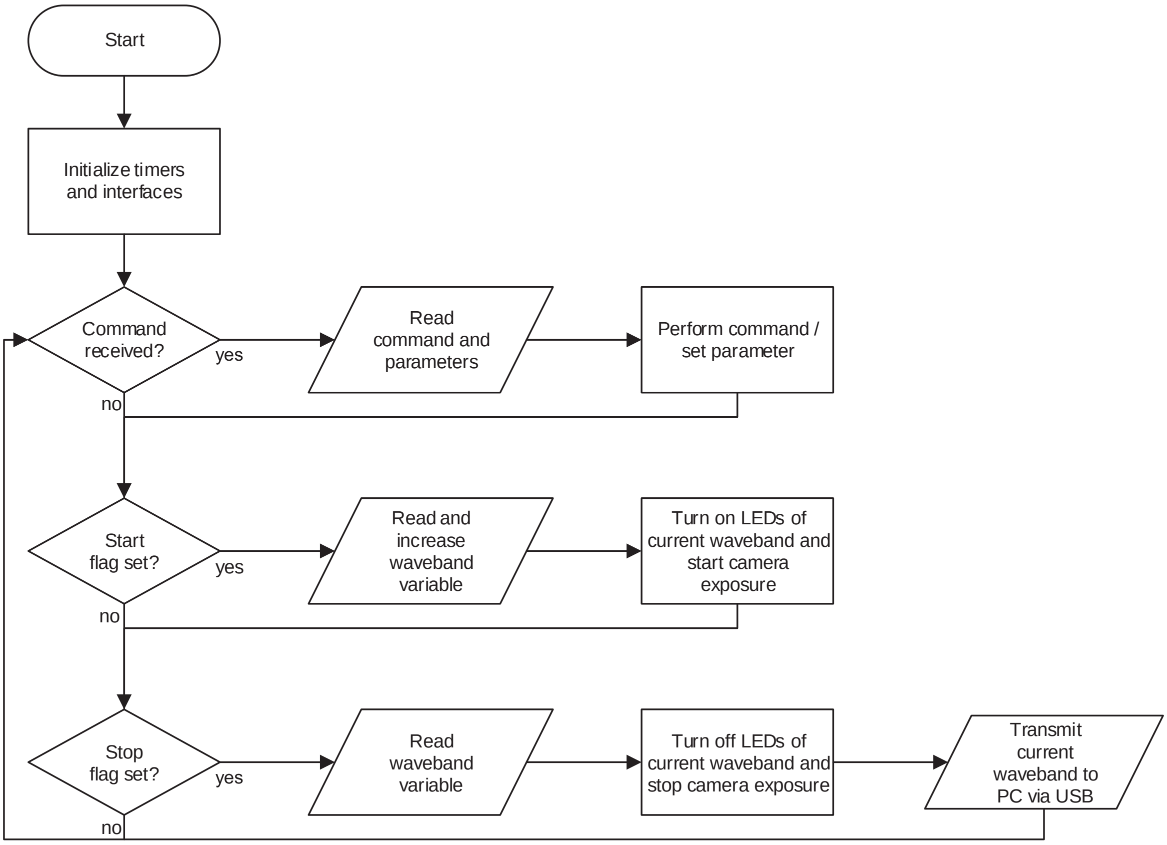

Start
Initialize timers and interfaces
Command received?
yes
no
Read command and parameters
Perform command / set parameter
Start flag set?
yes
no
Read and increase waveband variable
Turn on LEDs of current waveband and start camera exposure
Stop flag set?
yes
no
Read waveband variable
Turn off LEDs of current waveband and stop camera exposure
Transmit current waveband to PC via USB

Figure 6.5: Program flow of the microcontroller system.

6.2.3 Waveband Selection

In previous work [20, 25, 82], a large amount of human skin and material samples have been analyzed using both a visual (VIS) and a SWIR spectrometer. The captured spectra have been stored in a spectral remission database. In the context of this dissertation, remission spectra of additional skin and material samples, including a selection of facial disguises and masks, have been acquired and added to this database.

To evaluate the acquired data and to find a small set of wavebands that can be used successfully for material classification, the data mining software *AnaSpec*, initially developed by Schwaneberg *et al.* [82], has been applied on the extended remission database. It performs a brute force search over all possible combinations of a given number of wavebands and calculates the resulting normalized differences (see Section 5.2.1 on page 58) in order to find the set of wavebands that separates skin and other materials best. For this purpose, AnaSpec simulates the typical emission spectra of LEDs with the respective peak wavelengths λ_p and takes their expected full width at half-maximum (FWHM) into account. This is done by performing a convolution of a specific sample's remission spectrum with the LED's emission spectrum, which is simulated based on a database of several reference LED types.

In practice, the total number of LEDs used on the illumination module is limited, as costs, power supply and space need to be considered. Therefore, a compromise has to be made between the number of wavebands and the achievable radiated output power per waveband. For face verification, the demand for classification accuracy is higher than the demand for high operating distances. In addition, the selection of wavebands has to be restricted to those of LED chips that are actually available on the market.

In his work, Schwaneberg [25] selected a combination of four LEDs with wavebands of $\lambda_1 = 830\,\text{nm}$, $\lambda_2 = 1\,060\,\text{nm}$, $\lambda_3 = 1\,300\,\text{nm}$ and $\lambda_4 = 1\,550\,\text{nm}$. Here, a set of four wavebands was also found to be very promising, as this still allows for a sufficient number of LEDs per waveband in order to achieve an acceptable operation range in indoor scenarios. Taking the newly acquired material samples used to create facial disguises and masks, as well as the sensitivity curve of the camera's InGaAs detector into account, a good separation of skin and material samples can be achieved by choosing $\lambda_1 = 935\,\text{nm}$ for the first waveband and keeping the other wavebands as proposed by Schwaneberg.

6.2.4 Design of the Illumination Module

Within the scope of this dissertation, two illumination modules have been implemented that are both focused on the face recognition application scenario. Regardless of the application at hand, a uniform distribution of the LEDs around the camera lens, as well as similar viewing angles and radiant patterns of the different LED types are very important in order to achieve a homogeneous illumination of the scene. Otherwise, the extracted spectral signatures of an object would differ depending on the object's position in relation to the illumination module. To avoid this problem as much as possible, LEDs of the same model and manufacturer have been selected. In addition, optical simulations have been performed to find the optimal distribution for the different numbers of LEDs per waveband.

Selection of LEDs

The first evaluation model that was created in the context of this work has already been described in [21] and is targeted on face recognition applications. It consists of 90 LEDs in standard 5mm packages with plastic lenses in four distinct wavebands: $\lambda_1 = 935\,\text{nm}$, $\lambda_2 = 1\,060\,\text{nm}$, $\lambda_3 = 1\,300\,\text{nm}$ and $\lambda_4 = 1\,550\,\text{nm}$. The number of LEDs for each waveband was chosen with regard to both the expected radiated power of each waveband and a uniform distribution of the LEDs on the module.

The first waveband was designed with a much higher power output of $\sum \Phi_e(\lambda_1) = 300\,\text{mW}$ (at a forward current of $I_F = 100\,\text{mA}$) compared to the other channels with only 150 mW to 170 mW to compensate for the lower responsivity of the camera's InGaAs detector array in this spectral range. Nevertheless, it was found that the camera's responsivity in this waveband is even lower than expected and as a result, the noise level within the captured images is strongly increased. During practical use of the camera system, additional room for improvement has been recognized and a new revision of the illumination module has been developed.

In the second revision, the waveband at 935 nm was exchanged for a new waveband around 1 200 nm. According to the spectrometer data described in Section 6.2.3, the resulting combination is similarly well suited for face anti-spoofing and was found to be much more reliable in practice. In this revision, the illumination module contained 200 LEDs in a larger radius to improve both remission intensity and homogeneity of the illumination. The total radiated power per waveband was increased to values between 264 mW to 340 mW at a forward current of $I_F = 100\,\text{mA}$, which allows for continuous operation. Due to the pulsed operation of the module during multispectral image acquisition, the forward current can be doubled in practice, which increases the radiated power output to an unknown extent. For both illumination modules, specifications of the used LEDs can be found in Appendix A, Table A.2 on page 140.

Simulation of Designs and Placement Patterns

To find a suited placement pattern for the LEDs, the different LED types have been modeled as light sources using the optical simulation software FRED Optimum developed by Photon Engineering, LLC. Their typical peak wavelengths, spectral and radiant power distributions have been specified according to their datasheets. FRED performs ray tracing to simulate the propagation of light from each light source to a virtual target plane. It also provides a scripting language and batch processing capabilities to run a series of simulations with different parameters. This way, different placement patterns and varying positions for the LEDs can be compared by simulating the resulting intensity distribution for each waveband on a target plane. Ideally, the normalized intensity distributions of all wavebands should be identical, leading to a homogeneous "color" on the target. In addition, the overall illumination intensity should also be distributed as homogeneously as possible over the full plane in order to make the best of the camera's dynamic range. In contrast to this, an intensity peak in the center of the plane and strong light falloff to the outer edges, for example, reduces the dynamic range that is still available to detect actual differences in remission intensities of different material surfaces. Besides these two functional constraints, there is also a third, non-functional constraint: for easier handling and practical reasons, a smaller size of the illumination module is favorable.

With these requirements in mind, several designs and placement patterns have been simulated. Besides circular designs, which are advantageous with respect to size, rectangular designs have been created as well. For all simulations, the virtual target plane was adapted to the field of view of the camera with a $f = 50\,\text{mm}$ lens attached to it, *i.e.*, at an angle of view of $\hat{\alpha}_x \approx 18.2°$ along the x-axis and $\hat{\alpha}_y \approx 14.6°$ along the y-axis. Simulated light rays reaching the target plane are captured using an array of 41x33 simulated detectors. The different LED types in the four SWIR wavebands are visualized in yellow, red, green and blue. Please note that the coloring of the homogeneity plot does not correspond to the colors of the individual wavebands directly: here, the four colors are reduced to three (RGB) colors in order to illustrate their mixing ratio. For this purpose, the RGB channels are calculated as follows: $R = 0.5 \cdot (\lambda_1 + \lambda_4)$, $G = 0.5 \cdot (\lambda_2 + \lambda_3)$ and $B = 0.25 \cdot (\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)$. This way, any inhomogeneities would be noticeable by local changes in the color tones, while constant ratios between the four wavebands are shown as gray tones.

Figure 6.6 on the facing page illustrates the three most interesting designs for the first revision of the illumination module with a total of 90 LEDs and shows the results for spectral homogeneity and illumination distribution over the target plane. In all designs, the wider viewing angle of the 1 300 nm LEDs leads to a slight inhomogeneity of the spectral mixture that increases towards the edges of the plane. In addition, the lower number of 935 nm LEDs introduces an inconsistent intensity distribution of this waveband compared to the others. Using a rectangular design with groups of LEDs, this can be compensated well at the cost of a slightly worse mixture of the other channels. With respect to the size requirements, the design based on three small rings has been chosen for the implementation of the first revision. The final implementation of this *ring light* is shown in Figure 6.8a on page 81.

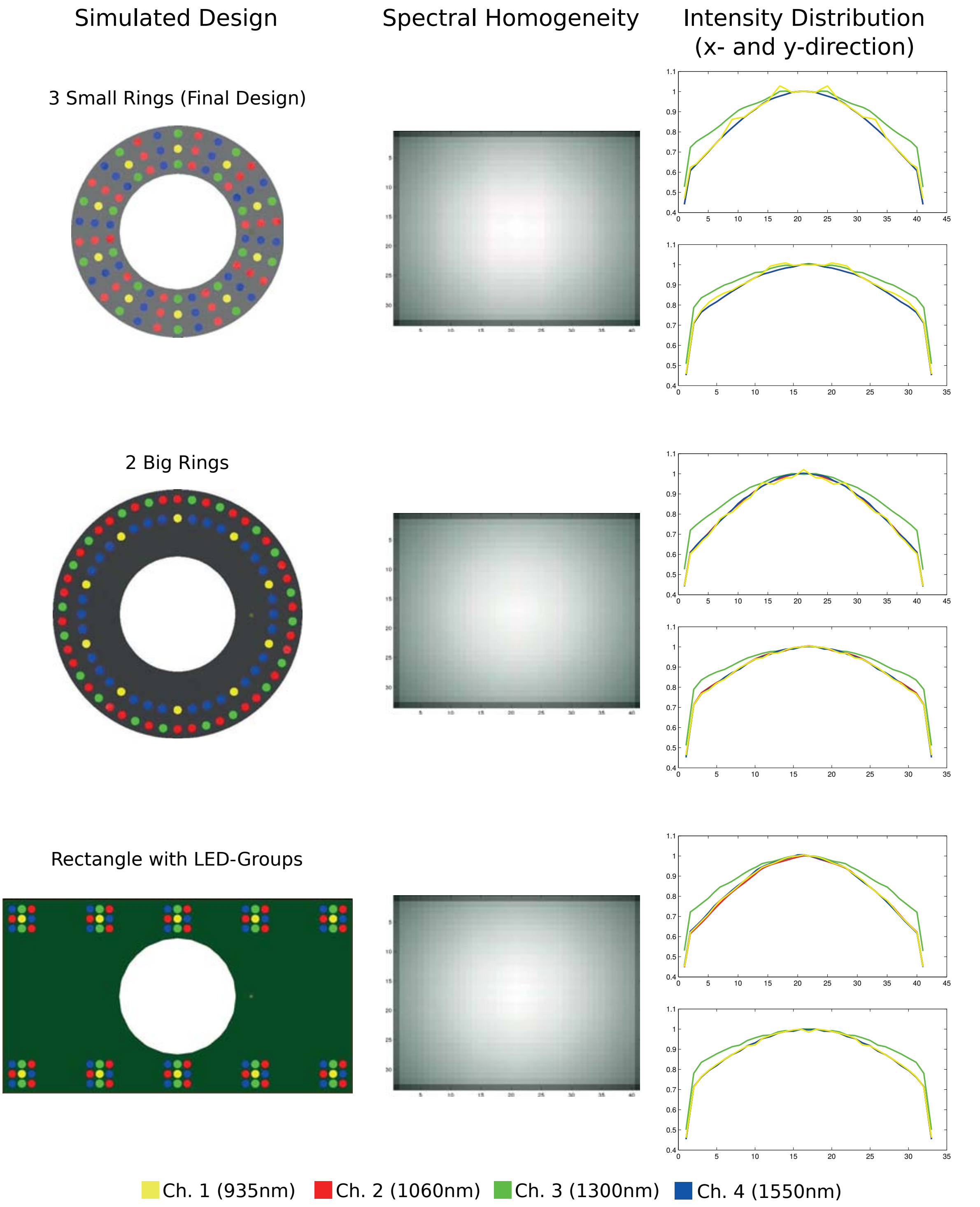


Figure 6.6: Designs and LED patterns for the first revision of the illumination module (left), resulting spectral homogeneity (middle) and distribution (right) projected on a virtual analysis (target) plane covering the cameras field of view.

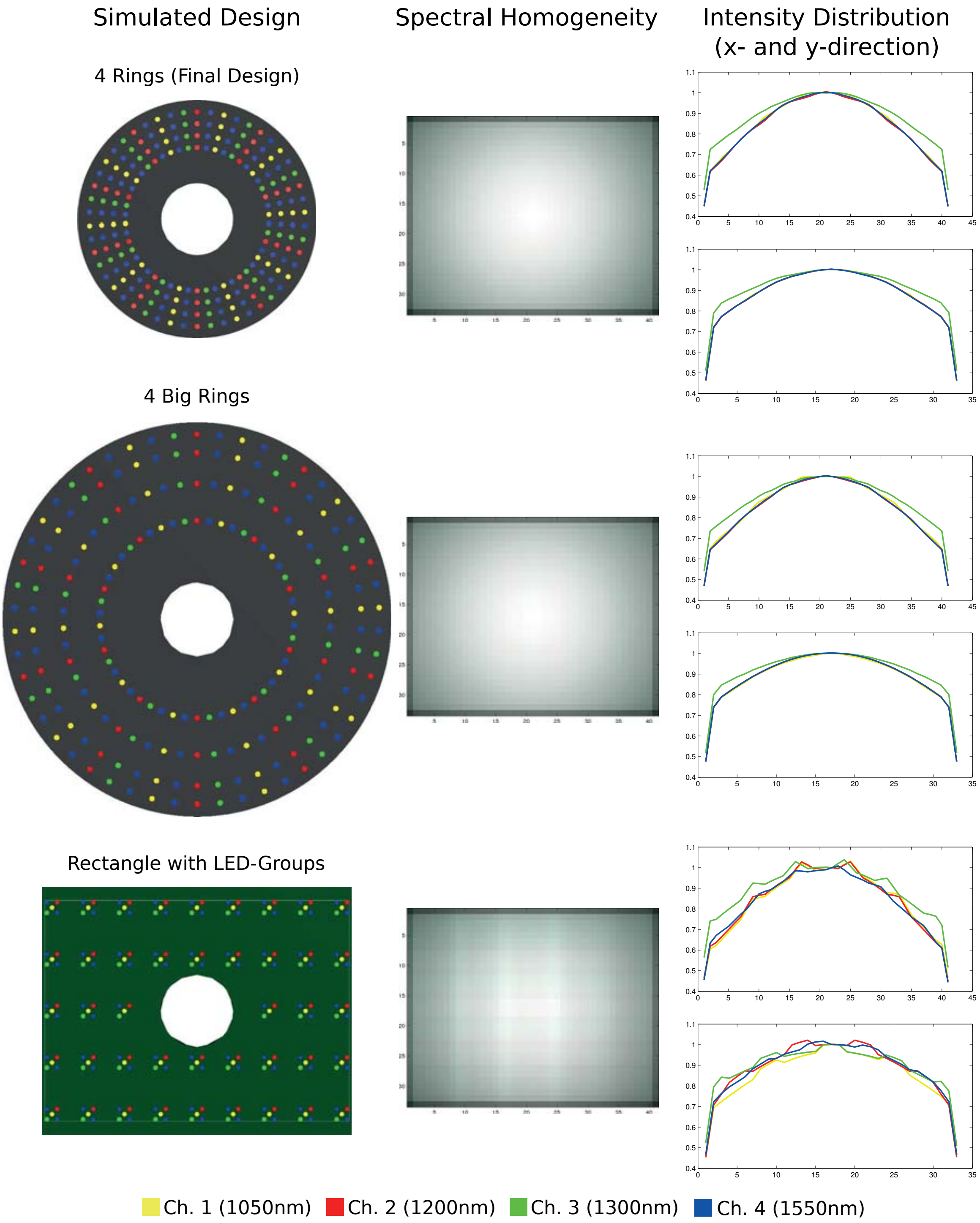
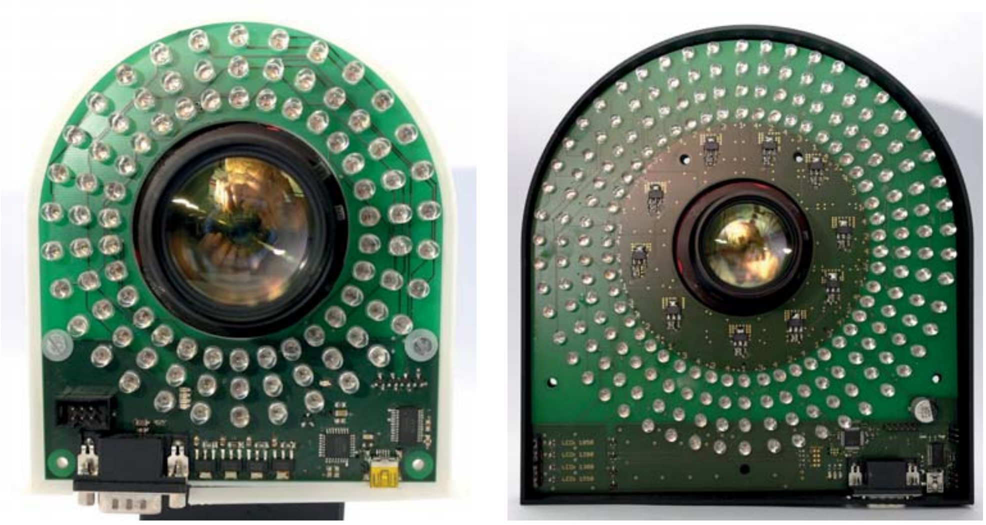


Figure 6.7: Designs and LED patterns for the second revision of the illumination module (left), resulting spectral homogeneity (middle) and distribution (right) projected on a virtual analysis (target) plane covering the cameras field of view.

As the second revision of the illumination module uses 200 LEDs in total, an increase of the size was inevitable for its implementation. Figure 6.7 on the facing page shows three designs for the second revision and their simulation results. The higher and more similar number of LEDs for each waveband allows for a more consistent distribution, which is clearly shown in the spectral homogeneity and spatial intensity distribution of the two circular designs. Compared to them and in contrast to the first revision, the use of a rectangular design produces significantly worse simulation results. As the bigger circular design does not show significant improvements over the smaller one, the latter one was chosen for the final implementation and is shown in Figure 6.8b.

Figure 6.8 presents the final implementations of both ring light revisions attached to the camera in frontal view. Implementation details and schematics of the full circuit design and the board layout for both revisions can be found in appendix A.



(a) Ring light rev. 1. Width: 0.11 m.

(b) Ring light rev. 2. Width: 0.19 m.

Figure 6.8: Comparison of the different illumination modules (scale varies).

6.2.5 PC-based Control and Analysis Software

The *Goldeye* cameras transmit each captured frame via gigabit ethernet to a connected computer, which is also connected to the microcontroller system embedded in the illumination module via USB. To control the camera systems operation, retrieve the captured images and perform the required image processing and analysis tasks illustrated in Figure 6.1 on page 70, a special software tool has been developed using C++

and the Qt framework[1]. In the following, the fundamental architecture of the software will be described, while Section 6.3 and Section 6.4 will present the implemented methods. For more information about the software itself, please refer to Appendix B, which includes details about the used libraries as well as UML class and sequence diagrams.

The *SkinCam* software is composed of several widely independent modules, which run in separate threads and communicate using the signal and slot architecture provided by Qt. Below the main program thread, which also runs the graphical user interface module, there are independent modules and threads for image acquisition, image (pre-) processing, as well as image analysis. Due to this division, the software is optimally suited for a computer with four logical processors. In order to use real time motion compensation as described in Chapter 4, a compatible graphics processing unit (GPU) is required.

The architecture has been designed with a focus on easy exchangeability of single modules and classes. For this purpose, interface classes have been specified which separate the basic functionality from actual implementations and specific hardware. The use of a different camera, for example, would only require to implement a new class that matches the specific camera's programming interface to the functions specified in the abstract camera interface class. Motion compensation or skin detection algorithms could be exchanged in the same way.

6.3 Image Processing

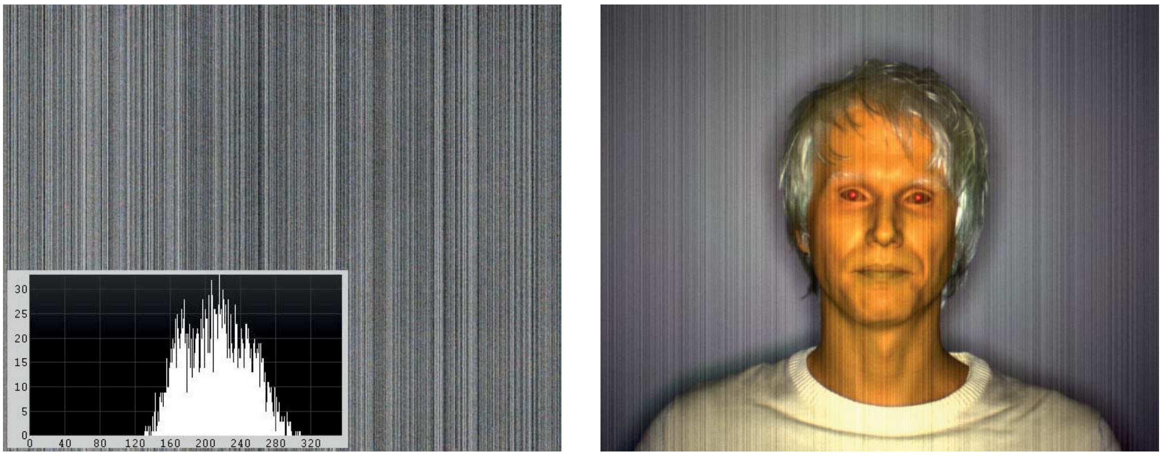
In the first step of image processing, each image that was captured by the SWIR camera is annotated with the ID of the waveband that has been active on the ring light during its exposure. Then, several processing steps are performed in order to calibrate, optimize and merge the images into a multispectral image cube.

6.3.1 Fixed Pattern Noise Correction

Both Goldeye cameras that are used in the context of this work feature an internal two-point non-uniformity correction, which can be adapted to varying exposure times. This correction method sets an individual offset and gain factor for each detector cell on the focal plane array (FPA) in order to compensate for varying sensitivity and dark

[1] Qt is a cross-platform application framework; see http://qt.io/

currents. Nevertheless, it was found that under-exposed images show significant fixed pattern noise that is not sufficiently accounted for by the internal correction and varies with the actual pixel intensities. As the proposed system design requires to capture a reference image without flashing the ring light that is subtracted from the individual waveband channels in order to create a multispectral image cube, this fixed pattern noise has noticeable influence on images taken in dark environments, as demonstrated in Figure 6.9.



(a) Fixed pattern noise on a dark image (accentuated) with histogram (actual 12 bit values).

(b) Fixed pattern noise on a multispectral image cube in false color representation.

Figure 6.9: Fixed pattern noise in images captured with the Goldeye P-032 and its influence on a multispectral image cube (wavebands at 1 060 nm, 1 300 nm and 1 550 nm).

To analyze the sensor's behavior in detail, the sensor area was homogeneously illuminated using an adjustable quartz halogen lamp through an integrating (Ulbricht) sphere positioned in a darkroom and 70 images with increasing brightness were taken. The measurement setup is shown in Figure 6.10 on the next page. The analysis of the acquired images proves that the fixed pattern noise is introduced by an insufficient non-uniformity correction: the darkest and the brightest pixels in each image vary from the average intensity, as shown in Figure 6.11 on the following page.

This analysis shows that the effect is non-linear, especially in the lower intensity range, and thus can not be compensated by a (linear) two-point correction function. To address this problem, three different approaches to compensate for the error of each individual pixel have been evaluated in terms of correction quality and runtime performance: fitting of a polynomial of up to the tenth degree to the full curves of each individual pixel, cubical spline interpolation, as well as piecewise linear interpolation

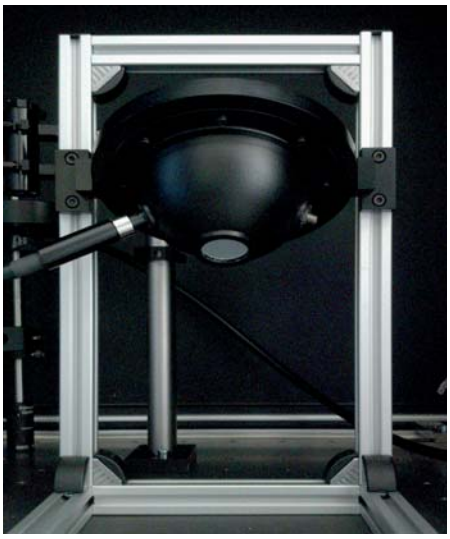


Figure 6.10: Measurement setup with an adjustable lamp connect via light guide to an integrating Ulbricht sphere placed in a controlled environment.

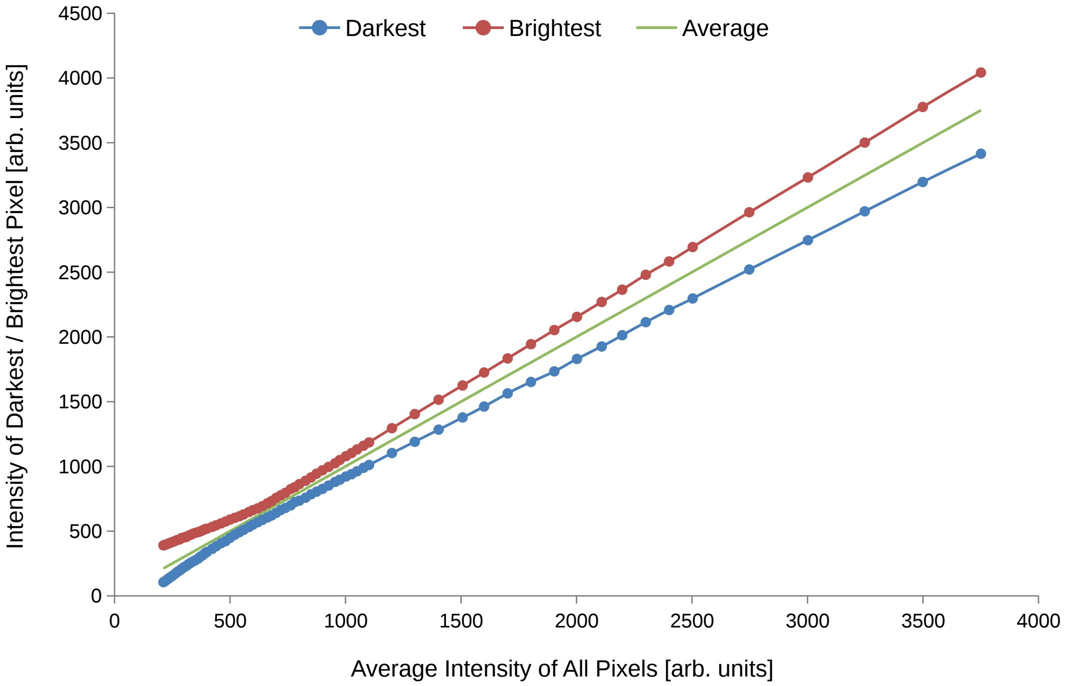


Figure 6.11: Intensity measured on darkest and brightest pixels compared to the average intensity.

using the ground truth data from the 70 captured images as a look up table to apply a multiple-point non-linearity correction to every single pixel.

Using a single thread on a standard desktop computer[1], the polynomial correction can be computed in about 2.5 ms on all 323 068 pixels of the input images. However, the polynomials are no good representation of the data and as a result, this method is not capable of correcting the fixed pattern noise completely. The cubical spline interpolation, on the other hand, matches the data very well and achieves a very good compensation of the noise, but the fastest available implementation still required 18 ms for the correction of all pixels, probably due to inefficient search algorithms. Therefore, the piecewise linear interpolation was implemented using a binary search algorithm, which allows to find the correct pair of ground truth data from the 70 reference images to interpolate between with only 5 comparisons, leading to a run time of about 5 ms with similarly good correction quality; see Section 7.1.1 on page 96.

6.3.2 Motion Compensation

As the proposed active camera system is based on field-sequential waveband capturing (FSWC), motion compensation is required to avoid motion artifacts at the edges of moving objects. In Chapter 4, different approaches to motion compensation for FSWC imagery have been presented. For the implementation of the *SkinCam* system, a trade-off between processing time and compensation quality on SWIR image sequences is necessary to maintain real-time operation. Based on the evaluation results presented in Section 7.2, the best suited combination of preprocessing methods and optical flow algorithm will be selected.

6.3.3 Camera and Ring Light Calibration

After the multispectral image cube has been properly aligned by compensating for object motion, the ambient illumination captured in the “dark” reference channel is subtracted from all waveband images. Then, lens distortion and differences in the illumination intensities can be corrected as last steps in the image preprocessing. For this purpose, three sets of multispectral image cubes are recorded for each lens: first, a checkerboard calibration pattern is captured from different perspectives and at different positions in the image. This set of images is used for the intrinsic camera calibration, as well as to calculate a correction matrix for the lens distortion for every

[1] intel Core i7 4771 CPU, Ubuntu 14.04 64bit, GCC4.8

waveband individually in order to compensate for different distortion characteristics due to lateral chromatic aberration of the lens; see Section 2.4 on page 17.

A second set of images is captured in front of a plain white surface. These images are used to determine both the vignetting effect of the lens and the light distribution pattern of the ring light for each waveband in order to calculate a respective correction matrix that allows to normalize the illumination intensity over the image area.

Finally, a third set of images is captured of a special “white reference” tile that is known to have uniform remission characteristics in the full SWIR spectral range. These images are used to detect absolute differences in illumination intensities between the different wavebands, which are stored as a vector of correction factors and applied on every waveband image as last step in the image preprocessing to achieve a proper balancing of the channels. This process is very similar to the “white balance” settings of common RGB cameras, but in contrast to these passive imaging systems, it is widely independent from ambient light.

6.4 Image Analyis

When preprocessing is finished, the skin detection method described in Section 5.2 on page 58 is applied on the final multispectral image cube. The *SkinCam* software allows to decide whether to use either the difference classifier or a machine-learning-based classifier, or both sequentially, and provides a user interface to change the classifiers’ configurations. The result of the classification process is a binary image, which is used for graphical highlighting of skin areas, as well as the face verification process. For this purpose, a channel within the wavelength range of 1 000 nm to 1 100 nm is extracted from each multispectral image cube and face recognition and verification are applied as described in Section 5.3 on page 61, as this range is best suited for face recognition due to the remission characteristics of human skin.

6.5 Eye Safety Evaluation

Eye safety is a critical aspect of high power SWIR illumination sources, as radiation with a wavelength of up to 1 400 nm can still penetrate the human eye and cause thermal damage to the retina. The directive 2006/25/EG [119] of the European Parliament defines binding permissible limits for illumination systems with pulsed light sources, which should be measured as specified by the applicable standards. For the camera system proposed in this dissertation, the standard IEC 62471 [120] has to be applied. To evaluate the eye safety of the SWIR ring light, the directive defines upper limits for the effective radiance L_R on the retina, which is weighted by a factor that depends on the wavelength of the radiation, as well as for the total irradiance E_{IR} on the cornea at a measurement distance of $d = 0.2\,\mathrm{m}$.

In the early development stages of this work, the necessary measurement setup to perform such measurements was not available. Therefore, the incident power of the SWIR radiation on the eye of an observer standing in the “sweet spot” of the ring light has been analyzed based on optical simulation and information given in the datasheets of the LEDs. For this purpose, a pupil diameter of $\varnothing_{\mathrm{pupil}} \geq 7\,\mathrm{mm}$ is assumed based on information in DIN 33402-2 [118] and a virtual target plane with this size is positioned at a distance of $d = 0.2\,\mathrm{m}$ to measure the incident radiation.

For the first revision of the ring light, the maximum incident power is achieved by the $\hat{\lambda}_1 = 935\,\mathrm{nm}$ waveband and reaches a total irradiance of $E_{IR} \approx 17.3\,\mathrm{W\,m^{-2}}$ at a “sweet spot” position. Due to the close distance to the ring light, there is no sweet spot in the center of the ring light, but rather directly in front of each of the 10 LEDs of this waveband. For the second revision of the ring light, the $\hat{\lambda}_3 = 1\,300\,\mathrm{nm}$ waveband is the most powerful and creates the highest total irradiance. Although the combined output power of this waveband is even higher than that of the 935 nm waveband of the first revision, the total irradiance at a “sweet spot” position only amounts to $E_{IR} \approx 7.64\,\mathrm{W\,m^{-2}}$, as the output power is distributed over 40 LEDs instead of just 10.

Using a simplified model[1] of the ring light, the plausibility of these results can be tested by using the specifications given in the LEDs datasheets: the typical radiant intensity of one 935 nm LED used in the first revision is given as $I_e = 100\,\mathrm{mW\,sr^{-1}}$ and that of one 1 300 nm LED used in the second revision as $I_e = 38\,\mathrm{mW\,sr^{-1}}$. If a worst case scenario is assumed in which all LEDs for this waveband are continuously powered and directly adjacent, the combined radiant intensity of n LEDs can be approximated as $I \approx I_e \cdot n$ and the radiating surface as $A \approx n \cdot A_{LED}$. Then, the effective radiance L_R and the total irradiance E_{IR} can be calculated as

$$L_R = \frac{I}{A \cos \varepsilon_1} \cdot R(\lambda)\,, \tag{6.2}$$

$$E_{IR} = \frac{I}{d^2} \tag{6.3}$$

[1] The model is simplified in the “safe direction” by assuming a worst case scenario. In its default operation mode, the ring light is a pulsed light source with even higher permissible limits.

Table 6.1: Effective radiance and total irradiance of the ring lights most critical wavebands on the eye of an observer in a distance of $d = 0.2m$ compared to applicable limits according to [119].

		L_R $[W/(m^2 \cdot sr)]$	E_{IR} $[Wm^{-2}]$
Ring Light Rev. 1	Simulation	-	17.3
$\hat{\lambda}_1 = 935\,\mathrm{nm}$	Calculation (worst case)	13 590	25
Ring Light Rev. 2	Simulation	-	7.64
$\hat{\lambda}_3 = 1\,300\,\mathrm{nm}$	Calculation (worst case)	1 216	38
Single SPAI LED module $\hat{\lambda}_2 = 1\,300\,\mathrm{nm}$	Simulation	-	5.90
	Calculation	515	5.57
	Measurement	-	5.69
	Limit ($t \geq 1000s$)	$\approx 545\,000$	100

with $R(\lambda)$ being a correction factor according to directive 2006/25/EG [119], $d = 0.2\,\mathrm{m}$ being the distance of an observer according to [120] and $\varepsilon_1 = 0°$ being the emission angle towards the observers eye; see [45].

Table 6.1 shows simulated and calculated results for both ring light revisions, compared to the applicable limits. As expected, the total irradiance calculated using the simplified “worst case” model is much higher than the results from the simulation: this worst case model assumes that all emitted radiation can reach the eye of the observer, but in practice, the distances between the single LEDs on the ring lights are too large in order for an observer in a close distance of $d = 0.2\,\mathrm{m}$ to be hit by the emitted radiation from all of them. For an observer in a greater distance, the numerical calculation and simulation results will be more similar, but also significantly smaller as the the irradiance decreases with the squared distance.

In the context of the research project *safe person detection in working areas of industrial robots (SPAI)*, a much more powerful ring light designed by Sporrer *et al.* [23] for safeguarding applications has been evaluated using a certified measurement setup by the BG ETEM[1]. In this evaluation, the $\hat{\lambda}_3 = 1\,550\,\mathrm{nm}$ waveband was found to produce the highest total irradiance of $E_{IR} = 18.36\,\mathrm{W\,m^{-2}}$ on an experimentally determined “sweet spot” at the distance of $d = 0.2\,\mathrm{m}$. In the theoretical worst case scenario, the total irradiance of this waveband has been calculated as $\hat{E}_{IR} \approx 35.65\,\mathrm{W\,m^{-2}}$. Similar to the differences in numerical calculation and simulation results, this large discrepancy can be explained by the size of the ring light (0.4 m in diameter; see Appendix A) and the LEDs angle of view (60°) compared to the small measurement distance: the

[1] German employer’s liability insurance association for energy, textiles, electronics and media

LED modules that are not directly in front of the measurement point have only minor influence on the measured irradiance in practice.

A second result of this practical evaluation supports that both the simulation and the numerical calculation lead to plausible values in simpler scenarios: the calculated expected irradiance of a single LED module with a peak radiant intensity of $I_e \approx 0.22\,\mathrm{W\,sr^{-1}}$ achieved very similar irradiance values in the simulation, as well as in the actual measurements performed with this module; see Table 6.1.

In conclusion, both the simulated and the calculated worst case values are by far below the permissible limits. Thus, the ring light is not expected to cause any damage to the human eye, even if the observer stares right into it for a very long time. This leaves some headroom for further increases of the output power in future work.

6.6 Depth from Chromatic Aberration

Depth information of a scene, *e.g.*, a captured face of a subject that passes an eGate system, can help to detect spoofing attacks. As described in Section 3.2.3, 3D depth imaging has been proposed in prior work to address the vulnerability of face recognition systems to presentation attacks using printed pictures or images and video sequences shown on a mobile device. Besides providing a depth profile that allows to reject flat objects as obvious two-dimensional presentation attacks, depth information also allows to estimate an expected size of the face and to check the plausibility of extracted features, such as the eye to eye width. In the following, a method is described that allows to estimate rather coarse depth information from the acquired multispectral images without the need for an additional 3D sensor.

As described in Section 2.4, longitudinal chromatic aberration leads to a shift in the focal length of a lens depending on the wavelength of the light. Due to the wide spectral range covered by the *SkinCam* system, this effect has a very noticeable influence on the acquired multispectral images: an object can only be correctly focused in one single waveband image and will suffer from out of focus blur in all other wavebands. As shown in Figure 6.12 on the following page, the focus is shifted from the front to the back for longer wavelengths. The effect can only be reduced by closing the aperture to increase the depth of field [45]. However, it can be used as an advantage in order to estimate a depth map of the captured scene, as shown by Atif [121] and Trouvé *et al*. [122] in the context of RGB imaging.

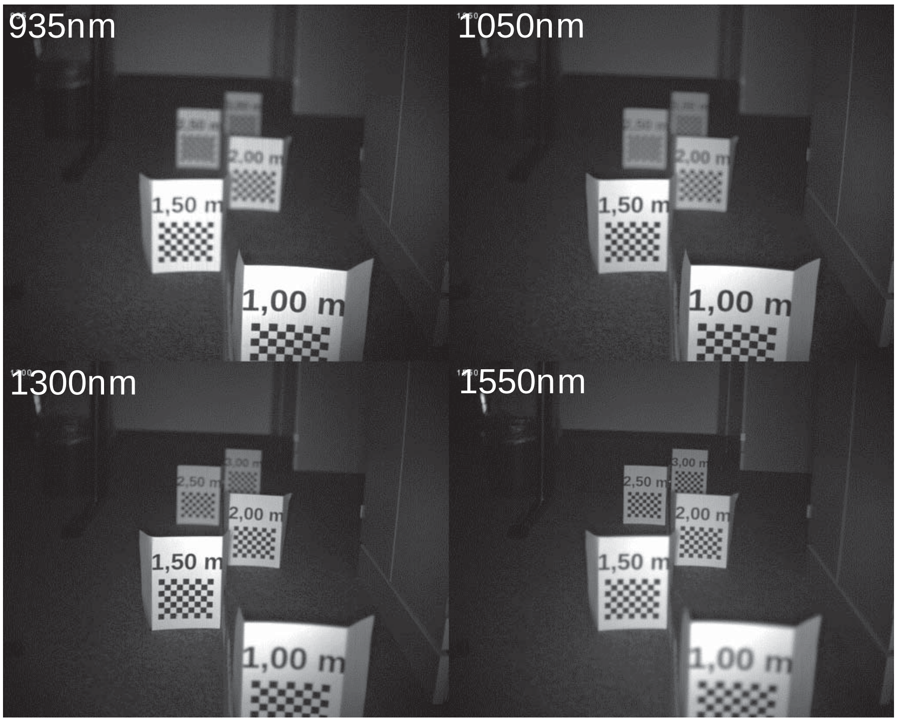


Figure 6.12: Focus shift due to chromatic aberration.

Estimating depth information based on the focus of an optical system is a known concept in the literature. If an object is correctly focused on the image plane of an optical system with image distance d_i, its distance d_o from the lens with focal length f can be determined using the thin lens formula [42]:

$$\frac{1}{d_i} - \frac{1}{d_o} = \frac{1}{f}. \tag{6.4}$$

In practice, however, finding the correct focus (*i.e.*, image distance) is a non-trivial problem [123]. As the image of incorrectly focused objects will be blurry, the correct focus setting can be found by analyzing the level of sharpness of the object's representation on the image plane while changing the image distance until the image is as sharp as possible, similar to the so-called contrast detection autofocus systems used in many cameras [124]. To automatically find the sharpest representation, a metric to measure the sharpness at a specific image area is necessary. Krotkov described and evaluated several different metrics or "criterion functions" for this purpose and found that the maximization of the gradient magnitude gives the most accurate and robust results [123]. The gradient magnitude G at a pixel coordinate (x, y) in image I is calculated using the Sobel operator with the convolution kernels S_x and S_y [46]:

$$G(x, y) = \sqrt{[S_x * I(x, y)]^2 + [S_y * I(x, y)]^2}, \tag{6.5}$$

$$S_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix}, \qquad S_y = \begin{bmatrix} +1 & +2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}. \tag{6.6}$$

Krotkov proposes to sum up the gradient magnitudes of all pixels within the analyzed image area that exceed a certain threshold T in order to find the correct focus setting with the highest value [123]:

$$\max \sum_x \sum_y G(x, y)^2 \qquad \text{with } G(x, y) \geq T. \tag{6.7}$$

By performing a sweep over the full focus range of a lens and determining the maximum gradient magnitude for each pixel, a depth map of the full image can be created [124]. Besides the need for a motor-driven lens that allows to adjust the focus electronically, performing a focus sweep takes a certain amount of time. An alternative to a full sweep has first been proposed by Pentland [124], who introduced the *focal gradient* as a source of depth information, which can be measured by either estimating the blur of (originally sharp) edges or by capturing the same scene twice with different aperture settings and estimating the change in the amount of blur due to increased depth of field.

Similarly, the different waveband images of a multispectral imaging system can be used instead of different aperture settings to estimate the focal gradient based on the change of edge sharpness [121, 122]. Here, sharpness is analyzed using the gradient magnitude as defined in Equation (6.6). To evaluate the focus shift of the *SkinCam* setup, the camera was mounted on an automated linear stage and a test pattern board was installed in a distance of 1.5 m in front of the linear stage. The aperture of the lens was fully opened and it was focused at a distance of 2 m for the lowest waveband. With this setup, the object distance between camera and test pattern was automatically adjusted in a range of 1.5 m to 3 m in increments of 10 mm. At each step, a multispectral image was captured and the sharpness of each waveband channel was analyzed. Figure 6.13 on the following page presents the resulting progression of the sharpness for both the 50 mm lens used in the context of this work and a 8 mm wide angle lens for comparison. While the influence of the chromatic aberration is very obvious for 50 mm lens with a pronounced shift of the maximum sharpness, it is much harder to assess for the 8 mm lens due to its a much larger depth of field.

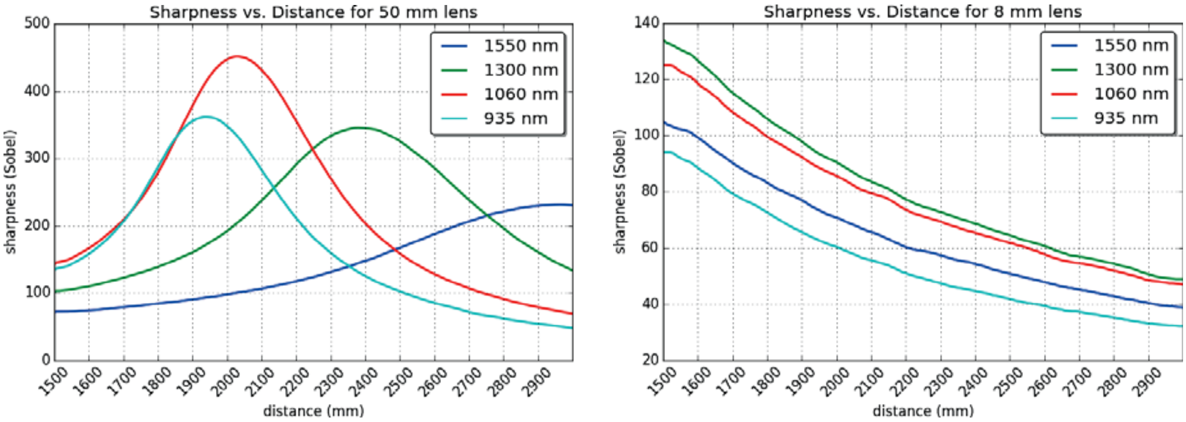


Figure 6.13: Progression of sharpness with increasing distance in the different wavebands for two different lenses (f=50mm and f=8mm) at wide aperture (f/1.4) [116].

Using the setup described above, a large amount of training data has been created using different kinds of target surfaces, including human faces. This data is used to train a model tree classifier that allows to predict the object distance based on the sharpness of object details; see Section 2.6.1. As this approach can only give an accurate estimate of sharpness if some kind of texture or structure is present in the analyzed image area, the resulting depth map will always be sparse and reliable depth information will only be available along edges of the input images. To remove outliers along these edges, a modified median filter kernel is used that discards all values below a certain threshold in order to smooth only the actual edges. Results of the depth estimation approach are presented in Section 7.3.

In order to create a dense depth map out of this sparse information, interpolation has to be applied in post processing. Approaches to interpolate dense depth maps from this sparse data have been described by [121], for example, but are not further discussed in the context of this work.

6.7 Summary

This chapter presents the system design and setup of the *SkinCam* camera system that implements the reference design of a skin detecting imaging system proposed in Chapter 3. The design consists of three building blocks: the camera system hardware, image processing and image analysis. The hardware setup combines a SWIR camera that is sensitive in a spectral range of 900 nm to 1700 nm with an active illumination module or *ring light* consisting of LEDs in four distinct wavebands around $\lambda_1 = 935\,\mathrm{nm}$, $\lambda_2 = 1060\,\mathrm{nm}$, $\lambda_3 = 1300\,\mathrm{nm}$ and $\lambda_4 = 1550\,\mathrm{nm}$ (revision 1), or $\lambda_1 = 1050\,\mathrm{nm}$, $\lambda_2 = 1200\,\mathrm{nm}$, $\lambda_3 = 1300\,\mathrm{nm}$ and $\lambda_4 = 1550\,\mathrm{nm}$ (revision 2), respectively. These wavebands have been chosen based on the analysis of spectral data from a large number of skin and material samples. The design of the illumination module and the distribution of the LEDs have been optimized according to the results of optical simulations. A microcontroller system is embedded in the ring light and controls the LEDs synchronized to the camera's exposure time. Image processing is performed on a connected computer and consists of a fixed pattern noise correction, intrinsic camera calibration, as well as a calibration of ring light homogeneity and "white balance". After this preprocessing, skin detection is applied on the multispectral images as described in Chapter 5.

In addition, an evaluation of eye safety according to applicable norms and standards is presented, as the ring light emits a significant amount of SWIR radiation. It is shown that the irradiance on the eyes of an observer is far below the permissible limits.

Finally, an approach to estimate depth from defocus in the captured images is described. Knowing the distance between the camera and a subject can, in practice, help to detect spoofing attacks based on plausibility of the sizes and proportions. The described approach is based on the analysis of changing edge sharpness due to focus shifts between the different wavebands, which is caused by longitudinal chromatic aberrations in the optical system and is shown to be very pronounced for lenses with a narrow depth of field.

Chapter 7

Concept Validation

In the following, the concept and reference design of a skin detecting camera system proposed in Chapter 3 is validated based on the system design and implementation described in Chapter 6. The results are presented in three separate sections: first, the system itself is evaluated with respect to calibration, influence of ambient light, operation range and distance estimation accuracy. Then, the skin detection method proposed in Chapter 5 is evaluated on the basis of a study with more than 150 participants. Finally, the performance of the anti-spoofing approach is evaluated on a dataset containing images of different masks and facial disguises.

Publications: A concept validation of an earlier development stage of the proposed system has already been presented in [21], including results on skin detection performance. Additional results on face verification performance were published in [22]. Detailed results of the motion compensation approaches are included in [24]. Results on the depth from chromatic aberration approach have already been presented by Velte [116].

7.1 System Evaluation

Figure 7.1 on the next page shows an example of the multispectral image cube acquired by the implemented *SkinCam* camera system with the first revision of the ring light after image preprocessing. The cube consists of four waveband images and a reference image which is used to compensate for ambient light. For comparison, a color image taken with a high quality RGB camera is given. Due to longitudinal chromatic aberrations of the camera's lens, it is impossible to have all waveband images perfectly

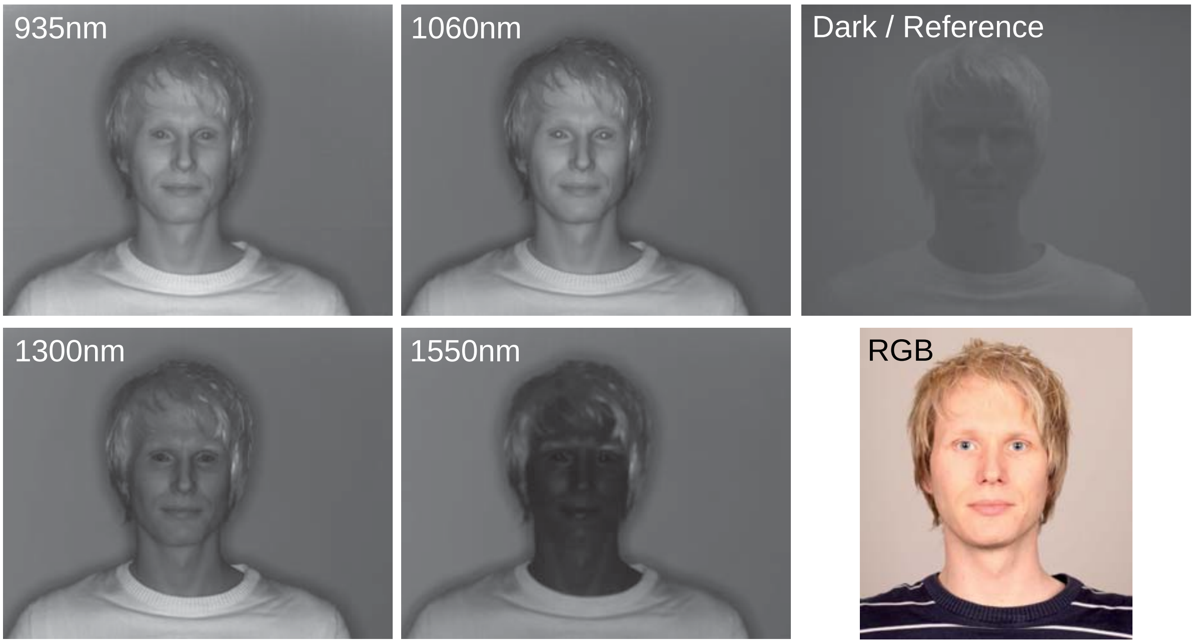


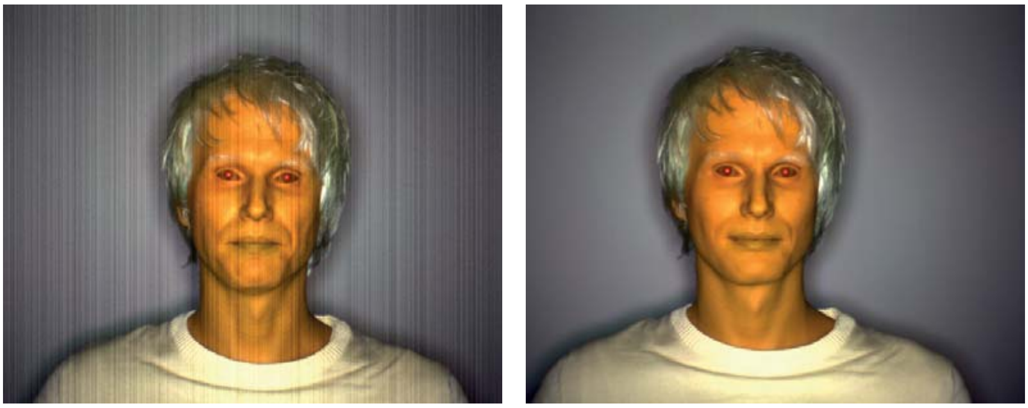
Figure 7.1: The multispectral image cube acquired by the SWIR camera system compared to an RGB color image.

focused at the same time; see Section 6.6. This effect can only be reduced by stopping down the lens to a smaller aperture. As the waveband image around 1 060 nm is best suited for face detection, all images are captured with this waveband image correctly focused while accepting a slight fall off in sharpness on the other waveband images.

7.1.1 Calibration Results

The first step in image processing is the correction of fixed pattern noise (FPN), as described in Section 6.3.1. In Figure 7.2 on the facing page, the effectiveness of the FPN correction method is demonstrated.

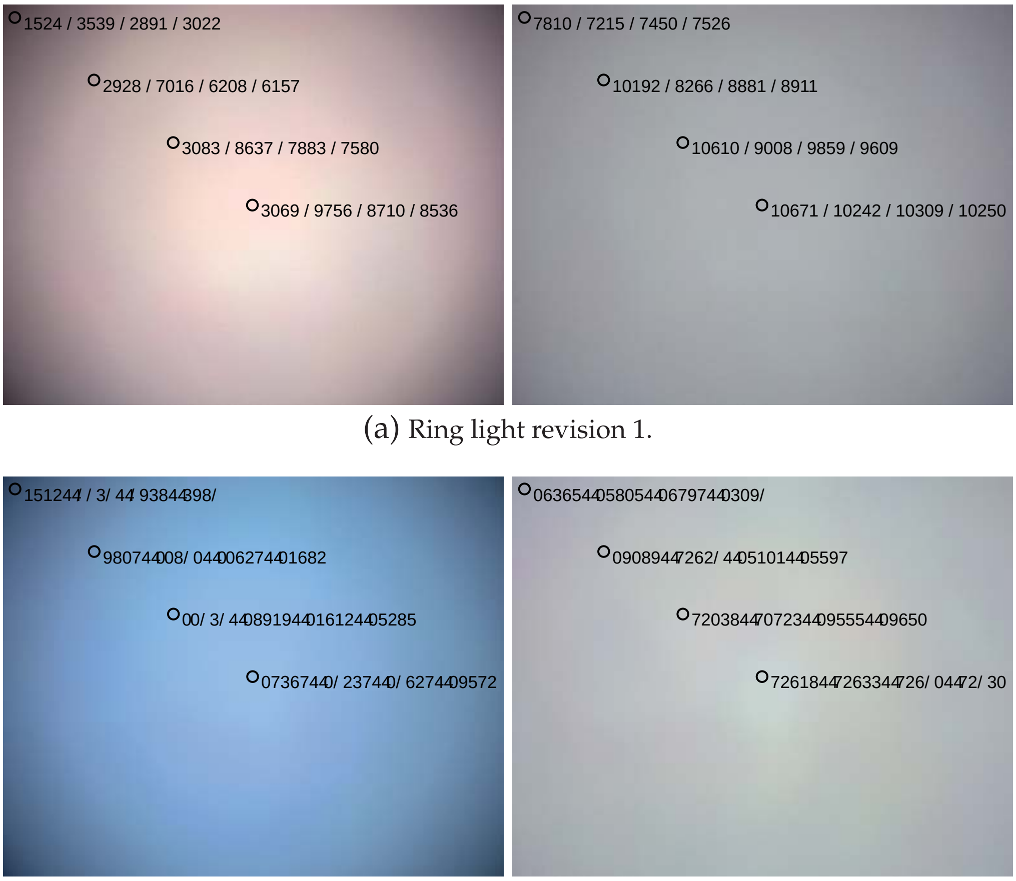
The evaluation of the illumination intensity and homogeneity of the ring light revealed that, despite coming from the same manufacturer and having similar packages, the different LED types have slightly different radiant patterns. However, this inhomogeneity, as well as different absolute intensities are compensated well by applying the generated calibration matrices and correction factors as described in Section 6.3.3. The calibration has been performed without ambient light at a distance of 1 m to a homogeneously white target plane and is tested at a distance of 2 m with bright ambient light. The results are shown in Figure 7.3 on the next page.



(a) Before FPN correction.

(b) After FPN correction.

Figure 7.2: Effectiveness of the fixed pattern noise (FPN) correction.



(a) Ring light revision 1.

(b) Ring light revision 2.

Figure 7.3: Ring light homogeneity before (left) and after (right) correction. Captured at a distance of 2m to a target plane with bright ambient light.

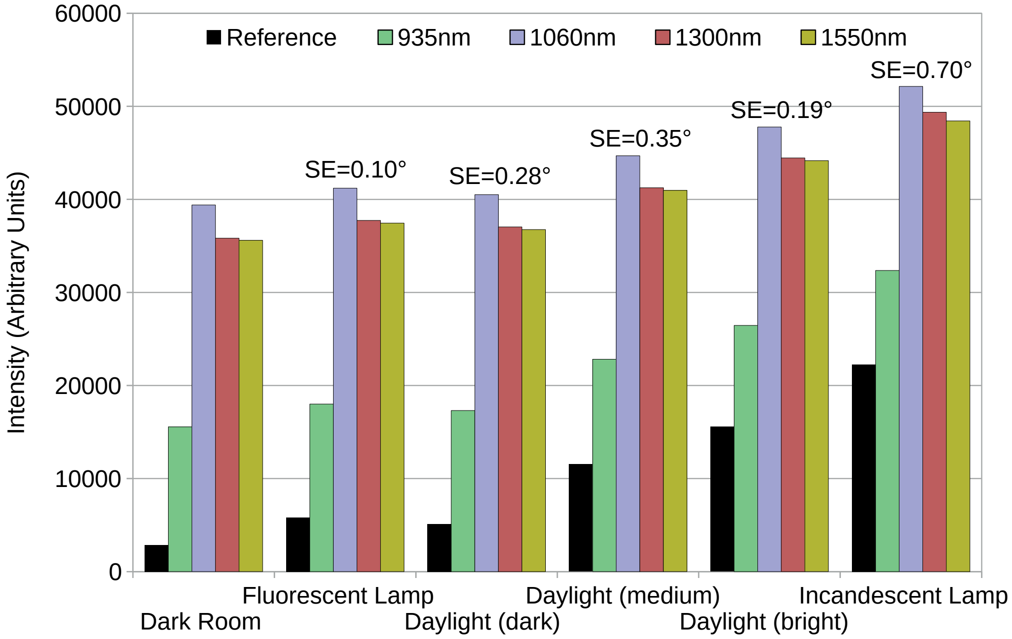


Figure 7.4: Remission intensities of the SWIR light pulses on a reference target and resulting spectral error (SE) with increasing ambient light.

7.1.2 Influence of Ambient Light

To evaluate the influence of ambient light on the camera system, a series of images of a reference target positioned in a distance of $\approx 1.5\,\mathrm{m}$ was captured with varying illumination conditions. The averaged illumination intensities measured on the reference target are shown in Figure 7.4. In this measurements, the ambient light is not yet subtracted from the signal pulses. Fluorescent lamps are barely visible for the short-wavelength infrared (SWIR) camera, while daylight and incandescent lamps increase the overall brightness significantly. Even without reaching saturation, the sensor shows some nonlinear behavior with increasing brightness levels: the signal strength decreases by up to $\approx 20\%$ between dark and bright ambient illumination. However, Figure 7.4 shows that the relative intensity differences between the wavebands stay almost the same with very low spectral error (SE), which is given in comparison to the spectral distribution measured without ambient light. As a result, the influence on the normalized differences and the classification results is only very small as long as the sensor is not saturated: the standard deviation of the normalized differences varies from 0.0022 to 0.0054. Saturation can be avoided easily by dynamically reducing the exposure time. In conclusion, ambient light can be widely neglected, but might reduce the maximum operation distance of the camera system.

7.1.3 Operation Range

The maximum operation distance of the camera system depends on several factors. The most important one is the radiated power of the ring light: with increasing distance to a target, the acquired remission intensities will decrease exponentially until they can no longer be distinguished from noise. In addition, with increasing ambient light the signal strength slightly decreases, while the shot noise (and, thus, the overall noise quantity) increases [125]. To evaluate the quality of the signal, both the noise level in terms of the standard deviation of the reference image and the average signal amplitudes for a target at different distances was measured in both dark and bright environments and the signal to noise ratio (SNR) was calculated according to Equation (2.2) on page 16. Results are presented in Table 7.1.

Table 7.1: Signal to noise ratio (SNR) of the ring light illumination for different target distances and ambient lighting conditions.

Dist.	**Amb.**	**Rev. 1 SNR [dB]**				**Rev. 2 SNR [dB]**			
[m]	**Light**	λ_1	λ_2	λ_3	λ_4	λ_1	λ_2	λ_3	λ_4
1.0	dark	25	30	30	29	31	32	31	32
	bright	24	28	28	28	29	30	29	30
1.5	dark	22	27	27	27	28	29	29	30
	bright	20	25	25	25	26	27	27	28
2.0	dark	20	25	25	25	26	27	27	28
	bright	18	23	23	23	24	25	25	26
2.5	dark	18	23	23	23	24	25	25	26
	bright	16	21	21	21	23	23	23	24
3.0	dark	17	22	22	22	23	24	24	25
	bright	15	20	20	19	21	22	22	23
3.5	dark	16	21	21	20	22	23	23	24
	bright	14	19	18	18	20	20	21	22
4.0	dark	15	20	20	19	21	22	22	23
	bright	13	18	17	17	19	19	20	21
4.5	dark	14	19	19	18	20	21	21	22
	bright	12	17	16	16	18	18	19	20
5.0	dark	13	18	18	18	19	20	21	21
	bright	11	16	16	16	17	18	18	19

In the experiments on skin detection, a $SNR \geq 20dB$ was enough to ensure reliable classification. For weaker signals, the classification performance started to decrease. As a result, in bright daylight conditions (overcast sky at noon) the first revision of the ring light can only be used at distances of up to 1.5 m without sacrificing the reliability of the first waveband, while the second revision can cover up to 3.5 m. In dark environments or indoor locations without incandescent lamps, both revisions achieve a higher SNR and the maximum operating distance increases by up to 1 m.

Besides the signal-to-noise ratio, the resolution and field of view of the camera system also put a limit on the operation range. For reliable face detection and recognition, current state-of-the-art algorithms require the image of a face to have an eye-to-eye resolution of ≥ 60 pixels [10]. For the current Goldeye cameras with a resolution of 636x508 pixels and the selected lens with an angle of view of $AOV \approx 18°$, this results in a maximum operation distance of $d_{max} \approx 2\,\text{m}$. To achieve a sufficiently high eye-to-eye resolution at longer distances, either a higher resolution camera (currently not available) or a lens with longer or variable focal length has to be used in practice.

7.2 Evaluation of Motion Compensation Performance

In the following, the field sequential waveband capturing (FSWC) motion compensation approaches described in Chapter 4 are evaluated using different datasets and quality measures. With a total number of 533 combinations of methods and preprocessing options, the total amount of results is very extensive. Here, only a representative selection is presented and the findings are summarized. Full results will be made available to the research community along with the *BRSU FSWC* datasets on the website of the Institute for Safety and Security Research (ISF) at the Bonn-Rhein-Sieg University of Applied Sciences (BRSU): `https://isf.h-brs.de`.

All of the described methods can be implemented using any kind of block matching or dense optical flow algorithm. In this work, GPU-accelerated implementations of Brox, TV-L1, Lucas-Kanade (LK), TGV2 and Huber-based (HL1, HGRAD, HQUADFIT) optical flow, as well as full search and fast approximate block matching algorithms from standard libraries (openCV 2.4.11 and FlowLib 3.0) are used. All Brox- and FlowLib-based optical flow algorithms have been applied twice, once with recommended (quality-oriented) parameters and once with parameters optimized for speed[1], which is denoted with a *. Optimal parameters for block matching have also been found experimentally[2]. Computational efficiency is measured on a standard desktop computer[3] using the *BRSU FSWC* VIS-D dataset.

[1] 3 instead of 10 inner/solver iterations and warps each

[2] search field parameter $p = 20$, block size $bs = 25$

[3] intel Core i7 4771 CPU, nVidia GTX 780, Ubuntu 14.04 64bit, GCC4.8, CUDA 7.5

7.2.1 Evaluation Setup

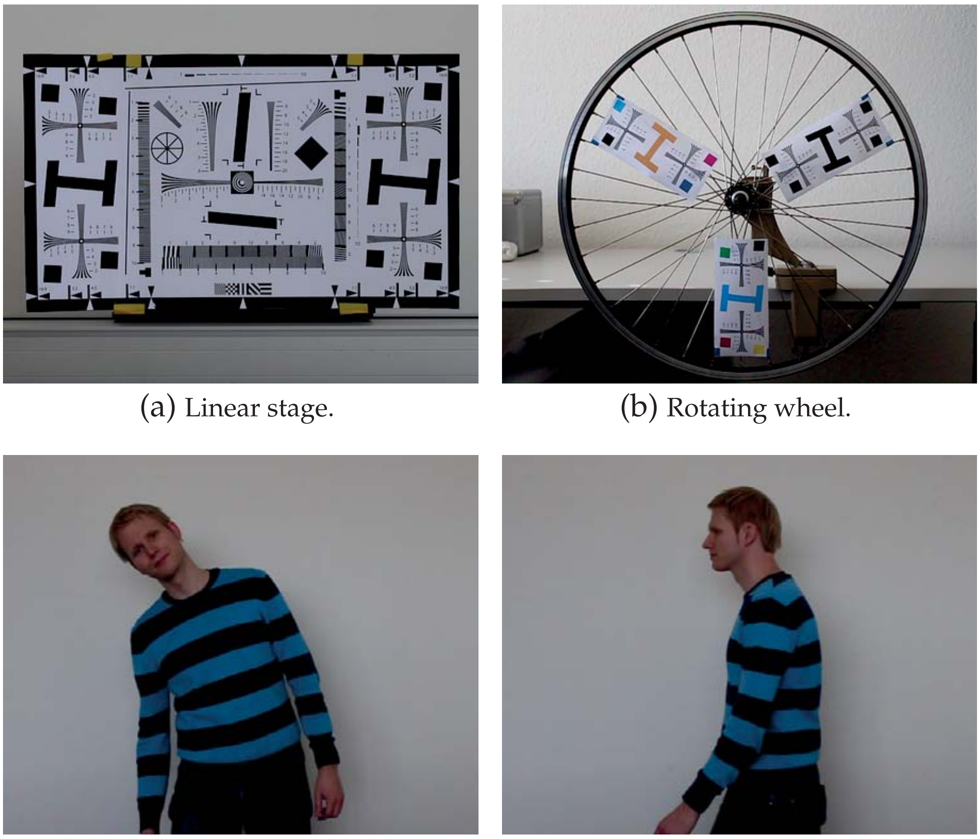
This section describes the datasets, creation of ground truth data and quality measures that are used to evaluate the described approaches for FSWC motion compensation. As there are no appropriate databases for comparing motion estimation and compensation techniques on actual FSWC imagery, a large variety of field-sequential multispectral SWIR and sRGB video sequences has been created in the context of this work. These FSWC datasets include different test scenarios with translational and rotational movement, and partially comprise ground truth data.

Datasets

Several different color/waveband-sequential image sequences are used as datasets for the evaluation of all described motion compensation approaches:

1. *Middlebury evaluation datasets:* all color sequences with 8 frames from the Middlebury optical flow accuracy and interpolation evaluation benchmark[1].
2. *BRSU FSWC datasets:* corresponding RGB and multispectral SWIR video sequences recorded using two cameras simultaneously. Some examples are shown in Figure 7.5 on the following page, a detailed overview is given in Appendix C.
    - *Linear stage:* video sequences of a test pattern board mounted on an automated linear stage, which performs translational movements with precise repeatability; see Figure 7.5 (a). As the movement of the linear stage is comparably slow, capturing it with high frame rates introduces only a minor amount of motion artifacts. Therefore, all sequences were captured twice, at 30 and 10 frames per second (FPS), from the same perspective.
    - *Rotating wheel:* video sequences of test pattern boards fixed on a rotating wheel; see Figure 7.5 (b). The rotation was initialized manually and images were captured at 30 FPS until the rotation stopped.
    - *Human movement:* video sequences showing upper-body shots of a person performing several different movements: walking by, moving sideways while looking forward, tilting the upper body, tilting and rotating the head, as well as waving with one and with both hands; see Figure 7.5 (c) and (d). All of these sequences were captured with 30 FPS.

[1] http://vision.middlebury.edu/flow/



(a) Linear stage.

(b) Rotating wheel.

(c) Head tilting.

(d) Walking.

Figure 7.5: Examples of the test scenarios included in the *BRSU FSWC* datasets.

The sequences from the BRSU FSWC datasets have been captured using both the active SWIR camera system developed in the context of this work and a high quality RGB camera with the same frame rate simultaneously. While the RGB camera uses a Bayer pattern to capture all channels simultaneously, the SWIR camera system is based on FSWC. For the creation of these datasets, it was configured to acquire three wavebands and the additional obligatory "dark" reference channel for each frame. The RGB images have been cropped and adjusted to match the field of view and image center of the SWIR images. Any negative effects caused by demosaicing of the RGB camera's Bayer pattern (see Section 2.1.3 on page 10) is accounted for by recording in high definition with 1920x1080 pixels and downsampling the images to the resolution of the SWIR camera's images, *i.e.* 636x508 pixels. All of these sequences will be made available to the research community.

Ground Truth

In order to perform an objective evaluation of the compensation quality, ground truth information is required. Unfortunately, it is not easily possible to get ground truth displacement vector fields for "real world" video sequences. Similarly, the amount of motion artifacts on FSWC imagery before and after motion compensation can be rated subjectively, but not objectively, unless there is a ground truth image of the exact same scene at the exact same time without any motion artifacts. Therefore, FSWC datasets comprising ground truth data have been created out of the RGB sequences, which were acquired using simultaneous color/waveband capturing: all RGB sequences were converted to waveband-sequential image sequences by (in turn) extracting only one of the channels from each frame, thus reducing the effective frame rate. These sequences are denoted as VIS. After applying motion compensation, all resulting multispectral image cubes can be compared to the respective original RGB frames, which serve as ground truth, using several quality measures; see Section 7.2.2.

To simulate an active camera system, a second set of waveband-sequential sequences has been prepared, denoted as VIS-D, by converting every fourth frame to a gray scale image with reduced brightness, which is used as "dark" reference, while the frames in between are used as spectral channels belonging to the same multispectral image cube, just as before. Again, after applying motion compensation on these sequences, they can be compared to the original frames.

However, for the video sequences acquired using the SWIR camera system, no similar ground truth data is available. Therefore, the corresponding RGB sequences are used in a cross-compensation approach. The SWIR camera system combines $n = 4$ channel images $C_{i,w}$ (*i.e.*, three wavebands plus dark reference) into one multispectral image cube M_i, while the RGB camera acquires all channels of each frame $\hat{M}_j$ simultaneously, *i.e.*, $C_{i,w}$ and $\hat{M}_{i\cdot 4+w}$ are acquired at the same time. The cross-compensation approach applies the optical flow $F_{(i,w)\to(i,0)}$ calculated for the SWIR channel $C_{i,w}$ to the RGB frame $\hat{M}_{i\cdot 4+w}$ and compares the result to $\hat{M}_{i\cdot 4}$, which corresponds to M_i and serves as ground truth.

To match the field of view of the RGB camera to the SWIR camera, the RGB imagery is shifted and cropped appropriately. However, the baseline between both cameras of $\approx 20cm$ induces a slightly different perspective and, thus, a mismatch in the motion fields. The application of extrinsic calibration in order to get a better matching has been tested, but did prove to be very unreliable due to missing depth information: a good calibration could only be found for one specific distance between camera and object. For other distances, the calibration introduced a significant additional error. Therefore, extrinsic calibration was discarded.

To estimate the mismatch of the motion fields, a second dataset was recorded in which the SWIR camera was replaced by another RGB camera. By applying the same cross-compensation procedure to this "double RGB" setup, a *baseline error* for the comparison of $\mathrm{IE}_{\mathrm{base}} \approx 2.7$ was found. As $\mathrm{IE}_{\mathrm{base}}$ is by far lower than the error of the best FSWC motion compensation method with IE ≈ 6.6, this cross-compensation approach appears to be valid within this range.

7.2.2 Quality Measures

The objective comparison of a compensated frame with the original frame is done using the following quality measures:

- *Interpolation error (IE)* is defined as the root mean square of the L2 norm of the vector of spectral channel differences between the interpolated, $\tilde{C}_{i,w}$, and ground truth images, $C_{i,w}$, of image cube M_i, analog to Baker *et al*. [36] and as used in the Middlebury evaluation:

$$\mathrm{IE}_i = \sqrt{\frac{1}{N}\sum_{x,y}\sum_{w}\left(\tilde{C}_{i,w}(x,y) - C_{i,w}(x,y)\right)^2}, \tag{7.1}$$

  where N is the number of pixels.

- *Peak signal to noise ratio (PSNR)* is based on the mean squared error (MSE) and commonly used to describe absolute differences in intensity values ($0 \leq \mathrm{Val} \leq \mathrm{maxVal}$) between of two images [126]:

$$\mathrm{MSE} = \frac{1}{wN}\sum_{x,y}\sum_{w}\left(\tilde{C}_{i,w}(x,y) - C_{i,w}(x,y)\right)^2 \tag{7.2}$$

$$\mathrm{PSNR} = 10 log_{10}\left(\frac{\mathrm{maxVal}^2}{MSE}\right). \tag{7.3}$$

- *Structural similarity index metric (SSIM)* describes the similarity of images based on structural information and is inspired by the human visual perception [126].

- *Spectral error (SE)* is defined as the root mean square of all pixel's spectral angular distance [32]. By representing spectra as vectors, the average angle between two vectors is calculated and amounts to 0° for similar and 90° for opposed spectra, independent from pixel intensities:

$$\mathrm{SE} = \sqrt{\frac{1}{N}\sum_{x,y}\left(cos^{-1}\left|\frac{p_1(x,y)\cdot p_2(x,y)}{||p_1(x,y)||\cdot||p_2(x,y)||}\right|\right)^2} \tag{7.4}$$

Here, N is the number of pixels, $p_1(x,y)\cdot p_2(x,y)$ is the dot product of the vectors describing the spectra of pixel (x,y) in image 1 and 2, and ||...|| means the square root of the sum of squares of all vector elements.

7.2.3 Performance Ranking

Analog to the *Middlebury* evaluation, all methods were ranked for each test sequence based on the described quality measures. As PSNR and IE are related and can be derived from each other, the ranking based on these measures will be identical. Therefore, only IE contributes to the overall ranking here. Table 7.3 on the following page shows the average ranks of the top-30 combinations of algorithms and approaches in the upper part (above the line). For comparison, results of the original algorithms without optimization for FSWC sequences and different inter-frame interpolation (IFI) optimizations have been added in the lower part of the table. Please note that this ranking does not allow to decide whether the compensation quality is sufficient or not. For this purpose, a second table presenting absolute error values of the same combinations (applied on the *BRSU FSWC* datasets only) is presented in Table 7.4 on page 107. In addition, Figure 7.6 on page 108 allows to easily compare the motion compensation performance to the computational efficiency of the different methods. Here, C2R methods have been left out in favor of better clarity, as the difference to the respective C2C methods is very small. Abbreviations are explained in Table 7.2.

Table 7.2: Abbreviations used in the context of the motion compensation evaluation.

Method based on		**Preprocessing**		**Algorithms**			
-B	all ch. bidir.	**N**	global norm.	**BM**	BM	**H2C**	HL2 Comp.
-U	all ch. unidir.	**L**	local norm.	**BM-m**	BM mod.	**HG**	HGRAD
-2	partial 2 ch.	**H**	histogram equ.	**FBM**	FastBM	**TC**	TGV2CENSUS
-1	partial 1 ch.	**C**	CLAHE	**Br**	Brox	**LK**	LK
-N	intens. norm.	**G**	gradient trans.	**TV**	TV-L1	**HQM**	HQUADFIT MIX
-TG	gradients			**HL**	HL1	**HQN**	HQUADFIT NCC
-TC	census			**FHL**	FAST HL1	**HQS**	HQUADFIT SAD
-C	correlation			**H1C**	HL1 Comp.	**PAC**	Pixelw. Art. Corr.

Table 7.3: Results of the top-30 (upper part) and selected additional approaches (lower part) on *Middlebury* and *BRSU FSWC* datasets; values are averaged ranks.

Approach			**Total**	**BRSU FSWC**			**Middlebury**		**Time**
Method	Preproc.	Alg.	**Avg.**	VIS	VIS-D	SWIR	VIS	VIS-D	**[s]**
C2C-TG		HG	54.00	33.64	44.00	41.11	118.73	32.52	0.235
C2R-TG		HG	58.55	33.64	44.00	41.11	118.73	55.27	0.237
C2C-TG	C	HG	61.90	43.97	41.06	30.93	150.72	42.82	0.237
C2C-TG	N+G	LK*	62.97	101.73	20.39	43.59	94.77	54.35	0.043
C2C-N	N	Br	63.22	66.30	30.21	108.26	86.18	25.17	0.322
C2R-TG	C	HG	66.20	43.97	41.06	30.93	150.72	64.33	0.240
C2C-TG	N+G	LK	66.55	83.00	31.15	44.78	121.87	51.97	0.069
C2R-TG	N+G	LK*	72.51	101.73	20.39	43.59	94.77	102.07	0.043
C2R-TG	N+G	LK	73.96	83.00	31.15	44.78	121.87	89.00	0.068
C2R-N	N	Br	74.22	66.30	47.24	128.41	86.18	42.97	0.320
C2C-TG	G	HG	75.30	55.27	47.79	37.37	165.82	70.27	0.230
C2C-TG	N+G	HQS	80.35	64.06	77.36	66.89	142.28	51.13	0.175
C2R-TG	G	HG	83.22	55.27	47.79	37.37	165.82	109.87	0.235
C2C-TG	C	HQM	84.48	125.45	54.15	29.00	172.07	41.72	0.156
C2C-TG	N	HG	85.43	52.61	32.21	170.93	122.60	48.80	0.267
C2R-TG	N+G	HQS	85.48	64.06	77.36	66.89	142.28	76.82	0.180
C2R-TG	C	HQM	86.49	125.45	54.15	29.00	172.07	51.78	0.158
C2R-TG	N	HG	89.74	52.61	32.21	170.93	122.60	70.35	0.269
C2C-TG	G	HQM	90.59	117.67	61.39	53.30	167.60	53.00	0.166
C2C-TG	L+G	Br	92.60	112.94	57.21	63.04	124.25	105.58	0.483
C2C-TG	N+G	Br	93.54	122.15	47.76	61.22	129.97	106.58	0.339
C2C-N	N	Br*	95.41	125.21	54.70	99.11	124.05	73.98	0.072
C2C-TG		HG*	99.53	77.73	83.70	92.93	146.23	97.07	0.044
C2C-N	C	Br	99.80	28.42	122.42	59.11	96.40	192.65	0.316
C2R-TG	G	HQM	100.12	117.67	61.39	53.30	167.60	100.65	0.171
C2C-TG	C	HG*	102.71	79.94	69.48	87.89	179.85	96.38	0.049
C2C-TG	N+G	HG	103.22	79.21	69.18	86.37	168.42	112.92	0.268
IFI-B		Br	103.54	118.91	162.76	55.70	68.27	112.08	0.659
IFI-B		HQS	105.49	124.88	175.79	46.93	59.72	120.15	0.302
C2R-N	N	Br*	107.63	125.21	65.21	128.30	124.05	95.37	0.067
IFI-U		Br	158.60	180.73	237.94	94.33	100.02	179.97	0.322
C2C		Br	193.28	61.58	250.30	285.56	89.67	279.30	0.315
C2C-TC		TC	200.58	215.21	169.52	249.93	206.30	161.95	0.256
C2C-TG	L+G	BME	211.44	182.15	119.30	265.81	286.13	203.80	47.801
IFI+C2R		Br+BM	221.05	188.03	301.79	157.96	176.62	280.85	11.944
IFI-2		Br	258.82		278.58			239.07	0.217
C2C-C		HQN	261.57	301.03	194.15	395.22	256.40	161.07	0.157
IFI-1		Br	267.30	288.61	376.42	195.00	189.80	286.68	0.116
C2C-TG	L+G	BM	287.65	276.09	206.21	391.56	316.27	248.12	11.861
PAC		PAC	390.56	371.09	419.42		358.67	413.07	0.014
C2C		FBM	424.26	378.70	481.64	436.67	354.15	470.17	0.195
C2C		LK	450.08	446.88	514.45	437.19	332.15	519.75	0.047
C2C		BM	470.91	449.52	508.61	486.70	421.37	488.35	11.496
C2C		TV	471.64	416.03	515.48	469.00	428.35	529.35	1.122
C2C		HL	476.52	427.45	528.67	464.67	442.15	519.67	0.125

Table 7.4: Average interpolation error of the top-30 (upper part) and selected additional approaches (lower part) on *BRSU human movement FSWC* datasets.

Approach			**Avg.**	**BRSU FSWC (VIS)**				**BRSU FSWC (SWIR)**			
Method	Preproc.	Alg.	**Rank**	IE	PSNR	SSIM	SE	IE	PSNR	SSIM	SE
2C-TG		HG	54.00	8.95	35.17	0.962	3.18	52.40	33.65	40.230	19.08
C2R-TG		HG	58.55	8.95	35.17	0.962	3.18	52.40	33.66	40.230	19.09
C2C-TG	C	HG	61.90	9.14	35.03	0.961	3.27	45.58	35.64	20.777	18.40
C2C-TG	N+G	LK*	62.97	9.48	34.72	0.958	3.57	22.58	36.98	22.879	8.84
C2C-N	N	Br	63.22	9.63	34.58	0.961	3.36	24.76	37.84	16.438	12.63
C2R-TG	C	HG	66.20	9.14	35.03	0.961	3.27	45.58	35.64	20.778	18.40
C2C-TG	N+G	LK	66.55	9.12	35.40	0.958	3.50	29.52	38.64	31.406	10.62
C2R-TG	N+G	LK*	72.51	9.48	34.72	0.958	3.57	22.58	37.00	22.879	8.84
C2R-TG	N+G	LK	73.96	9.12	35.40	0.958	3.50	29.51	38.66	31.406	10.62
C2R-N	N	Br	74.22	9.63	34.58	0.961	3.36	28.58	39.18	16.437	12.77
C2C-TG	G	HG	75.30	9.62	35.17	0.959	3.29	27.10	40.41	21.616	18.47
C2C-TG	N+G	HQS	80.35	9.54	34.94	0.958	3.29	11.76	36.60	16.066	19.80
C2R-TG	G	HG	83.22	9.62	35.17	0.959	3.29	27.10	40.42	21.616	18.48
C2C-TG	C	HQM	84.48	11.13	33.78	0.953	3.54	53.32	33.91	33.652	10.93
C2C-TG	N	HG	85.43	9.20	34.82	0.961	3.31	34.99	38.34	38.156	18.78
C2R-TG	N+G	HQS	85.48	9.54	34.94	0.958	3.29	11.74	36.61	16.067	19.80
C2R-TG	C	HQM	86.49	11.13	33.78	0.953	3.54	53.25	33.94	33.653	10.91
C2R-TG	N	HG	89.74	9.20	34.82	0.961	3.31	34.98	38.35	38.156	18.78
C2C-TG	G	HQM	90.59	10.98	34.15	0.952	3.58	43.34	36.24	42.659	10.79
C2C-TG	L+G	Br	92.60	10.25	34.24	0.956	3.48	21.56	41.11	20.165	18.59
C2C-TG	N+G	Br	93.54	10.38	34.13	0.957	3.58	19.55	40.21	11.610	11.29
C2C-N	N	Br*	95.41	10.45	33.75	0.956	3.76	30.89	41.87	26.773	19.42
C2C-TG		HG*	99.53	9.55	34.56	0.958	3.45	42.09	32.20	50.737	13.64
C2C-N	C	Br	99.80	8.86	35.43	0.963	3.13	15.41	35.66	16.364	12.38
C2R-TG	G	HQM	100.12	10.98	34.15	0.952	3.58	43.28	36.26	42.660	10.77
C2C-TG	C	HG*	102.71	9.66	34.52	0.957	3.47	31.97	34.57	33.193	15.32
C2C-TG	N+G	HG	103.22	9.93	34.73	0.957	3.42	22.06	38.23	20.268	17.86
IFI-B		Br	103.54	12.46	33.03	0.948	3.34	39.65	32.83	30.248	12.38
IFI-B		HQS	105.49	12.56	33.03	0.947	3.42	20.84	36.49	14.848	22.96
C2R-N	N	Br*	107.63	10.45	33.75	0.956	3.76	33.29	41.95	26.774	19.41
IFI-U		Br	158.60	13.44	32.32	0.945	3.76	17.52	31.14	9.345	20.16
C2C		Br	193.28	9.54	34.79	0.961	3.36	12.07	31.78	0.935	5.58
C2C-TC		TC	200.58	12.02	32.13	0.946	4.25	13.90	32.58	15.665	15.53
C2C-TG	L+G	BME	211.44	11.38	32.78	0.953	4.12	46.82	37.27	46.926	17.32
IFI+C2C-N	G	-	221.05	12.75	32.34	0.949	3.73	47.47	33.46	28.706	19.70
C2C-C		HQN	261.57	14.13	30.80	0.932	4.76	64.85	30.46	47.943	24.02
IFI-1		Br	267.30	17.57	29.54	0.934	4.54	28.69	28.45	13.289	23.61
C2C-TG	L+G	BM	287.65	13.13	31.30	0.939	4.72	60.21	32.88	52.289	16.66
PAC		PAC	390.56	20.48	27.92	0.913	5.63	-	-	-	-
Uncompensated			***412.43***	***25.60***	***25.70***	***0.910***	***6.28***	-	-	-	-
C2C		FBM	424.26	19.26	27.40	0.919	7.02	47.78	33.70	28.834	34.74
C2C		LK	450.08	27.35	24.18	0.891	9.16	73.15	23.45	26.433	26.75
C2C		BM	470.91	25.76	24.76	0.895	8.42	59.44	31.44	23.521	37.11
C2C		TV	471.64	25.45	24.93	0.909	8.57	62.29	17.11	0.707	26.37
C2C		HL	476.52	25.12	24.94	0.909	8.41	65.70	16.57	0.679	26.67

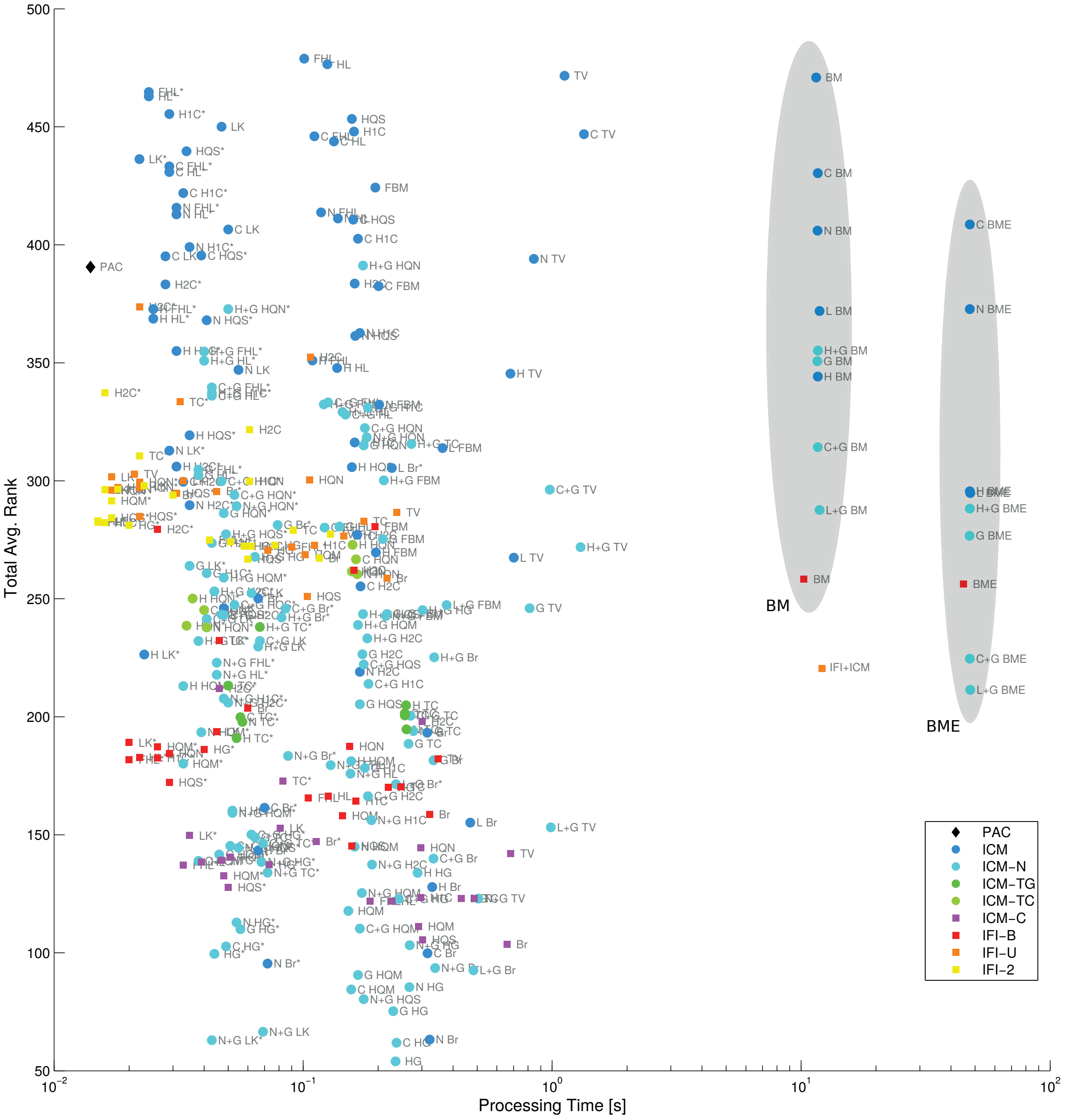


Figure 7.6: Scatter plot showing accuracy and processing time of the described FSWC motion compensation approaches.

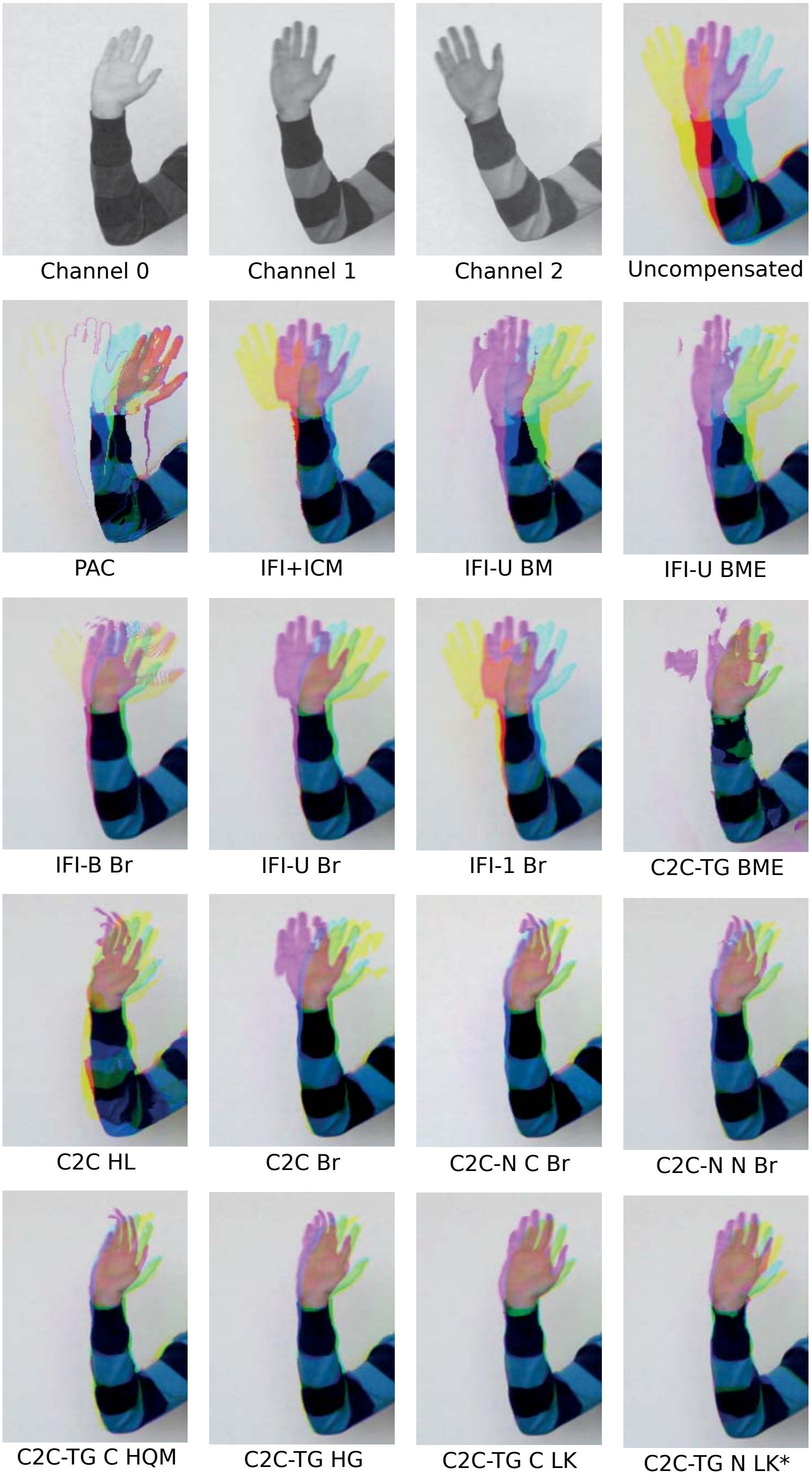
To illustrate the performance of the different approaches, the first frames from the *human movement* sequence “waving with one hand” and a selection of compensation results achieved by different approaches are shown in Figure 7.7 on the following page.

Varying the number of optical flow calculations in IFI

Table 7.3 on page 106 includes IFI results using all channels bi- (IFI-B) and unidirectional (IFI-U), first and last channel (IFI-2), as well as first channel only (IFI-1) interpolation based on the Brox algorithm. Brox was found to perform best for IFI, followed by Huber-based HQUADFIT SAD and (FAST) HL1, which run up to 3 times faster. For all algorithms, a reduction of the number of optical flow calculations decreases the processing time almost proportionally, while increasing the error at the same time, as shown in Table 7.5. In addition, Figure 7.8 on page 111 presents a diagram showing the relative error reduction on the y-axis versus the frame rate on the x-axis for the FAST HL1 algorithm with default and speed-optimized parameters, which lead to a much shorter processing time with only minor impact on the compensation quality. Depending on the performance requirements of a particular (real time) application at hand, this approach allows to find an acceptable trade-off between quality and speed.

Table 7.5: Processing time and compensation quality of IFI methods using different dense optical flow algorithms.

Method	Time [s]	IE	SSIM	SE
Uncompensated		32.38	0.832	10.11
TV-L1 IFI-B (6xOF)	0.682	19.89	0.883	6.71
TV-L1 IFI-U (3xOF)	0.348	20.80	0.880	7.34
TV-L1 IFI-2 (2xOF)	0.238	22.62	0.871	7.38
TV-L1 IFI-1 (1xOF)	0.129	27.24	0.852	8.55
Brox IFI-B (6xOF)	0.659	17.99	0.891	5.69
Brox IFI-U (3xOF)	0.322	19.28	0.886	6.57
Brox IFI-2 (2xOF)	0.217	21.18	0.878	6.67
Brox IFI-1 (1xOF)	0.116	26.49	0.856	8.41
FAST HL1 IFI-B (6xOF)	0.227	19.10	0.885	6.36
FAST HL1 IFI-U (3xOF)	0.126	19.78	0.882	6.93
FAST HL1 IFI-2 (2xOF)	0.090	21.39	0.874	6.99
FAST HL1 IFI-1 (1xOF)	0.051	26.67	0.852	8.48



Channel 0 Channel 1 Channel 2 Uncompensated

PAC IFI+ICM IFI-U BM IFI-U BME

IFI-B Br IFI-U Br IFI-1 Br C2C-TG BME

C2C HL C2C Br C2C-N C Br C2C-N N Br

C2C-TG C HQM C2C-TG HG C2C-TG C LK C2C-TG N LK*

Figure 7.7: Examples of a multispectral FSWC frame from the *BRSU* dataset “waving with one hand” before and after motion compensation.

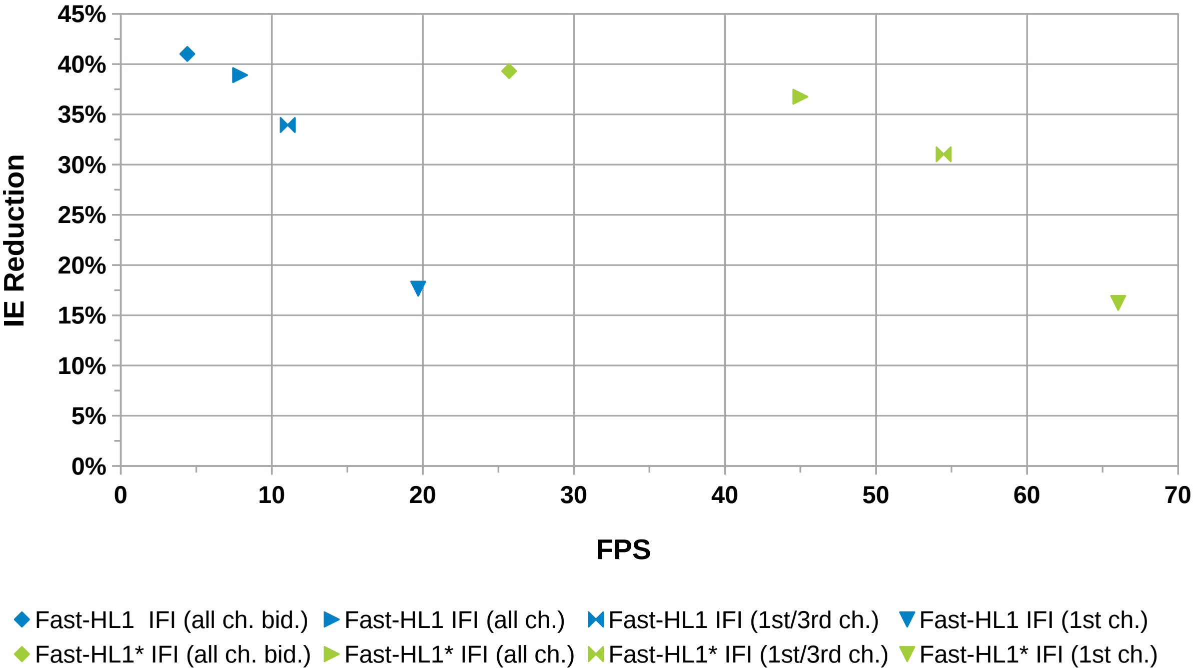


Figure 7.8: Relative interpolation error (IE) reduction versus frame rate of IFI methods using FAST HL1 (* = speed-optimized) applied on the BRSU FSWC datasets.

Differences of ICM C2C and C2R

From the results presented in Table 7.3 on page 106 it is obvious that the channel to channel (C2C) variants of the inter-channel matching (ICM) methods perform consistently better than the channel to reference (C2R) variants. This has been found to be true for all combinations of preprocessing approaches and optical flow algorithms throughout this evaluation and leads to the conclusion that smaller object displacements have a noticeable influence on the compensation quality.

Comparing IFI and ICM

On the *Middlebury* VIS sequences (created without "dark" reference), bidirectional IFI (IFI-B) can not be matched by any ICM approach. Compared to IFI-U, however, the best ICM methods perform slightly better at comparable processing times. On the *Middlebury* VIS-D (created with an additional "dark" reference channel) and all *BRSU FSWC* sequences, which have a higher amount of non-linear motion, *i.e.*, changes of motion direction and speed, several ICM methods perform significantly better than the best IFI approach.

Handling of Inconsistent Intensities

Without any preprocessing of the input images, only the Brox algorithm is capable of handling the inconsistent intensities with ICM methods to some degree. Combined with normalized intensities (ICM-N), Brox also performs very well in the overall ranking and is only surpassed by methods based on transformation to gradients (ICM-TG) using the HGRAD algorithm, as well as (surprisingly) the comparably simple Lucas-Kanade optical flow algorithm. Neither census transform (ICM-TC) nor NCC (ICM-C) can deliver similarly good results. When taking processing time into account as well, ICM-TG with speed-optimized HGRAD or the Lucas-Kanade optical flow after normalization and transformation to gradients delivers outstanding results.

Due to the computationally very expensive calculation of mutual information, see Section 4.3.3 on page 51, this approach has been evaluated in an independent second run using downscaled images from the *Middlebury* dataset with a resolution of only 320x240 pixels. Block size and search range have been reduced to $bs = 11$ and $p = 10$ accordingly. Table 7.6 presents the results of block matching results using mutual information as inverse cost function compared to the transformation to gradients. It is shown that MI-based compensation does not work well and even increases the amount of motion artifacts.

Table 7.6: Results of block matching based on mutual information.

Method	**IE**
Uncompensated	33.23
C2C BM	40.47
C2C-C Mut.Inf. BM	42.78
C2C-TG G BM	30.40

Combination of IFI and ICM

The combination of IFI and ICM (see Section 4.2.3 on page 49) is implemented using the Brox optical flow algorithm for the IFI step (all channels unidirectional, IFI-U) and the full search block matching for the ICM step. Here, the cost function for the block matching algorithm, which is based on the sum of absolute differences (SAD) [56], has been modified by adding additional costs depending on the deviation $\vec{d}$ from the initial displacement vector, weighted by a factor k_v:

$$\mathrm{SAD}_d(\Delta x, \Delta y) = \sum_{j=0}^{bs} \sum_{i=0}^{bs} \Big| f(x+i, y+j) - g(x+i+\Delta x, y+j+\Delta y) \Big| + k_v \cdot \Big| \vec{d}(x, y) \Big|. \tag{7.5}$$

As shown in Table 7.3 on page 106, this approach gives mediocre results on all datasets: in almost all cases, the ICM matching in the second step fails to improve on the initial IFI flow estimation, probably due to the generally low performance of the block matching algorithm. In future work, the approach should be implemented with a better optical flow algorithm for the ICM step, which might lead to better results.

Extended Cost Function for BM

As shown in both Table 7.3 on page 106 and Figure 7.6 on page 108, the extended cost function (ECF) does increase the performance of full search block matching at the cost of drastically increased processing times. However, neither original nor ECF-enhanced block matching get near the results of the dense optical flow approaches in terms of accuracy or speed.

Pixelwise Artifact Correction

The pixelwise artifact correction (PAC) performs comparably bad regarding the compensation quality, but the approach is computationally very effective and fast, although it is the only algorithm that does not rely on GPU acceleration; see Table 7.3 on page 106 and Figure 7.6 on page 108.

7.2.4 Influence of Low-Intensity Reference Channels

Active multispectral imaging systems capture an additional reference image without active illumination. The intensity of this “dark” reference represents the amount of ambient light that is captured by the camera and has to be subtracted from all channels. If its intensity is very low, the performance of ICM methods will decrease, as the matching of the spectral channels to the reference channel gets more difficult due to the stronger variations in intensity. To evaluate the influence of the illumination, one of the *human movement* sequences from the *BRSU FSWC* dataset has been used to create waveband-sequential sequences with increasingly bright reference channels

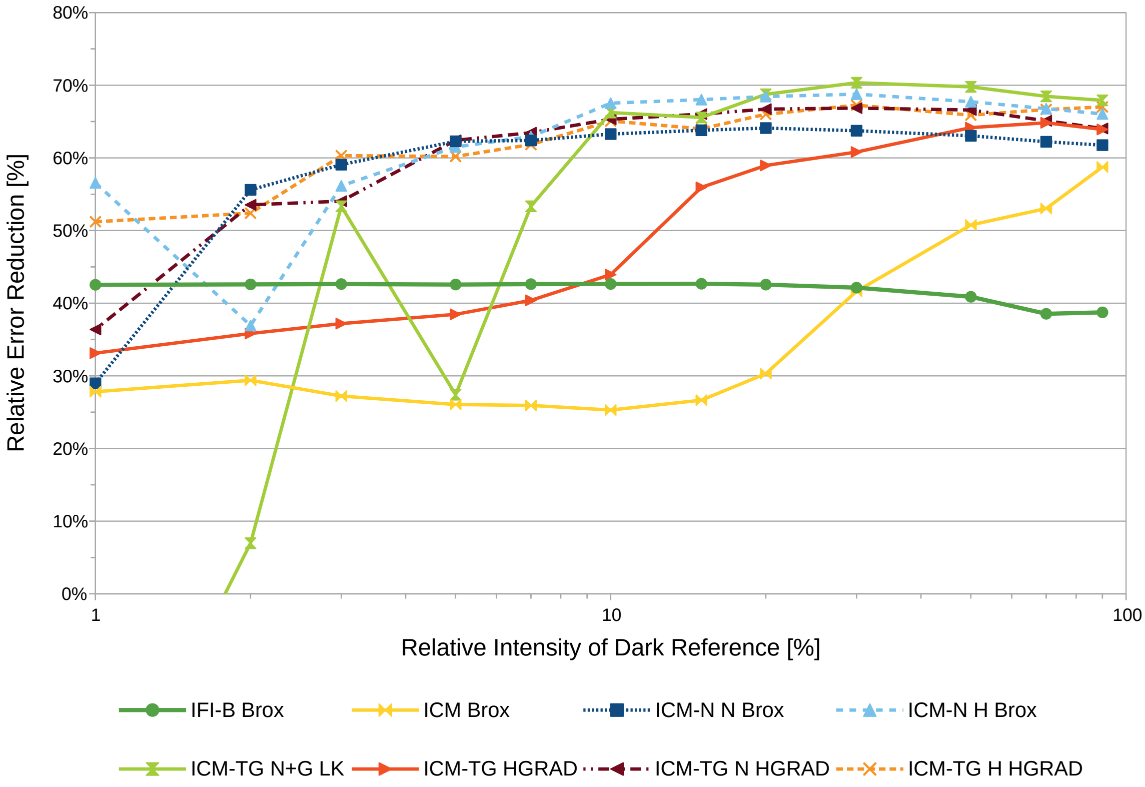


Figure 7.9: Relative IE reduction of IFI and ICM methods depending on the relative intensity of the “dark” reference.

between 0% and 90% of the original image’s intensity. Figure 7.9 presents the relative IE reduction achieved by compensating these sequences using IFI and ICM methods.

While the IFI method achieves an almost constant error reduction of ≈ 43%, ICM directly based on intensity does not work well for reference images with intensities below 30% without applying any sort of preprocessing. However, if the intensity is normalized by, *e.g.*, global linear normalization or histogram equalization prior to motion estimation, an intensity of ≥ 1% is enough for ICM to achieve better results than IFI. Transforming the intensity information into gradients also benefits strongly from a normalizing preprocessing step.

7.3 Distance Estimation Accuracy

To evaluate the error of the depth estimation approach described in Section 6.6, a variety of test data with ground truth information was acquired by capturing images of four different subjects, as well as several test patterns and different objects in the

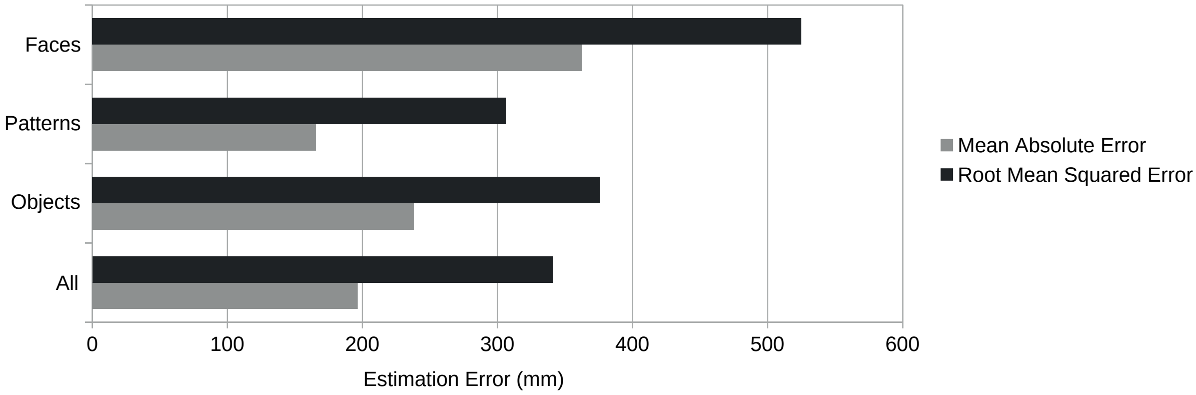
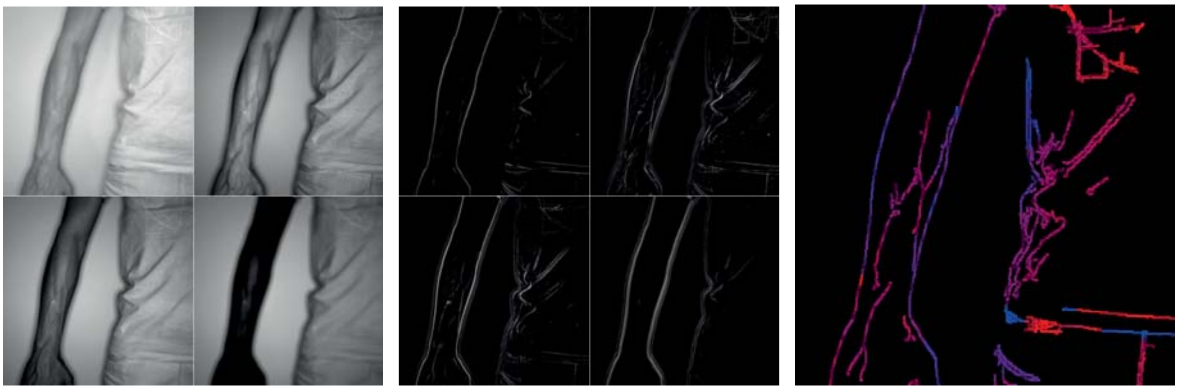


Figure 7.10: Distance estimation error using a model tree predictor.



(a) Input images. (b) Sobel-filtered images. (c) Depth estimation results.

Figure 7.11: Illustration of the distance estimation approach based on varying edge sharpness due to chromatic aberration. Distance is illustrated by color.

distance range of 1.5 m to 3 m using an automated linear stage setup. The f=50 mm lens was set to the widest aperture setting (f/1.4). As shown in Figure 7.10, the resulting estimation error of the used model tree predictor is comparably high for targets without sharp edges, such as human faces, while the distance to targets with pronounced edge detail, such as printed patterns, can be predicted with higher accuracy. Overall, a mean absolute error of 196 mm was achieved.

Figure 7.11 illustrates the results of the approach in practice: despite the application of median filtering, the depth estimation along the edges of the upper body and arm of the subject shows several discontinuities, depicted by a change in color from red to blue. Due to these rather imprecise depth estimation results with frequent outliers, this approach is currently not used for any further image analysis such as face anti-spoofing, as the amount of false rejections would have been too high. However, it was shown in related work by Atif [121] that this approach can be further improved and, thus, might be of use in future work.

7.4 Evaluation of Skin Detection Performance

7.4.1 Study Design

In order to evaluate the robustness of the skin detection approach and to gather training data for the classification algorithms, a study was designed in order to acquire images of a large number of persons with both the *SkinCam* camera system and a Canon EOS 50D RGB camera, as well as spectrometer data in the spectral range of 660*nm* to 1700*nm* using a TQ irSys 1.7 spectrometer. A subset of the resulting database, which includes all spectrometer data, as well as those images of participants who agreed to publication, is available to the research community; see Appendix C.

In the following, we present data from 152 participants, consisting of 79 women and 74 men. 42 of the 74 men had a beard or facial hair of some kind, while 63 of the 79 female participants have been wearing some sort of make-up or cosmetics. As this will be a common situation in real-life applications, testing the influence of facial hair and make-up was part of this study. Most of these datasets have been acquired in the laboratory at the BRSU. However, some additional datasets were acquired at the VISION 2014 trade fair and at the Institute for Occupational Safety and Health (IFA), both with a reduced measurement setup. Therefore, multispectral SWIR images were taken of all 152 persons, while RGB-color images in the visual (VIS) spectrum and spectrometer data have only been acquired for 137 and 101 persons, respectively. As most participants were students at the BRSU, the most common skin types are 2 and 3 and most of our participants were in their early twenties. The average age is ≈ 30. The respective frequency distributions are shown in tables 7.7 and 7.8.

Table 7.7: Age distribution of study participants.

Age	**< 18**	**18-24**	**25-34**	**35-44**	**45-54**	**≥ 55**
N	2	67	43	13	16	12

Table 7.8: Skin type distribution of study participants.

Skin Type	**1**	**2**	**3**	**4**	**5**	**6**
N	3	45	92	9	3	1

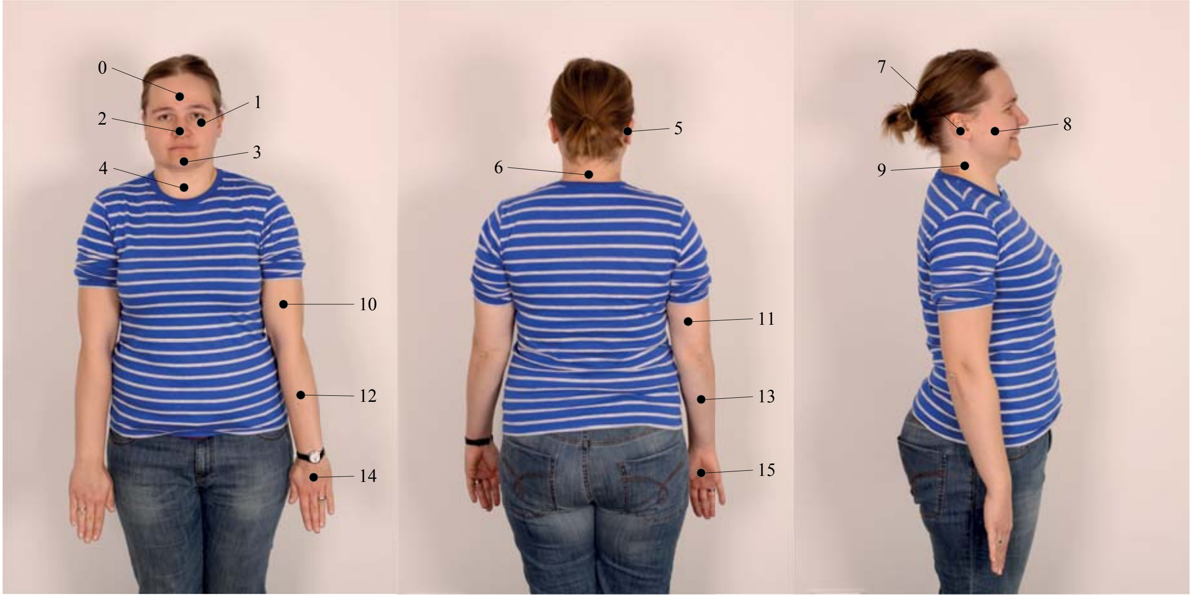


Figure 7.12: Spectrometer measurement points on face and arms.

For each subject, spectrometer data was acquired at 16 measuring points on face and arms, shown in Figure 7.12. These points have been chosen as they cover all skin regions that are typically expected in the field of view of a camera meant for face detection. With both the RGB camera and the *SkinCam* system, seven portrait pictures were taken for each subject: three frontal shots with different facial expressions, two shots from an angle of ±45° and two profile shots from an angle of ±90°. Subjects wearing glasses were asked to take them off for these shots and an additional image with glasses on was captured for comparison.

7.4.2 Data Analysis

In the following, both spectrometer and camera data is analyzed in detail in order to proof the validity of the skin detection approach. For this purpose, the datasets acquired during the study have been complemented with data acquired from a variety of material samples, including different plastics, textiles, metal and wood. This data does **not** contain any material samples designed with the intention to look similar to real skin, but includes rather typical samples of clothing, interior or workpieces that are found at offices as well as factories or workshops, for example. Therefore, the classification performance presented here is rather general and not representative for face verification applications. The classification performance for anti-spoofing is evaluated in greater detail in Section 7.5.

Spectrometer Data

For this evaluation, spectrometer data from only 12 of the 16 measuring points is used: these include 0, 4, 6, 8, 9, 10, 11, 12, 13, 14 and 15. The remaining measuring points 1, 2, 5 and 7 have been left out as it was found that the amount of faulty measurements was very high at these points, probably due to incorrect positioning of the measuring probe. In total, 1 111 skin samples have been combined with 335 samples of different materials. The spectrometer data was convoluted with the remission spectra of LEDs in the respective wavebands of both ring light revisions in order to simulate the expected spectral signatures $\vec{s}$ of the camera system as described in Section 6.2.3.

In a first step, the normalized differences $d(g_a, g_b)$ between all $n = 4$ wavebands of the spectral signatures $\vec{s}$ with $1 \leq a < n - 1$ and $a < b < n$ are calculated for all samples and a principle component analysis is applied on the dataset. In the following, only the results for the waveband set of the first ring light, R_1, are presented. However, classification results are identical for both sets of wavebands ($R_1 = \{935\,\text{nm}, 1\,060\,\text{nm}, 1\,300\,\text{nm}, 1\,550\,\text{nm}\}$, $R_2 = \{1\,050\,\text{nm}, 1\,200\,\text{nm}, 1\,300\,\text{nm}, 1\,550\,\text{nm}\}$). Figure 7.13 on the next page presents a plot of the two main components, which already separate most of the samples. Using difference filters by specifying minimum and maximum thresholds for each normalized difference, all skin samples can be separated perfectly from all material samples with a precision of 1.0.

Camera Data

To analyze the data acquired with the camera system, the spectral signatures of skin and a variety of other materials similar to those included in the spectrometer dataset have been extracted from the images taken during the study with the help of a software tool as described in Section 5.2.2. Pixels showing skin are stored as positive examples, “non-skin” pixels as negative examples. Again, a principle component analysis was applied on this dataset in a first step. The two main components are illustrated in Figure 7.14 on the facing page: here, no perfect separation of both classes is possible, as a few material samples overlap with the area of the skin samples. Using the thresholding filter on normalized differences, almost all samples can be separated from each other, as shown in Table 7.9 on page 120. An even better result is achieved by using a support vector machine (SVM) as classifier, which is evaluated on this dataset with 10-fold cross validation and leaves only one false positive (FP) result. In conclusion, an almost perfect classification of skin and material samples is possible with the proposed camera system.

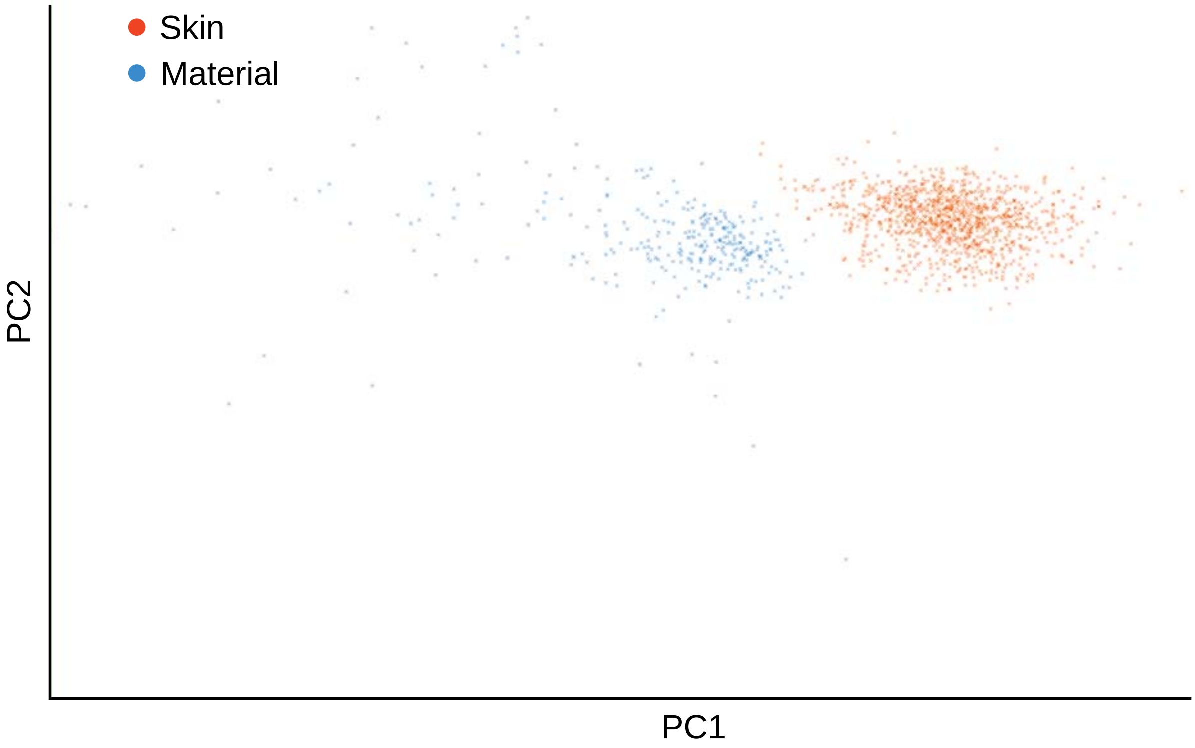


Figure 7.13: Plot of the spectrometer data mapped by its two main components.

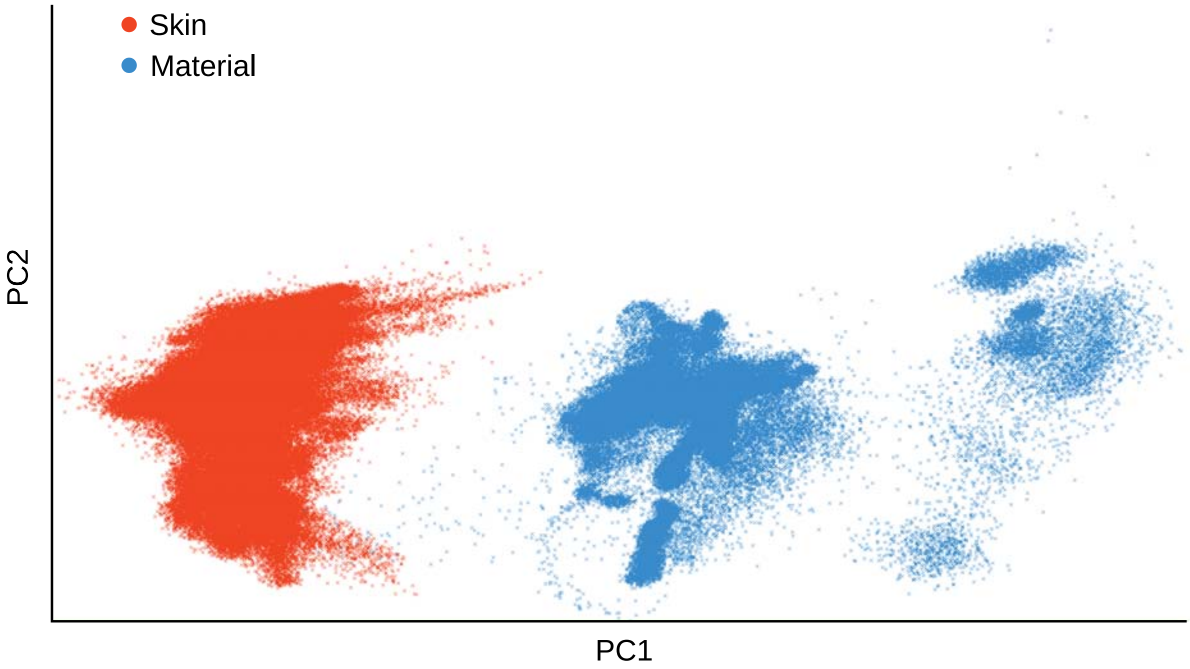


Figure 7.14: Plot of the spectral data extracted from individual pixels of the camera dataset mapped by its two main components.

Table 7.9: Confusion matrix and classification results for the spectral data extracted from individual pixels of the camera dataset.

(a) difference filter classifier

		Predicted Class	
		Skin	**Mat.**
Actual	**Skin**	158706	0
Class	**Mat.**	8	96246
	Precision	0.99995	
	Accuracy	0.99997	

(b) SVM classifier

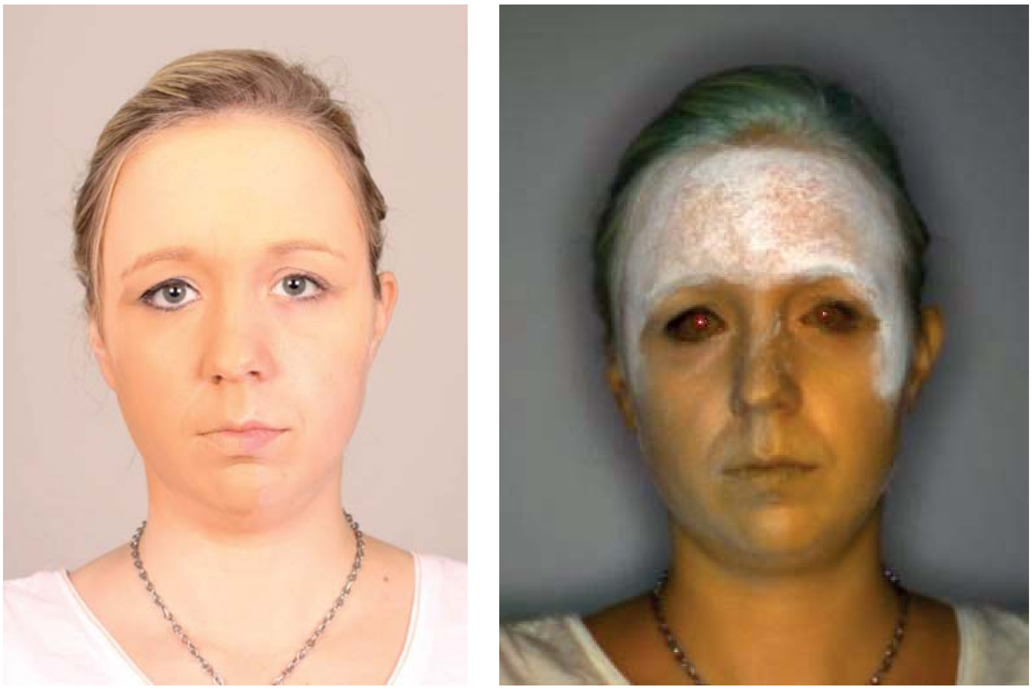
		Predicted Class	
		Skin	**Mat.**
Actual	**Skin**	158706	0
Class	**Mat.**	1	96253
	Precision	0.99999	
	Accuracy	0.99999	

Influence of Make-Up

In the context of the study, no significant influence of make-up on the skin classification results was found. However, heavy theater make-up or multiple layers of powder could potentially be a problem to skin detection. To evaluate this in detail, two subjects were asked to use very large amounts of make-up and powder and additional images were acquired. It was found that several layers of powder reflect a larger amount of SWIR radiation and appear very bright in the SWIR images, while eyeliner and eye shadow (both black and white) are more absorbing than skin in the SWIR spectral range. An example is shown in Figure 7.15 on the facing page. Both types of make-up might lead to false negative classifications and could, in theory, reduce the availability of the face verification method due to false rejections. In practice, however, this problem seems rather uncritical due to the large amounts of make-up that are necessary to “hide” the skin completely.

7.5 Evaluation of Face Anti-Spoofing

To analyze the anti-spoofing and face verification performance of the presented camera system, a database of images showing various spoofing attacks has been created and is used in conjunction with the face images acquired during the study; see Section 7.4.1. In the first step, the performance of the different classifiers for the per-pixel material classification is evaluated on the combined databases. Then, the usability and quality of the acquired SWIR images for face detection and recognition is tested and finally, the spoof detection performance is evaluated for the two attack scenarios “counterfeiting” and “disguise”, which have been introduced in Section 3.2.3.



(a) VIS (RGB color) (b) SWIR (false color; 1060, 1300 and 1550nm)

Figure 7.15: Example of a subject wearing several layers of make up and powder.

As described in Section 5.3, the *SkinCam* system was designed without a specific face recognition software in mind. Scientific open source solutions can be used as well as commercial off the shelf solutions. In this evaluation, FaceVACS[1] is used in version 8.9.

7.5.1 Dataset and Test Design

For training and test of the classifiers and the spoof detection performance, a variety of material that can be used to create spoofing attacks as well as different commercially available masks and facial disguises have been acquired, including heavy make-up and (fake) facial hair. In addition, several photo-fakes and masks, which mimic the face of one of the test subjects, were manufactured in the context of the spoof detection at biometric face recognition systems (FeGeb) research project conducted together with the German Federal Office for Information Security[2]. Different materials have been used for these masks, including special silicon mixtures, plastics, hard resin,

[1] Cognitec Systems GmbH, Dresden, Germany (http://www.cognitec.com)

[2] Bundesamt für Sicherheit in der Informationstechnik (BSI)

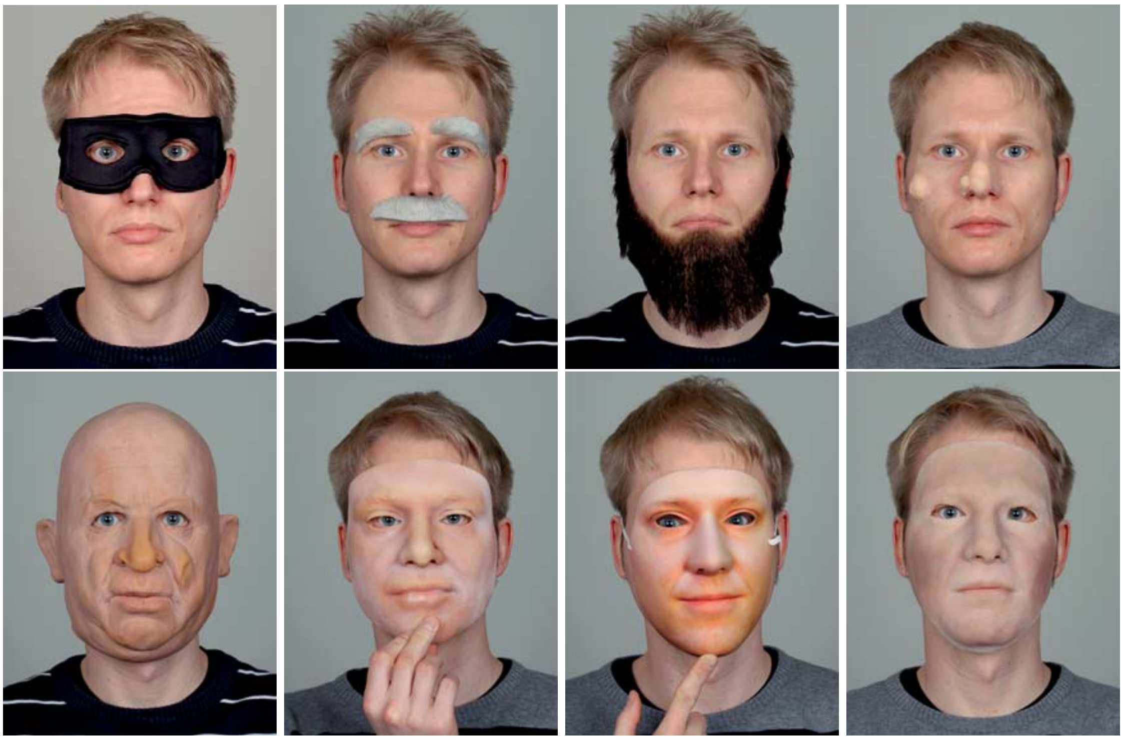


Figure 7.16: Examples of evaluated spoofing attacks.

textiles and paper. Make-up and paint have been applied to the masks to make them more realistic. A database of multispectral SWIR and RGB color images of these spoofing attacks has been created and is provided to the research community to further promote anti-spoofing research. Figure 7.16 shows a selection of the considered spoofing attacks.

Normalized spectral signatures have been extracted from all skin and material samples of both the face and the spoof database and were split up in distinct datasets for training and testing. To proof the universal validity of the classifiers, the "skin" samples used for training have been extracted from images of different persons than those used for the test dataset. Due to the limited number of masks and facial disguises, the same level of separation can not be achieved for the set of spoofing attacks: here, different images of the same spoofs have been used to extract samples for either test or training dataset. Both datasets contain roughly the same number of "skin" and "non-skin" samples. As light make-up, facial cream or tattoos should not be rejected as a spoofing attack per se, no such samples have been used for training, but are included in the test dataset. Using the Weka machine learning environment, support vector machine, binary decision tree and random forest classifiers have been trained and tested. Optimal parameters for the classifiers were experimentally found by testing the resulting models using cross-validation.

7.5.2 Classification Accuracy

To evaluate the overall classification accuracy of the different classifiers, all learned classification models have been applied on the same test dataset. The individual results in the form of confusion matrices are shown in Table 7.10.

Table 7.10: Confusion matrices and classification results of the spectral data extracted from individual pixels of the spoofing attack dataset.

(a) difference filter classifier

		Predicted Class	
		Skin	**Mat.**
Actual	**Skin**	320322	65
Class	**Mat.**	60975	265909
	Precision	0.9998	
	Accuracy	0.9057	

(b) SVM classifier

		Predicted Class	
		Skin	**Mat.**
Actual	**Skin**	320008	378
Class	**Mat.**	5319	321562
	Precision	0.9988	
	Accuracy	0.9912	

(c) binary decision tree (J48)

		Predicted Class	
		Skin	**Mat.**
Actual	**Skin**	319437	949
Class	**Mat.**	6726	320155
	Precision	0.9794	
	Accuracy	0.9881	

(d) random forest

		Predicted Class	
		Skin	**Mat.**
Actual	**Skin**	319518	868
Class	**Mat.**	5246	321635
	Precision	0.9796	
	Accuracy	0.9906	

Due to the way the normalized difference filter classifier is trained, the amount of false negative results is very low with this classifier. However, a large amount of material samples is falsely classified as "skin", which potentially is a thread for the safety of the anti-spoofing method. The best results in terms of both precision and accuracy are achieved by the SVM classifier. Compared to the other classifiers, the drawback of the SVM classifier is its much higher processing time: classifying the test set with the SVM takes about 20 times longer than with the decision tree. By combining the difference filter classifier sequentially with the SVM classifier, the amount of false positives can be reduced only slightly to 5274, while the processing time is reduced significantly to approx. 50% for the given test set and even further for datasets including less skin samples.

Both binary decision tree and random forest can be evaluated in less than one second on the test set. While the results of the simple binary decision tree are noticeably worse compared to the SVM, the random forest achieves the lowest number of false positives on the test set at the cost of slightly higher false negatives.

In a more detailed analysis, the different spoof samples have been classified individually. Results of the SVM classifier are included in Table 7.12 on page 126 and Table 7.14 on page 128. All machine-learning-based classifiers are able to distinguish most of these material samples from skin perfectly. However, artificial blood applied on a fake scar was found to be particularly hard to distinguish from skin, probably due to its high water content. Fortunately, this material is difficult to be applied for spoofing attacks due to its liquid character.

7.5.3 False Rejection Rate

In the field of face recognition, the rejection of a valid face is called a false rejection, similar to a false negative classification in machine learning. To evaluate the false rejection rate (FRR) of the anti-spoofing methods *masking (A)* and *ROI matching (B)* that have been proposed in Section 5.3, both methods are tested on the full dataset of authentic (not disguised) faces. All 137 subjects for whom VIS (RGB-color) images have been captured are enrolled in FaceVACS using three frontal images with varying facial expressions for each subject. With a few exceptions due to faulty data there are also three multispectral SWIR face images available for each subject, resulting in a total number of 404 SWIR images. For this test, only the 1 060 nm waveband has been used, as it was found to be best suited for face recognition. As a cooperative user scenario is expected, the influence of glasses has not been tested and only images without glasses are used for training and testing. Results are shown in Tab. 7.11.

If the correct subject is predicted with a higher probability than any other subject, this is denoted as a rank-1 identification. With a rank-1 identification rate of 100%, the recognition performance surpasses that presented in prior work. However, for

Table 7.11: False rejection rate and face verification performance of both methods using FaceVACS (trained on VIS images, queried with SWIR images).

404 images in total	**(A) Masking**	**(B) ROI**
Rank-1 Identification Rate	100 %	100 %
Above Verification Threshold	95.79 %	95.05 %
False Rejection Rate (FRR)	4.21 %	4.95 %

some images, the *matching score* is slightly below FaceVACS' internal threshold for a successful verification result. Surprisingly, the matching score is slightly higher for the masking method, which removes or "blacks out" non-skin pixels in the image, while the ROI method uses the unmasked image for recognition. This leads to the slightly lower false rejection rate of the masking method. Please note that the ROI method allows to use VIS images as input for the face recognition system as well. Additional experiments have shown that this reduces the false recognition rate significantly, especially under good lighting conditions, as the error of matching VIS and NIR images is removed. As described in Section 5.3.2, the ROI template and its acceptance threshold has been designed to accept all of these faces. Therefore, no face images were falsely rejected due to an incorrect spoofing detection. Furthermore, neither facial hair nor make-up had any noticeable influence on the results.

7.5.4 Spoof Detection and False Acceptance Rate

To evaluate the anti-spoofing performance of both methods, two attack scenarios are considered: disguise of the own identity and counterfeiting of a foreign identity.

Counterfeiting Scenario

In this scenario, an attacker tries to imitate the identity of a specific person to attack face recognition systems working in the *face verification* mode, for example in order to pass an automated border control gate using a fake passport. Here, a *false acceptance* occurs if the attacker is falsely verified as the person he claims to be using a spoofing attack, without the attack being detected by anti-spoofing methods. Table 7.12 on the next page presents a list of spoofing attacks from the created dataset that have been designed to counterfeit another person's face.

While the 2D attacks can be produced very easily by capturing an image of a face with a camera, the production of 3D masks is more complicated. Here, three approaches have been tested with the same subject. The first and rather traditional approach was to apply plaster on the subjects face to create a cast, which resembles a "negative impression" of the face. Then, two "positive" masks have been created by filling the cast with silicon and pressing it on the attackers face until the silicon hardened. Obviously, this procedure can not easily be performed without the original subject noticing it. The second approach was to take several images of the face from different perspectives in order to create a rough 3D face model, which was used as a basis for a 3D print with colored resin. Finally, the third approach was the use of

Table 7.12: Evaluated spoofs for the counterfeiting scenario with false positive rate (FPR) of the pixel-level classifier, number of false acceptances and total false acceptance rate (FAR) using FaceVACS compared to anti-spoofing methods (A) and (B). 2D attacks include prints and images shown on mobile devices. ** = with makeup.*

Description / no. of images		**Classifier FPR**	**FaceVACS**	**(A) Masking**	**(B) ROI**
Full 2D attacks	9	0.0	9	0	0
Partial 2D attacks	12	0.0	12	0	0
Full mask 1, silicon	3	0.0	0	0	0
Full mask 2, silicon *	3	0.0	0	0	0
Full mask 3, silicon *	3	0.0	0	0	0
Full mask 4, plastic	3	0.0	0	0	0
Full mask 5, hard resin	3	0.0	3	0	0
Full mask 6, hard resin	3	0.0	3	0	0
Sum / FAR$_{cf}$	**39**	**0.0**	**69.2%**	**0.0%**	**0.0%**

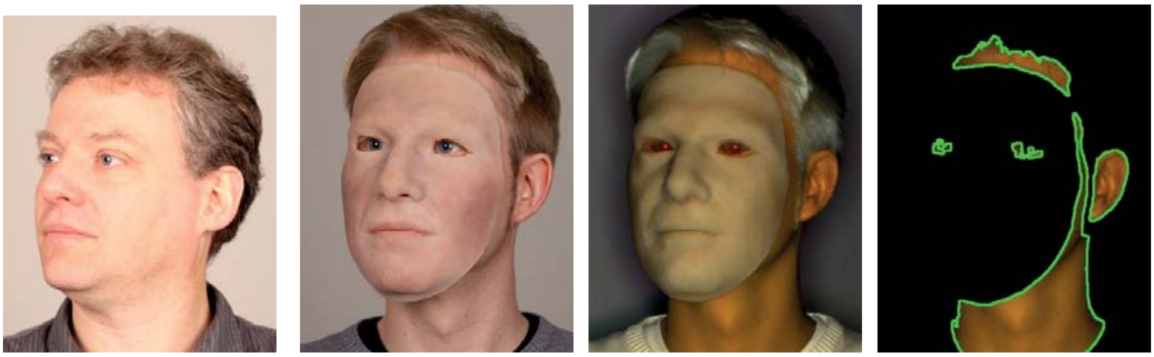
a sophisticated 3D scanner to achieve a detailed 3D model of the face. From this model, two versions were printed on another 3D printer: one “positive” mask, which was manually colored in a final step, and one “negative” mask, which was created in software and served as a cast to produce another silicon mask.

For each spoof, multiple images have been captured with three different attackers and (if meaningful) variations of the attack. Without an additional anti-spoofing method, all 2D attacks and the 3D-printed hard resin masks achieve scores similar to or even higher than real faces using FaceVACS. The quality of the plastic and silicon masks was not high enough for them to exceed the verification threshold, although they missed it by just a few percent, probably due to the manual coloring. Therefore, it must be expected that an attacker who is more skilled in mask-making might eventually be able to produce a silicon mask that gets accepted by FaceVACS.

Both proposed anti-spoofing methods successfully reject all evaluated attacks and, with regard to this scenario, achieve an FAR$_{cf}$ = 0%. An example of the multispectral SWIR image of a silicon mask and its classification result compared to the corresponding RGB color image is shown in Figure 7.17 on the facing page. Furthermore, Table 7.13 shows a qualitative comparison of these results to prior work.

Disguise Scenario

This scenario focuses on situations in which an attacker does not need to counterfeit a specific person’s identity, but simply tries to disguise his own, for example, because his face is known and “blacklisted”. Therefore, for this scenario, a *false acceptance*



(a) Original RGB image. (b) Spoof RGB image. (c) Spoof SWIR image. (d) Spoof classification.

Figure 7.17: (a): Original face; (b)-(d): RGB color image, SWIR false-color image and classification result of an attackers face with a silicon mask of the original face.

Table 7.13: Qualitative comparison to reported results of existing approaches on different datasets. * = Based on Rank-1-Identification; ** = only 2D attacks

Method	FAR	FRR
(A) Masking	0.0	4.21 / 0.0*
(B) ROI	0.0	4.95 / 0.0*
Buciu *et al.* [100]**	2.5	6.2
Kose *et al.* [103]	14.0 / 9.1	9.8 / 18.8
Wang *et al.* [105]	3.0	3.3
Yi *et al.* [127]**	0.0	6.0

occurs if the attacker can hide his identity without the spoof being detected. Possible examples for this scenario are face recognition systems used to protect public or critical infrastructure, which might rely on a database of known persons who are not allowed to enter. A known hooligan, for example, who is not allowed to enter a football stadium, might try to attack such a system by disguising his face in a way that is still detected as a face, but not recognized as his.

Obviously, the attacks evaluated in the counterfeiting scenario also allow an attacker to disguise his identity. Additionally, several partial disguises and alterations to a face have been tested, which are listed in Table 7.14 on the next page. For each spoof, the table lists the number of correct identifications (*i.e.* the attacker did not succeed with hiding his identity) and false acceptances using the standard FR system without anti-spoofing, as well as both anti-spoofing methods. For the latter, the number of detected spoofing attempts is given as well.

Without any anti-spoofing module, FaceVACS is easily tricked by all full face attacks, but does perform well on the partial attacks: in all but two cases, it identifies

Table 7.14: Evaluated spoofing attacks for the disguise scenario, number of correct identifications, detected spoofs and false acceptances for FaceVACS without anti-spoofing (using RGB images for query) compared to both anti-spoofing methods (using SWIR images for query). ** = see Tab. 7.12, ** = with and without makeup, *** = with and without artificial blood.*

		Classifier	**FaceVACS**		**(A) Masking Method**			**(B) ROI Method**		
Description / number of images		**FPR**	Attacker ident.	False Acc.	Attacker ident.	Spoof det.	False Acc.	Attacker ident.	Spoof det.	False Acc.
Full face counterfeiting attacks *	27	0,0%	0	27	0	27	0	0	27	0
Full face masks, latex	6	0,0%	0	6	0	6	0	0	6	0
Partial face 2D attacks	12	0,0%	12	0	6	6	0	0	12	0
Partial mask 1, unknown mat.	3	0,0%	3	0	0	3	0	0	3	0
Partial mask 2, cotton on plastic	3	0,0%	3	0	2	0	1	0	3	0
Partial mask 3, cotton on plastic	3	0,0%	3	0	0	0	3	1	3	0
Fake nose 1, foam plastic	3	0,0%	3	0	2	1	1	2	3	0
Fake nose 2, rubber latex	3	0,0%	3	0	3	0	0	3	3	0
Soft nose putty (1) **	3	0,3%	3	0	3	0	0	3	3	0
Soft nose putty (2) **	3	0,0%	3	0	3	0	0	3	3	0
Liquid rubber latex **	2	0,0%	2	0	2	0	0	2	0	0
Fake scar, latex **	2	0,0%	2	0	2	0	0	2	0	0
Fake scar w. art. blood	1	15,9%	1	0	1	0	0	1	0	0
Fake mustache, blond	2	0,0%	2	0	2	0	0	2	0	0
Fake eyebrows / mustache, gray	3	0,0%	3	0	1	1	1	1	2	1
Fake full beard, black	3	0,0%	3	0	3	0	0	3	0	0
Fake glasses / nose / eyebrows	3	0,0%	2	1	0	2	1	0	3	0
Headscarf	2	0,0%	1	1	1	1	0	1	1	0
Makeup / facial cream / tattoos	18	–	18	0	15	3	0	18	3	0
Total	**102**	**1,6%**	**65,7%**	**34,3%**	**45,1%**	**49,0%**	**6,9%**	**41,2%**	**73,5%**	**1,0%**

the attacker in spite of them. For the partial 2D attacks (*i.e.* parts of the attackers face are covered by a cropped image of another person's face, which is also known to the face recognition system), FaceVACS listed both the attacker and the imitated person as matches. However, by combining several of these attacks, a successful disguise might still be possible. In total, FaceVACS achieves an $\text{FAR}_{\text{dg}} \approx 34\%$.

Method (A) can detect spoofing attacks only by comparing face detection results before and after masking out non-skin areas. Therefore, the detection of partial disguises is not reliably possible with this method. At the same time, the disguises are more successful on SWIR images than on the VIS images used for the evaluation of FaceVACS. In combination, the attacker managed to disguise his true identity in $\text{FAR}_{\text{dg}} \approx 7\%$ of the query images without the attack being detected.

Method (B) detects most of the partial disguises and misses only those that are too small or out of the specified regions, which is uncritical as these attacks are unlikely to successfully disguise the identity. Only in one image of a face with fake eyebrows and mustache, the attack is not detected and the attacker is not identified correctly. By using a VIS image as input for the FR software instead of the SWIR image, which is optionally possible with this method, the attacker is identified correctly in this case, though. In total, the ROI method achieves an $\text{FAR}_{\text{dg}} = 1\%$.

Receiver operating characteristics (*i.e.*, ROC curves) for both scenarios are shown in Figure 7.18. It has to be noted that the printed 2D attacks and hard resin masks achieve recognition scores similar to real faces, while all other masks achieve far lower scores. Thus, the ROC curves of FaceVACS appear as a step function, as either all valid faces are falsely rejected or all of these spoofs are falsely accepted without a slow transition.

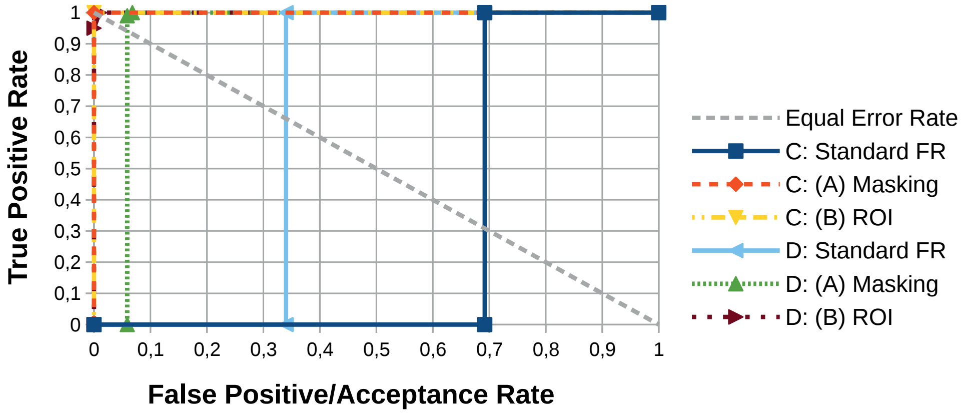


Figure 7.18: Receiver operating characteristic (ROC) curves for the counterfeit (C) and disguise (D) scenario.

7.6 Summary

In this chapter, the proposed system design and preprocessing methods, as well as skin detection and anti-spoofing performance are validated. In all of these aspects, the design goals specified in Section 3.1 can be met by the realized camera system.

Results of the calibration methods for fixed pattern noise and ring light homogeneity are presented and the influence of ambient light and the maximum operation range are analyzed. Ambient light only has a minor impact on the acquired spectral signatures and limits the operation range to approx. 1.5 m for the first ring light revision and approx. 3.5 m for the second revision. Thus, the second revision easily covers the typical operating distance of eGate systems.

Although not being a specified design goal, the accuracy of the depth estimation method presented in Section 6.6 is evaluated as well. While the estimation is too imprecise to provide three-dimensional geometry information of a face, it can still be used as a validity check to support face anti-spoofing. However, due to a high number of outliers, this method was not further used in this work, as the false rejection rate would suffer strongly. In future work, this method will be further improved.

Furthermore, it is shown that motion artifacts at the edges of moving objects can be significantly reduced to avoid disturbances of the spectral signatures using a suited motion compensation approach. In total, 533 combinations of FSWC motion compensation approaches, implemented using state-of-the-art block matching and dense optical flow algorithms, are evaluated on a created database of FSWC image sequences including ground truth information and on appropriate datasets from the *Middlebury optical flow evaluation*. It is shown that the best ICM methods achieve significantly higher accuracy compared to IFI methods, especially in the presence of non-linear motion. A very good compromise between quality and real-time capable processing times can be achieved by the runtime-optimized FlowLib implementation of the Lukas-Kanade optical flow applied on normalized gradient images in a channel-to-channel matching approach (denoted as *C2C-TG N+G LK** in Section 7.2). For multispectral image cubes with four wavebands plus reference channel, a frame rate of up to 18 FPS can be achieved.

The performance and robustness of the skin classification approach is evaluated based on data acquired during an extensive study that includes spectral remission data and multispectral images of 152 participants and 355 material samples. Skin and material samples can be separated almost perfectly using a machine-learning-based classifier. Furthermore, neither typical make-up, nor the skin type, gender or age of a subject have a significant influence on the classification results.

Finally, the proposed anti-spoofing approach is validated. For this purpose, the dataset is extended by multispectral images of different spoofing attacks, including (partial) disguises such as fake facial hair or fake noses, as well as full face masks made from silicon or latex. The results show that the proposed system achieves unprecedented performance in skin and spoof detection.

Chapter 8

Conclusions and Outlook

8.1 Summary

In this dissertation, a reference design and concept for an active multispectral short-wavelength infrared (SWIR) imaging system optimized for skin detection and face verification called *SkinCam* is presented and validated. Its fine-grained skin classification on pixel-level allows to detect and repel spoofing attacks at biometric face recognition systems in order to increase both their security and acceptance. Despite the significant progress in the field of face recognition, such attacks are still a serious problem for the current state of the art [1, 12, 93].

The proposed skin detection method is based on the analysis of several narrow wavebands in the SWIR spectral range. This so-called spectral signatures are well suited to distinguish skin from other materials as the remission spectra of authentic human skin are very characteristic in this spectral range and independent of the skin type, age or gender of a subject. The spectral signatures are acquired by using active (pulsed) illumination in the distinct wavebands in combination with a camera that is sensitive to the full SWIR spectrum. This method of acquiring multispectral image cubes is denoted as field sequential waveband capturing (FSWC). Compared to other designs of multispectral imaging systems, FSWC-based imaging with active illumination has several advantages: it eliminates the influences of ambient light almost completely and allows for flexible waveband configuration as well as a fast acquisition without degrading image resolution. In addition, the frontal illumination avoids shadows and ensures reliable face recognition with constant conditions.

The proposed method raises two major challenges: first, due to the sequential waveband acquisition, an efficient and real-time capable motion compensation method is required. Second, an accurate pixel-level material classifier must be combined with state-of-the-art face recognition methods without opening up new ways to attack the recognition system. With respect to these challenges, the following methodological contributions are presented in this work:

First approaches to motion compensation for FSWC-based imaging; see Chapter 4.

The major problem in motion compensation for FSWC imaging originates from the varying remission intensities in the subsequent spectral channels, as most state-of-the-art algorithms for motion estimation assume constant intensities of corresponding pixels. This work presents two fundamental concepts to solve this problem: inter-frame interpolation (IFI) and inter-channel matching (ICM). While IFI estimates motion vector fields between corresponding channels of successive multispectral cubes to avoid any inconsistencies of pixel intensities, ICM estimates motion fields between neighboring channels within a multispectral cube to achieve shorter displacement vectors. As this requires appropriate handling of the inconsistent intensities, different approaches based on normalization, intensity transformation and correlation are introduced. All methods and variations are extensively evaluated on a newly created database of FSWC imagery, partially comprising ground truth.

Cross-modal methods to integrate multispectral SWIR skin authentication into existing face verification systems; see Chapter 5. Based on the extraction of spectral signatures from individual pixels of multispectral SWIR images, a two-stage skin classification approach is proposed. It consists of the coarse-grained thresholding on normalized differences between the pixel intensities of all spectral channels and a subsequent fine-grained classifier based on machine learning techniques. For this purpose, binary decision trees, random forests and support vector machines (SVMs) have been evaluated, with SVMs showing the best overall classification performance.

To integrate this per-pixel skin classification with existing state-of-the-art face recognition methods, two different methods are proposed which allow to verify a face captured using *SkinCam* against a known face given by a previously captured SWIR or even a visual (VIS) spectrum image, for example, from an already existing face database. The first method masks out non-skin pixels from a captured SWIR image prior to face recognition in a preprocessing step to ensure that no (possibly forged) non-skin pixels are used in the recognition process. It requires that a given face recognition system is capable of handling SWIR imagery as input. In contrast to this, the second method allows to use two cameras to capture both a VIS and a SWIR image of the same face simultaneously and can thus easily be used to enhance any given face recognition system. It applies skin classification on the SWIR image and performs anti-spoofing based on a generic region of interest (ROI) in a postprocessing step, while the face recognition system can be fed with either the SWIR or the VIS image.

Furthermore, this work presents the system design, implementation and in-depth validation of a camera system with active LED illumination that is based on the proposed reference design; see Chapter 6. This includes the following, rather engineering oriented contributions:

The camera system design, setup and implementation details. The design is separated into three major building blocks which are described in detail: the camera system hardware, image processing and image analysis. The developed camera system is controlled with a microcontroller embedded into the illumination module and acquires four-band multispectral image cubes in real time with up to 20 FPS. Wavebands of the ring light have been selected based on the analysis of spectral data with a special software tool. Its design was optimized using optical simulations to achieve a homogeneous illumination. Image processing and analysis is performed in software on a connected PC and includes a fixed pattern noise correction, motion compensation, intrinsic camera and ring light calibration, as well as two-stage skin classification and cross-modal face verification methods.

An eye safety evaluation for the illumination module. As the emitted SWIR radiation is not visible to the human eye, an evaluation of the eye safety is indispensable. Using both simulations and calculations based on a theoretical worst case scenario, the irradiance on the eye of an observer is estimated according to applicable norms and shown to be far below permissible limits.

A depth estimation method based on chromatic aberrations. Longitudinal chromatic aberration lead to a focus shift of the camera system depending on the waveband. This inevitable flaw of each optical system is used to estimate the distance between the camera and an object in the image within a limited distance range. However, this method showed to be rather imprecise in practice and needs to be improved in future work in order to be of use for further image analysis.

Finally, the proposed concept and methods are validated based on an in-depth evaluation of the implemented camera system; see Chapter 7. Results of the calibration methods are presented, the influence of ambient light and the maximum operation range of the developed system are analyzed and reach all specified design goals. In addition, an extensive study was carried out in the context of this work in order to acquire spectral data and multispectral images of 152 participants and 355 material samples to evaluate the skin classification performance. This dataset was further extended by multispectral images of different spoofing attacks used to validate the proposed anti-spoofing approach. It was shown that the proposed system achieves unprecedented performance in skin and spoof detection.

All databases that have been created in the context of this work, *i.e.*, the FSWC motion compensation sequences, skin and face database, as well as the database of spoofing attacks, are available to the research community on the website of the Institute for Safety and Security Research (ISF) of the Bonn-Rhein-Sieg University of Applied Sciences (BRSU): `https://isf.h-brs.de`.

8.2 Outlook to Other Applications

Besides its use for face verification, reliable image-based skin detection can also help to increase the safety at robot workplaces up to the required reliability values, especially when it comes to human-robot collaboration in so-called joint-action scenarios. The feasibility of this approach has been investigated in the context of the research project *safe person detection in working areas of industrial robots (SPAI)*. The proposed safeguarding concept is demonstrated in Figure 8.1 on the next page: a slow down or stop signal for the robot is triggered as soon as the safety zone (yellow) is entered by a person. As soon as a person or his/her limbs enter the dangerous zone (red), any dangerous motion must already be stopped. This concept introduces slightly different requirements compared to the face verification application scenarios:

1. **Sufficiently fast processing and reaction time.** The industrial standard ISO-13855 [128] defines an expected approaching speed of a human towards a dangerous zone with $v < 2\,\mathrm{m\,s^{-1}}$ and provides a method to calculate the required minimum size of the observed safety zone for protective devices based on its reaction time, which can be adapted here.

2. **Sufficient operating range and spatial resolution.** The angle of view and operating range of the safeguarding camera system has to be wide enough to cover the complete safety zone. Based on the specification of a state-of-the-art *vision-based protective device*, the SafetyEYE[1], an angle of view of $\alpha \approx 65°$ and an operating range of $d_{max} = 7\,\mathrm{m}$ is regarded as feasible in practice. At the same time, the

[1] Pilz GmbH & Co. KG (Ostfildern, Germany)

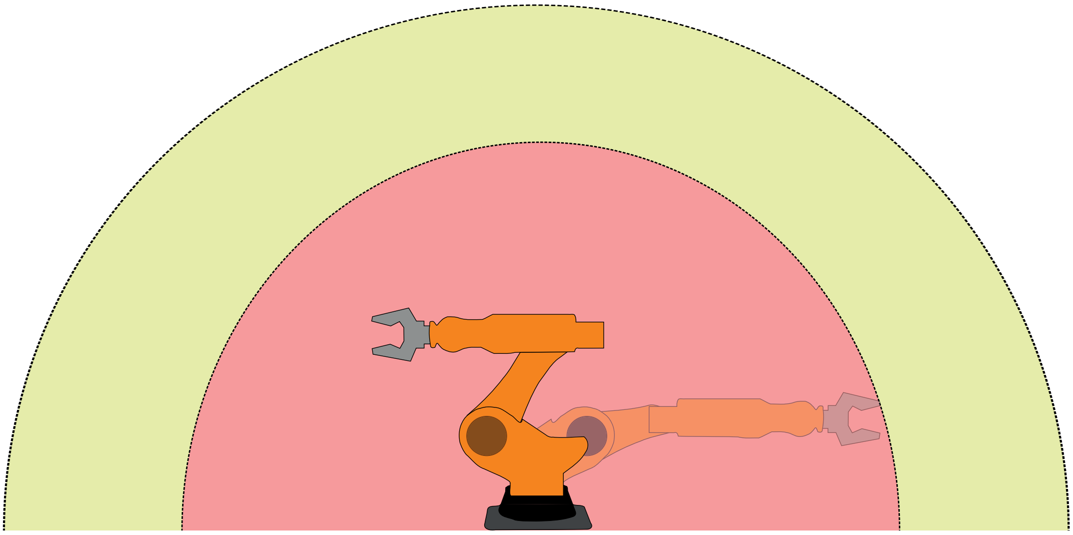


Figure 8.1: Concept for safeguarding of a robot arm with a safety (yellow) and a dangerous (red) zone.

spatial resolution must be fine-grained enough to capture image details such as hands or, depending on the specific application, even fingers of a person everywhere within the safety zone.

3. **Avoidance of occlusions.** Occlusions might impose a safety risk if skin is occluded by another object. Therefore, the camera system should be set up at a place that reduces the risk of occlusions. An alternative might be the use of multiple cameras that are set up at different locations around the safety zone or to move the camera mechanically according to the movements of the robot in order to keep a clear field of view.

A modified system setup based on the camera system described in this dissertation that addresses the safeguarding application has been presented by Sporrer *et al.* [23]. It includes wide angle optics and a specifically optimized ring light, which fulfills the requirements specified above. Image analysis has been extended by a method that matches pixels classified as skin to additional binary masks that define either a warning or a safety zone. In future work, this safeguarding system will be extended with additional modalities and evaluated in greater detail.

Appendix A

Hardware Documentation

A.1 Camera Specifications

Table A.1 provides technical specifications of the Allied Vision cameras (Goldeye P032 and Goldeye G032) that are used in this work, as well as a comparison to the Xenics Cheetah 640 CL. The data is taken from technical datasheets available at www.alliedvision.com and www.xenics.com.

Table A.1: Technical specifications of Allied Vision Goldeye P-032, G-032 and Xenics Cheetah SWIR cameras.

	P-032	**G-032**	**Cheetah 640 CL**
FPA Resolution	636x508	636x508	640x512
FPA Operating Temp.	268 K	278 K	268 K
Max. Frame Rate	30 FPS	100 FPS	1730 FPS
Cell size	25x25 μm	25x25 μm	20x20 μm
Dark Noise	400 e^-	400 e^-	400 e^-
Saturation Capacity	1.9 Me^-	1.9 Me^-	1.1 Me^-
Dynamic Range	73 dB	73 dB	63 dB
Peak Quantum Efficiency	74 %	74 %	80 %

A.2 LED Specifications

Table A.2 presents the specifications of the LED types used for the different illumination modules built in the context of this work and, for comparison, those built by Sporrer [23] to address the safeguarding applications. It states the number (n), peak wavelength (λ_p), FWHM ($\Delta\lambda_{0.5}$), viewing angle (φ), radiated power (Φ_e) and total radiated power ($\sum \Phi_e$) of the used LEDs.

Table A.2: Specifications of the different illumination modules and used LEDs.

(a) Ring light for face recognition, rev. 1. Power rating at $I_F = 100\,\text{mA}$.

LED type	n	λ_p [nm]	$\Delta\lambda_{0.5}$ [nm]	φ [°]	Φ_e [mW]	$\sum \Phi_e$ [mW]
ELD-935-525	10	935	65	20	30	300
ELD-1060-525	30	1060	50	20	5.5	165
ELD-1300-525	20	1300	70	25	8.5	170
ELD-1550-525	30	1550	130	20	5.0	150

(b) Ring light for face recognition, rev. 2. Power rating at $I_F = 100\,\text{mA}$.

LED type	n	λ_p [nm]	$\Delta\lambda_{0.5}$ [nm]	φ [°]	Φ_e [mW]	$\sum \Phi_e$ [mW]
EOLD-1050-525	40	1050	80	20	8.0	320
EOLD-1200-525	40	1200	70	20	7.0	280
EOLD-1300-525	40	1300	70	25	8.5	340
EOLD-1550-525	80	1550	130	20	3.3	264

(c) Ring light for safeguarding, rev. 1. Power rating at $I_F = 600\,\text{mA}$.

LED type	n	λ_p [nm]	$\Delta\lambda_{0.5}$ [nm]	φ [°]	Φ_e [mW]	$\sum \Phi_e$ [mW]
L-970-66-60	8	970	40	60	500	4000
L-1300-66-60	8	1300	80	60	140	1120
L-1550-66-60	16	1550	100	60	60	960

(d) Ring light for safeguarding, rev. 2. Power rating at $I_F = 600\,\text{mA}$.

LED type	n	λ_p [nm]	$\Delta\lambda_{0.5}$ [nm]	φ [°]	Φ_e [mW]	$\sum \Phi_e$ [mW]
L-1050-66-60	16	1050	55	60	120	1920
L-1300-66-60	16	1300	80	60	140	2240
L-1550-66-60	32	1550	100	60	60	1920

A.3 Ring Light and Embedded System Design

For both ring light revisions, the LED placement patterns have been transferred to the EAGLE CAD software and the required electrical circuits have been added, as well as the already described components of the embedded microcontroller system; see Section 6.2.2 on page 73. In order to provide the LEDs with an accurate and constant forward current of up to $I_F = 200\,\mathrm{mA}$, adjustable versions of the LM1117 voltage regulators are used. The LM1117ADJ creates a constant voltage of $V_{\mathrm{out}} = 1.25\,\mathrm{V}$ at its output pin[1]. By connecting their output and ground pins with a high precision resistor, a constant current flow $I = 1.25\,\mathrm{V}/R$ is established between the ground pin of the voltage regulator and the actual ground - here, the LEDs are added in series connection. The supply voltage $V_{\mathrm{supply}} = 12\,\mathrm{V}$ of the ring light is provided by the AC adapter of the camera. As the forward voltages of the LEDs can be up to $V_F \leq 1.5\,\mathrm{V}$, depending on the LED type, and sum up when using this type of connection, at most 8 LEDs can be connected in series. Thus, several of these supply lines have been implemented in parallel.

The following Figures A.1, A.2, A.3, and A.4 present the schematics / circuit diagrams and board layouts of both ring light revisions with the embedded microcontroller systems.

[1] See LM1117 datasheets, *e.g.*, by Texas Instrument

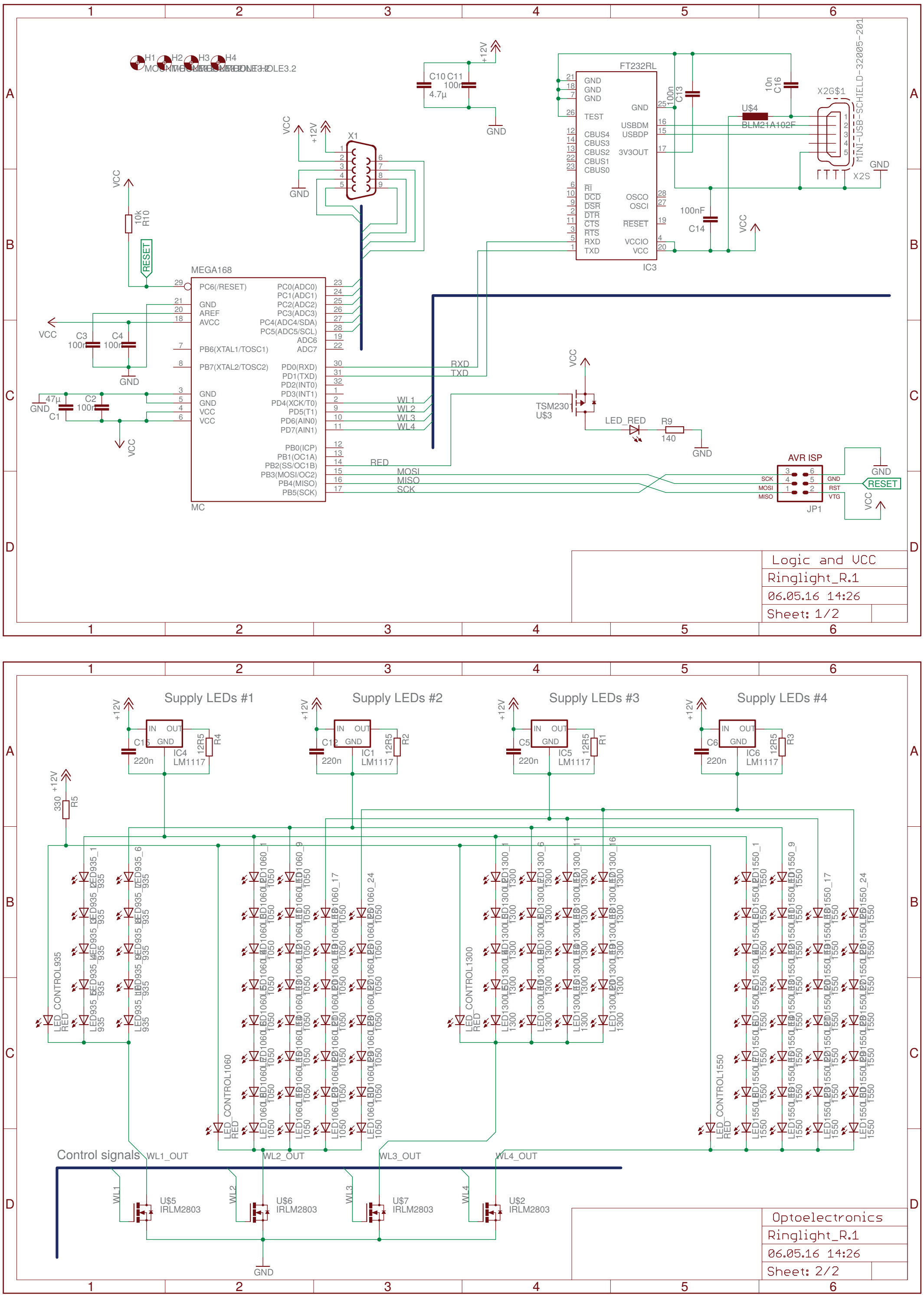


Figure A.1: Schematics of the first ring light revision.

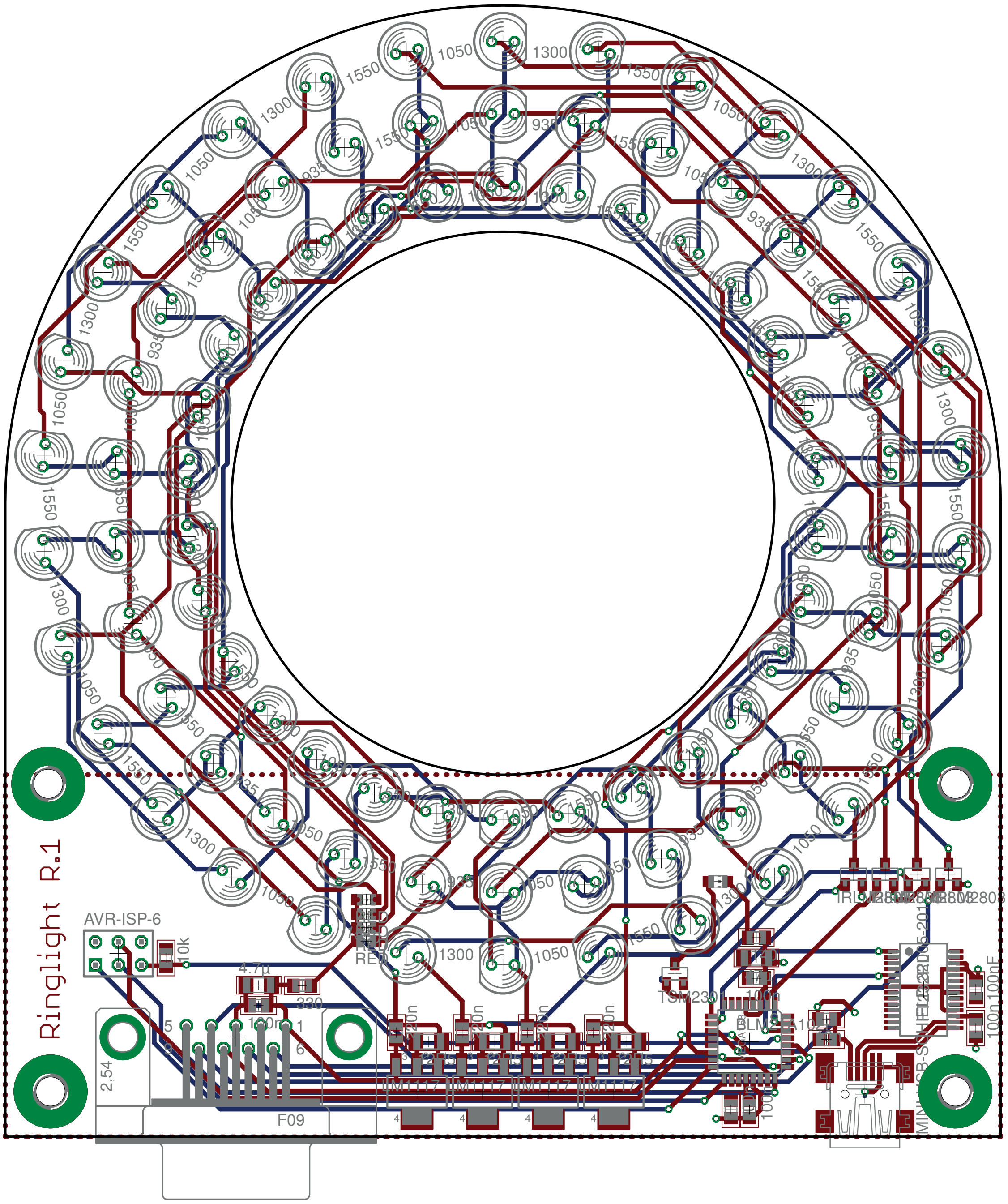


Figure A.2: Printed circuit board layout of the first ring light revision.

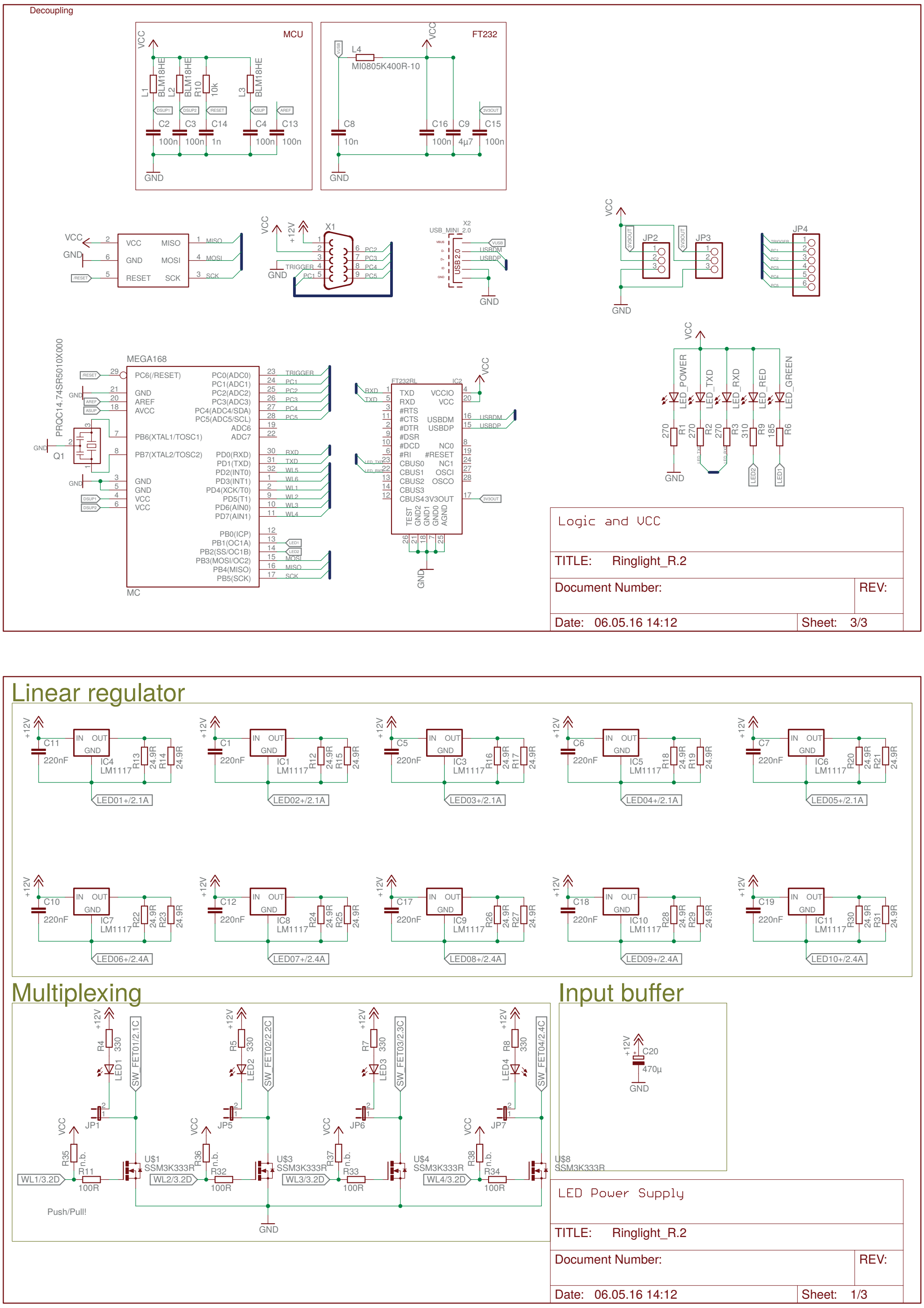


Figure A.3: Schematics of the second ring light revision (LED lines not shown here).

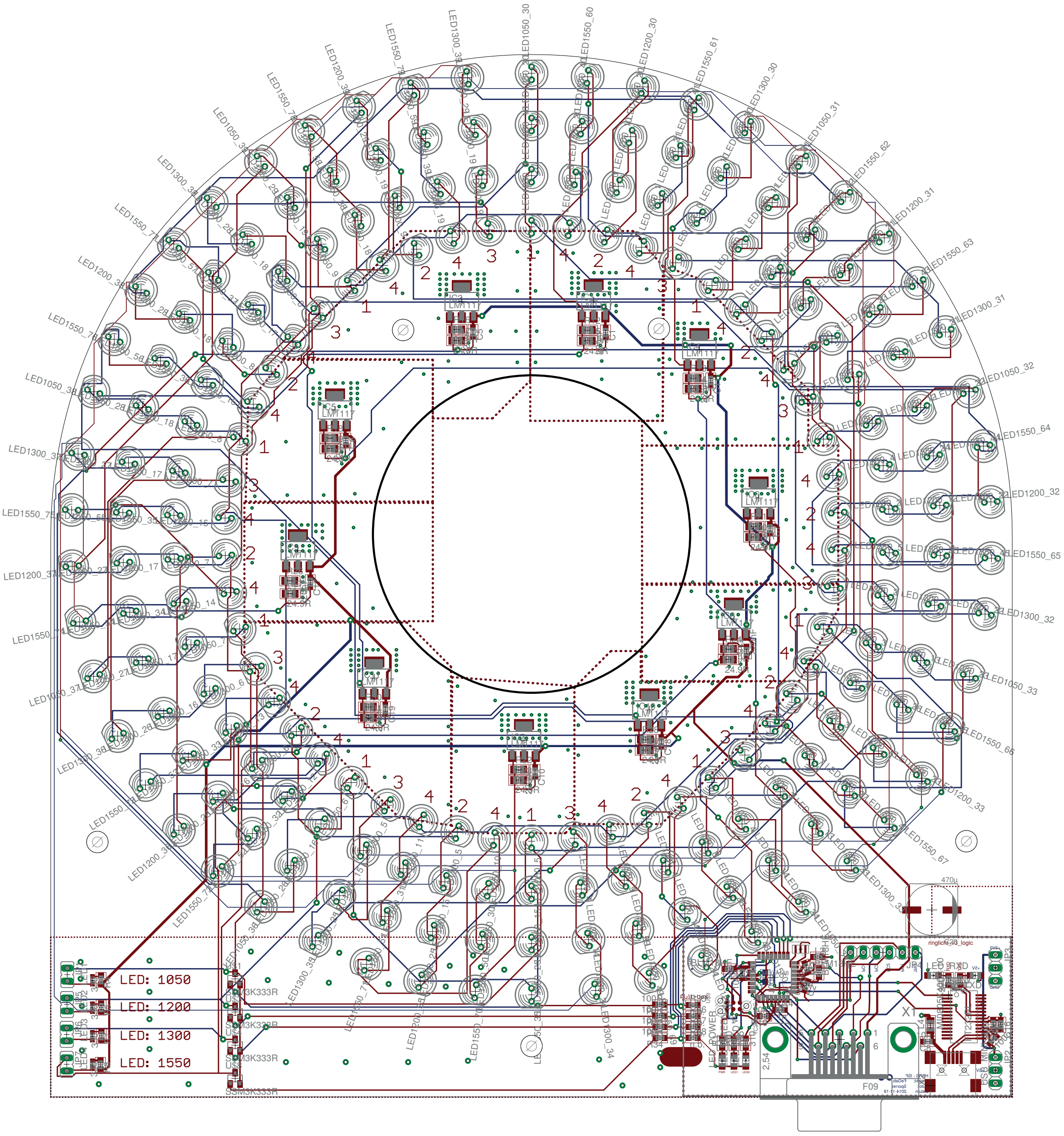


Figure A.4: Printed circuit board layout of the second ring light revision.

Appendix B

Software Documentation

B.1 Software Design

The fundamental design of the SkinCam control software is illustrated in Figure B.1 on the next page. The presented UML class diagram is strongly simplified and shows only the most important attributes and operations in favor of better clarity. Furthermore, Figure B.2 on page 149 shows the basic process flow and inter-thread communication during image acquisition.

B.2 Libraries and Tools

Table B.1: Development tools and libraries used in this work.

Tools	**Version**
GNU C++ Compiler (g++)	4.8.2
Weka	3.6.10
Libraries	**Version**
Qt Framework	5.4.1
openCV	2.4.11
nVidia CUDA	7.5
Allied Vision Vimba API	1.3
libSerial	0.6
cLandmark	2015-11
flowLib	2.2
libSVM	3.12

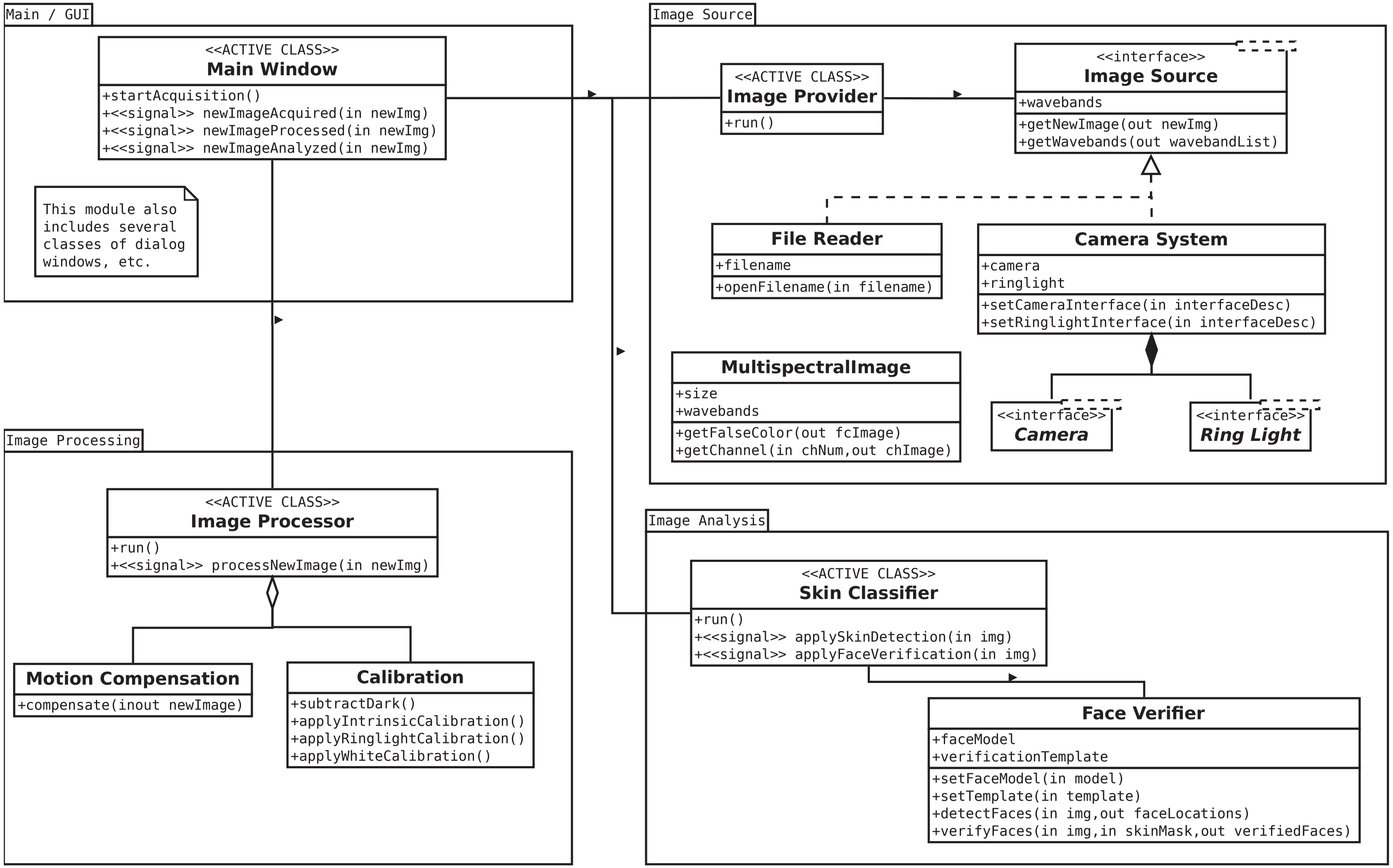

Main / GUI
<<ACTIVE CLASS>>
Main Window
+startAcquisition()
+<<signal>> newImageAcquired(in newImg)
+<<signal>> newImageProcessed(in newImg)
+<<signal>> newImageAnalyzed(in newImg)
This module also includes several classes of dialog windows, etc.
Image Source
<<ACTIVE CLASS>>
Image Provider
+run()
<<interface>>
Image Source
+wavebands
+getNewImage(out newImg)
+getWavebands(out wavebandList)
File Reader
+filename
+openFilename(in filename)
Camera System
+camera
+ringlight
+setCameraInterface(in interfaceDesc)
+setRinglightInterface(in interfaceDesc)
MultispectralImage
+size
+wavebands
+getFalseColor(out fcImage)
+getChannel(in chNum,out chImage)
<<interface>>
Camera
<<interface>>
Ring Light
Image Processing
<<ACTIVE CLASS>>
Image Processor
+run()
+<<signal>> processNewImage(in newImg)
Motion Compensation
+compensate(inout newImage)
Calibration
+subtractDark()
+applyIntrinsicCalibration()
+applyRinglightCalibration()
+applyWhiteCalibration()
Image Analysis
<<ACTIVE CLASS>>
Skin Classifier
+run()
+<<signal>> applySkinDetection(in img)
+<<signal>> applyFaceVerification(in img)
Face Verifier
+faceModel
+verificationTemplate
+setFaceModel(in model)
+setTemplate(in template)
+detectFaces(in img,out faceLocations)
+verifyFaces(in img,in skinMask,out verifiedFaces)

Figure B.1: (Simplified) class diagram describing the modules of the *SkinCam* control software.

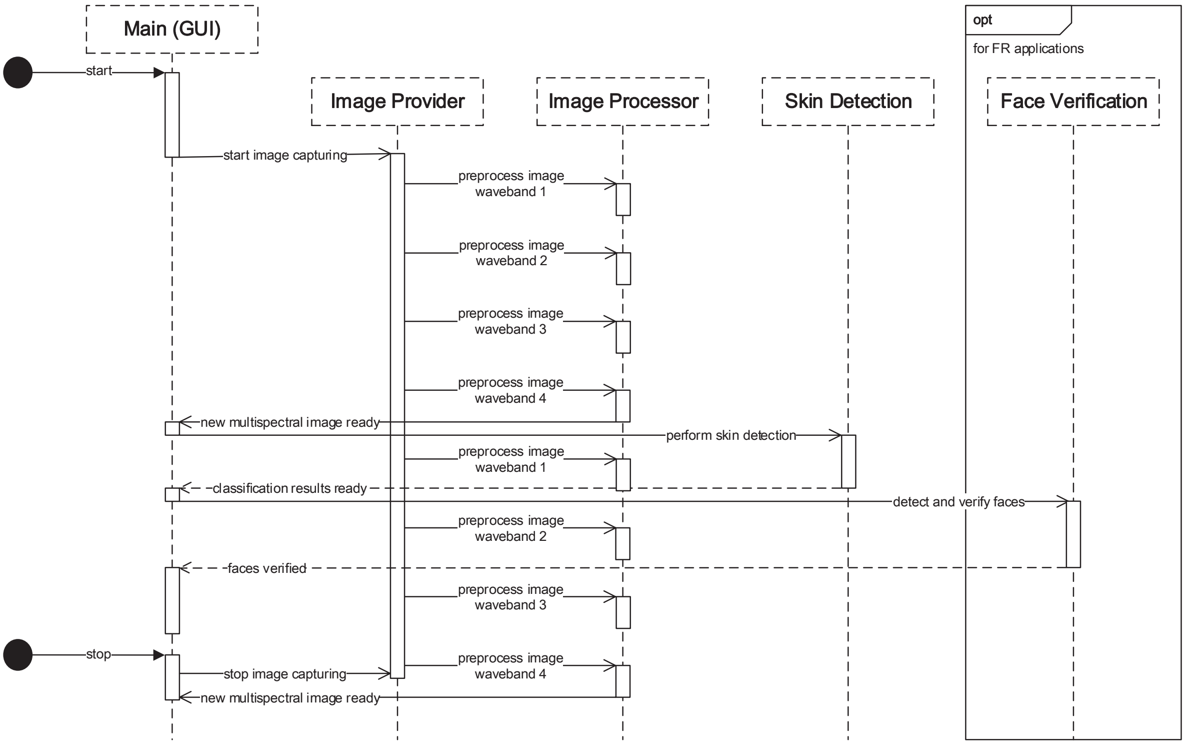

Main (GUI)
start
Image Provider
Image Processor
Skin Detection
opt
for FR applications
Face Verification
start image capturing
preprocess image waveband 1
preprocess image waveband 2
preprocess image waveband 3
preprocess image waveband 4
new multispectral image ready
perform skin detection
preprocess image waveband 1
classification results ready
detect and verify faces
preprocess image waveband 2
faces verified
preprocess image waveband 3
stop
stop image capturing
preprocess image waveband 4
new multispectral image ready

Figure B.2: Sequence diagram describing the relations between tasks and threads of the *SkinCam* control software.

Appendix C

Datasets

In the context of this work, three databases have been created and are available for download on the website of the Safety and Security Research Institute (ISF) of the Bonn-Rhein-Sieg University of Applied Sciences (BRSU) on `http://isf.h-brs.de/`.

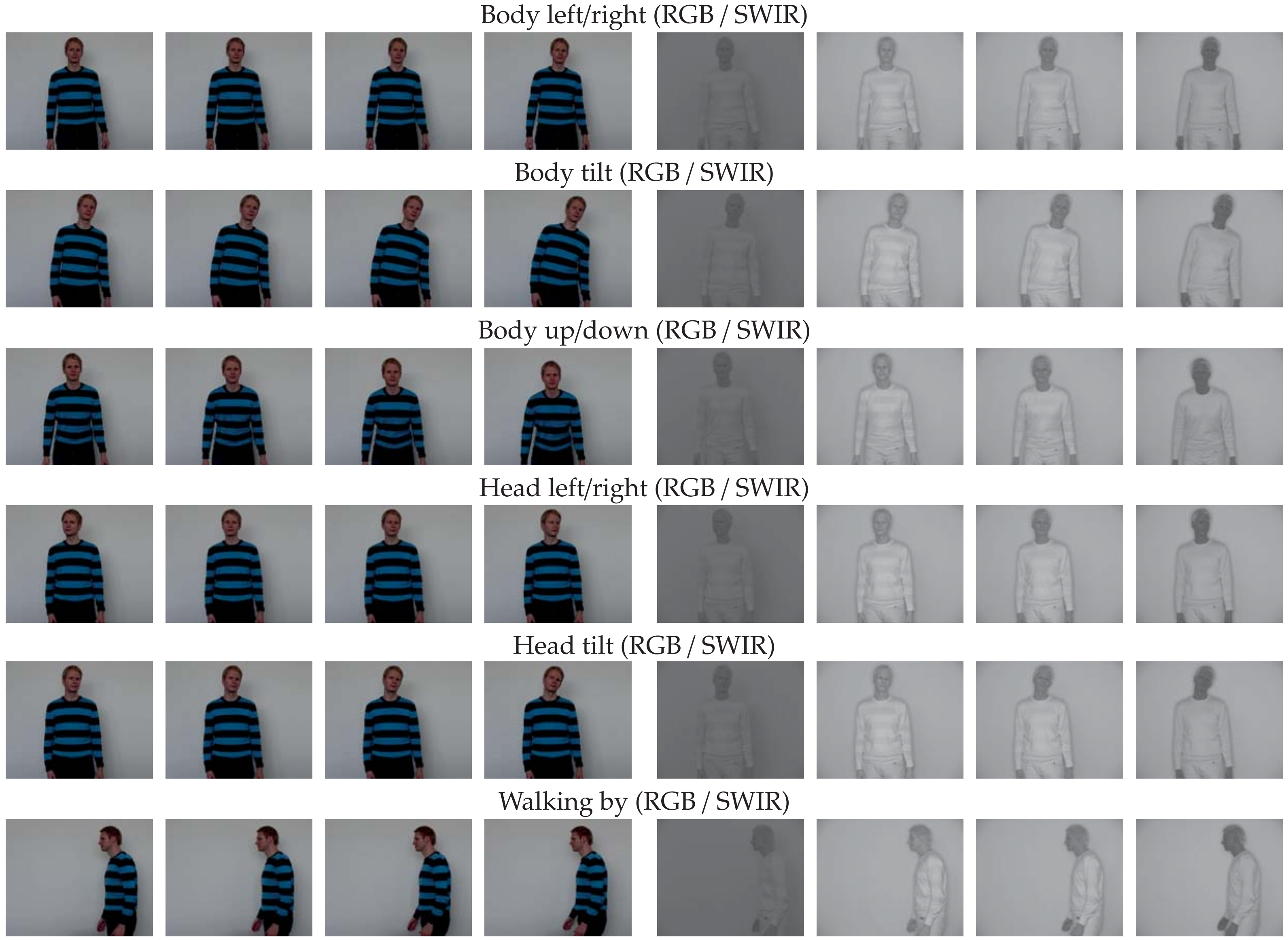
C.1 FSWC Optical Flow Database

This database contains a set of corresponding RGB color and multispectral short-wavelength infrared (SWIR) video sequences recorded using two cameras simultaneously and is provided to other researchers in order to promote research on motion compensation for FSWC based imaging systems; see Section 7.2 on page 100. Examples for each optical flow dataset are shown in Figure C.1 and Figure C.2. All sequences have been captured twice, once with 30 FPS and once with 60 FPS. The RGB and SWIR cameras had a slightly different perspective, but the sequences have been centered, resampled and cropped to match each other as good as possible.

C.2 Skin/Face Database

The multispectral skin/face database consists of two parts:

Spectro: Contains spectrometer measurements acquired from 120 subjects, measured at 16 positions on face, neck, arms and hands with a spectral range of 660 nm to 1700 nm using a irSys 1.7 spectrometer by TQ Group GmbH; see Section 7.4.2 on page 118. These remission measurements have been performed using a



Body left/right (RGB / SWIR)

Body tilt (RGB / SWIR)

Body up/down (RGB / SWIR)

Head left/right (RGB / SWIR)

Head tilt (RGB / SWIR)

Walking by (RGB / SWIR)

Figure C.1: FSWC optical flow database: datasets part 1

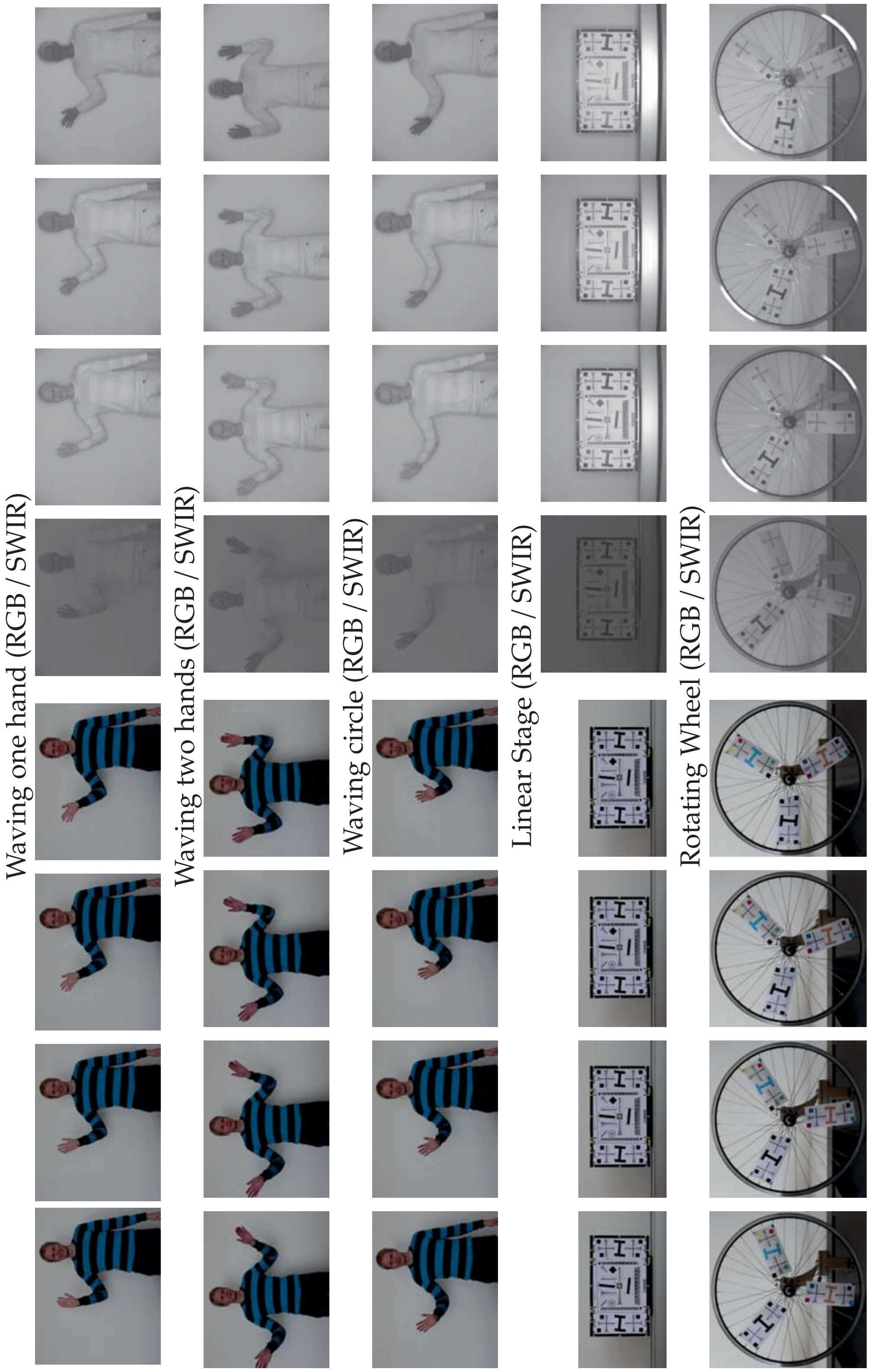

Waving one hand (RGB / SWIR)
Waving two hands (RGB / SWIR)
Waving circle (RGB / SWIR)
Linear Stage (RGB / SWIR)
Rotating Wheel (RGB / SWIR)

Figure C.2: FSWC optical flow database: datasets part 2

special two-way lightguide and a standard halogen lamp. For every data set, information about gender, age and skin type are provided.

Face: Contains color (RGB) and multispectral SWIR images consisting of the wavebands $\lambda_1 = 935\,\text{nm}, \lambda_2 = 1\,060\,\text{nm}, \lambda_3 = 1\,300\,\text{nm}$ and $\lambda_4 = 1\,550\,\text{nm}$, as well as the corresponding spectrometer measurements, of (currently) more than 150 subjects. The RGB images were taken with a Canon EOS 50D with a f=50 mm lens. SWIR images were acquired with the *SkinCam* system described in this work. As not all of the participants agreed to publication of the acquired images, only a limited number of datasets is available to the public, including (currently) more than 50 subjects. For every dataset, information about gender, age and skin type are provided. Examples of the acquired images are shown in Figure 2.4 on page 14.

C.3 Spoofing Attack Database

The spoof database contains RGB color and multispectral SWIR images consisting of the wavebands $\lambda_1 = 935\,\text{nm}, \lambda_2 = 1\,060\,\text{nm}, \lambda_3 = 1\,300\,\text{nm}$ and $\lambda_4 = 1\,550\,\text{nm}$ as well as the corresponding spectrometer measurements of a variety of spoofing attacks, including (partial) facial disguises and masks; see Section 7.5.1 on page 121. The RGB images were taken with a Canon EOS 50D with a 50 mm lens. SWIR images were acquired with the *SkinCam* system described in this work. The dataset can be used for training and testing of spoof detection methods using both modalities. Examples of the included spoofing attacks are shown in Figure 7.16 on page 122.

Bibliography

[1] N. Erdogmus and S. Marcel, “Spoofing face recognition with 3D masks,” *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 7, pp. 1084–1097, 2014.

[2] T. Bourlai and B. Cukic, “Multi-spectral face recognition: Identification of people in difficult environments,” in *Proc. Int. IEEE Conf. on Intelligence and Security Informatics (ISI)*, 2012, pp. 196–201.

[3] R. S. Ghiass, O. Arandjelović, A. Bendada, and X. Maldague, “Infrared face recognition: A comprehensive review of methodologies and databases,” *Pattern Recognition*, vol. 47, no. 9, pp. 2807 – 2824, 2014.

[4] T. Bourlai, N. Kalka, A. Ross, B. Cukic, and L. Hornak, “Cross-spectral face verification in the short wave infrared (SWIR) band,” in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, 2010, pp. 1343–1347.

[5] I. Pavlidis and P. Symosek, “The imaging issue in an automatic face/disguise detection system,” *Proc. IEEE Workshop on Computer Vision Beyond the Visible Spectrum*, p. 15, 2000.

[6] A. S. Nunez, “A physical model of human skin and its application for search and rescue,” Ph.D. dissertation, Air Force Inst. of Tech. Wright-Patterson AFB OH School of Engineering, 2009.

[7] A. Nunez and M. Mendenhall, “Detection of human skin in near infrared hyperspectral imagery,” in *Proc. IEEE Int. Geoscience and Remote Sensing Symposium*, vol. 2, 2008, pp. II–621 –II–624.

[8] T. Bourlai, N. Narang, B. Cukic, and L. Hornak, “On designing a SWIR multiwavelength facial-based acquisition system,” in *Infrared Technology and Applications XXXVIII*, vol. 8353, 2012, pp. 83 530R–83 530R–14.

[9] M. Bertozzi, R. Fedriga, A. Miron, and J.-L. Reverchon, “Pedestrian detection in poor visibility conditions: Would SWIR help?” in *Image Analysis and Processing*

(ICIAP), ser. Lecture Notes in Computer Science, A. Petrosino, Ed. Springer Berlin Heidelberg, 2013, vol. 8157, pp. 229–238.

[10] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, 2nd ed., S. Z. Li and A. K. Jain, Eds. Springer, 2011.

[11] K. A. Nixon, V. Aimale, and R. K. Rowe, "Spoof detection schemes," in *Handbook of biometrics*, A. K. Jain, P. Flynn, and A. A. Ross, Eds. Springer, 2008, pp. 403–423.

[12] N. Kose and J.-L. Dugelay, "On the vulnerability of face recognition systems to spoofing mask attacks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 2357–2361.

[13] A. Opitz and A. Kriechbaum-Zabini, "Evaluation of face recognition technologies for identity verification in an egate based on operational data of an airport," in *Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, 2015, pp. 1–5.

[14] K. Patel, H. Han, A. Jain, and G. Ott, "Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks," in *Proc. Int. Conf. on Biometrics (ICB)*, 2015, pp. 98–105.

[15] D. Reinert, O. Schwaneberg, N. Jung, S. Ullmann, W. Olbert, D. Kamin, and R. Kohler, "Finger and hand protection on circular table and panel saws," *Safety Science*, vol. 47, no. 8, pp. 1175 – 1184, 2009.

[16] C. Lenz, S. Nair, M. Rickert, A. Knoll, W. Rosel, J. Gast, A. Bannat, and F. Wallhoff, "Joint-action for humans and industrial robots for assembly tasks," in *IEEE Int. Symp. on Robot and Human Interactive Communication*, 2008, pp. 130–135.

[17] R. D. Kilmer, "Safety sensor systems for industrial robots," in *Proc. SME Conf. on Robots*, 1982, pp. 479–491.

[18] A. Clodic, R. Alami, C. Vesper, E. Pacherie, B. Mutlu, and J. A. Shah, "FJA@HRI15: Towards a framework for joint action," in *Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction*, ser. HRI'15 Extended Abstracts. ACM, 2015, pp. 259–260.

[19] M. Kudo and T. Inoue, "Multi-optical-path photoelectric safety apparatus," US Patent 6,979,814, 2005.

[20] H. Steiner, O. Schwaneberg, and N. Jung, "Advances in active near-infrared sensor systems for material classification," in *Imaging Systems and Applications*. Optical Society of America, 2012, p. ITu2C.2.

[21] H. Steiner, S. Sporrer, A. Kolb, and N. Jung, “Design of an active multispectral SWIR camera system for skin detection and face verification,” *Journal of Sensors*, vol. 2016, no. 1, 2016, article ID 9682453, Special Issue on Multispectral, Hyperspectral, and Polarimetric Imaging Technology.

[22] H. Steiner, A. Kolb, and N. Jung, “Reliable face anti-spoofing using multispectral swir imaging,” in *Proc. Int. Conf. on Biometrics (ICB)*, 2016, pp. 1–8.

[23] S. Sporrer, H. Steiner, M. Velte, and N. Jung, “NIR camera based person detection in the working range of industrial robots,” in *Proc. Int. Conf. on Safety of Industrial Automated Systems (SIAS)*, 2015, pp. 147–152.

[24] H. Steiner, N. Jung, and A. Kolb, “Real-time motion compensation for field sequential multispectral imaging,” 2017, in preparation.

[25] O. Schwaneberg, “Concept, system design, evaluation and safety requirements for a multispectral sensor,” Ph.D. dissertation, University of Siegen, Fakultät IV: Naturwissenschaftlich-Technische Fakultät, 2013.

[26] A. Gowen, C. O’Donnell, P. Cullen, G. Downey, and J. Frias, “Hyperspectral imaging – an emerging process analytical tool for food quality and safety control,” *Trends in Food Science & Technology*, vol. 18, no. 12, pp. 590 – 598, 2007.

[27] J. Brauers, T. Aach, and S. Helling, “Multispectral image acquisition with flash light sources,” *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 31 103–1–31 103–10, 2009.

[28] N. Gat, “Imaging spectroscopy using tunable filters: a review,” in *Proc. SPIE Wavelet Applications VII*, vol. 4056, 2000, pp. 50–64.

[29] C.-I. Chang, *Hyperspectral Data Processing: Algorithm Design and Analysis*. John Wiley & Sons, 2013.

[30] B. Labitzke, “Visualization and analysis of multispectral image data,” Ph.D. dissertation, University of Siegen, Fakultät IV: Naturwissenschaftlich-Technische Fakultät, 2013.

[31] D. W. Allen, “An overview of spectral imaging of human skin toward face recognition,” in *Face Recognition Across the Imaging Spectrum*, Thirimachos Bourlai, Ed. Springer, 2016, ch. 1-20.

[32] M. Petrou and C. Petrou, *Image Processing: The Fundamentals*, 2nd ed. John Wiley & Sons, 2010.

[33] J. M. Desse, P. Picart, and P. Tankam, "Sensor influence in digital 3L holographic interferometry," *Measurement Science and Technology*, vol. 22, no. 6, p. 064005, 2011.

[34] S. Helling, E. Seidel, and W. Biehlig, "Algorithms for spectral color stimulus reconstruction with a seven-channel multispectral camera," *Conf. on Colour in Graphics, Imaging, and Vision*, vol. 2004, no. 1, pp. 254–258, 2004.

[35] S. Daly and X. Feng, "Method and system for field sequential color image capture using color filter array," US Patent 6,690,422, 2004.

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

[37] J. Jacquez, J. Huss, W. McKeehan, J. Dimitroff, and H. Kuppenheim, "Spectral reflectance of human skin in the region 0.7-2.6μm," *J. of Applied Physiology*, vol. 8, no. 3, p. 297, 1955.

[38] T. L. Troy and S. N. Thennadil, "Optical properties of human skin in the near infrared wavelength range of 1000 to 2200 nm," *J. of Biomedical Optics*, vol. 6, no. 2, pp. 167–176, 2001.

[39] I. V. Meglinski and S. J. Matcher, "Quantitative assessment of skin layers absorption and skin reflectance spectra simulation in the visible and near-infrared spectral regions," *Physiological Measurement*, vol. 23, no. 4, p. 741, 2002.

[40] T. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Archives of Dermatology*, vol. 124, no. 6, p. 869, 1988.

[41] J. L. Miller, *Principles of Infrared Technology: A Practical Guide to the State of the Art*. Springer Science & Business Media, 1994.

[42] E. Hering and R. Martin, Eds., *Photonik*. Springer, 2006.

[43] S. Kavusi and A. El Gamal, "Quantitative study of high-dynamic-range image sensor architectures," in *Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications*, vol. 5301, 2004, pp. 264–275.

[44] F. Koppens, T. Mueller, P. Avouris, A. Ferrari, M. Vitiello, and M. Polini, "Photodetectors based on graphene, other two-dimensional materials and hybrid systems," *Nature nanotechnology*, vol. 9, no. 10, pp. 780–793, 2014.

[45] F. Pedrotti, L. Pedrotti, W. Bausch, and H. Schmidt, *Optik für Ingenieure*, 3rd ed. Springer, 2002.

[46] B. Jähne, *Digital Image Processing*. Springer Berlin Heidelberg, 2005.

[47] D. Fortun, P. Bouthemy, and C. Kervrann, "Optical flow modeling and computation: A survey," *Computer Vision and Image Understanding*, vol. 134, pp. 1 – 21, 2015.

[48] M. Jakubowski and G. Pastuszak, "Block-based motion estimation algorithms — a survey," *Opto-Electronics Review*, vol. 21, no. 1, pp. 86–102, 2012.

[49] B. K. Horn and B. G. Schunck, "Determining optical flow," *Proc. SPIE*, vol. 0281, pp. 319–331, 1980.

[50] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Conf. on Artificial Intelligence (IJCAI)*, ser. IJCAI'81. Morgan Kaufmann Publishers Inc., 1981, pp. 674–679.

[51] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. Europ. Conf. on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, T. Pajdla and J. Matas, Eds. Springer Berlin Heidelberg, 2004, vol. 3024, pp. 25–36.

[52] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," in *Proc. DAGM Conf. on Pattern Recognition*. Springer Berlin Heidelberg, 2007, pp. 214–223.

[53] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic Huber-L1 optical flow," in *Proc. British Conf. on Machine Vision (BMVC)*, 2009.

[54] M. Werlberger, "Convex approaches for high performance video processing," Ph.D. dissertation, Institute for Computer Graphics and Vision, Graz University of Technology, 2012.

[55] B.-D. Choi, J.-W. Han, C.-S. Kim, and S.-J. Ko, "Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 17, no. 4, pp. 407–416, 2007.

[56] E. Cuevas, D. Zaldivar, M. A. Pérez-Cisneros, and D. Oliva, "Block matching algorithm based on differential evolution for motion estimation," *Engineering Applications of Artificial Intelligence*, vol. abs/1405.4721, 2013.

[57] T. M. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997.

[58] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, pp. 81–106, 1986.

[59] J. R. Quinlan, *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.

[60] J. R. Quinlan, "Learning with continuous classes," in *Proc. Conf. on Artificial Intelligence*, 1992.

[61] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[62] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[63] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2001.

[64] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[65] P. Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, 2012.

[66] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine learning*, vol. 31, no. 1, pp. 1–38, 2004.

[67] S. Z. Li and A. K. Jain, *Handbook of Face Recognition*, 2nd ed. Springer, 2011, ch. 1.

[68] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[69] G. Shakhnarovich and B. Moghaddam, "Face recognition in subspaces," in *Handbook of Face Recognition*, 2nd ed., S. Z. Li and A. K. Jain, Eds. Springer, 2011, ch. 2.

[70] J.-K. Kämäräinen, A. Hadid, and M. Pietikäinen, "Local representation of facial features," in *Handbook of Face Recognition*, 2nd ed., S. Z. Li and A. K. Jain, Eds. Springer, 2011, ch. 4.

[71] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 1701–1708.

[72] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Li, "Face matching between near infrared and visible light images," in *Advances in Biometrics*, ser. Lecture Notes in Computer Science, S.-W. Lee and S. Li, Eds. Springer Berlin Heidelberg, 2007, vol. 4642, pp. 523–530.

[73] B. Klare and A. Jain, "Heterogeneous face recognition: Matching NIR to visible light images," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, 2010, pp. 1513–1516.

[74] D. Goswami, C. H. Chan, D. Windridge, and J. Kittler, "Evaluation of face recognition system in heterogeneous environments (visible vs NIR)," in *Proc. IEEE Int. Conf on Computer Vision (ICCV) Workshops*. IEEE, 2011, pp. 2160–2167.

[75] J.-Y. Zhu, W.-S. Zheng, J.-H. Lai, and S. Z. Li, "Matching NIR face to VIS face using transduction," *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 3, pp. 501–514, 2014.

[76] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106 – 1122, 2007.

[77] M. J. Mendenhall, A. S. Nunez, and R. K. Martin, "Human skin detection in the visible and near infrared," *Applied Optics*, vol. 54, no. 35, pp. 10 559–10 570, 2015.

[78] M. Hacskaylo, "Automatic human body detector," US Patent 4,500,784, 1985.

[79] G. Kilgore and P. Whillock., "Skin detection sensor," US Patent 11/264,654, 2008.

[80] G. E. Determan and D. J. Wunderlin, "Encoded binary liveness detector," US Patent 2008/0 203 307 A1, 2008.

[81] Z. Zhang, D. Yi, Z. Lei, and S. Li, "Face liveness detection by learning multispectral reflectance distributions," in *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2011, pp. 436–441.

[82] O. Schwaneberg, H. Steiner, P. H. Bolívar, and N. Jung, "Design of an LED-based sensor system to distinguish human skin from workpieces in safety applications," *Applied Optics*, vol. 51, no. 12, pp. 1865–1871, 2012.

[83] I. Pavlidis, P. Symosek, B. Fritz, and N. Papanikolopoulos, "A near-infrared fusion scheme for automatic detection of vehicle passengers," in *Proc. IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (CVBVS)*, 1999, pp. 41 –48.

[84] I. Pavlidis, P. Symosek, B. Fritz, M. Bazakos, and N. Papanikolopoulos, "Automatic detection of vehicle occupants: the imaging problem and its solution," *Machine Vision and Applications*, vol. 11, pp. 313–320, 2000.

[85] J. Dowdall, I. Pavlidis, and G. Bebis, "Face detection in the near-IR spectrum," *Image and Vision Computing*, vol. 21, no. 7, pp. 565 – 578, 2003.

[86] H. Chang, A. Koschan, M. Abidi, S. Kong, and C.-H. Won, "Multispectral visible and infrared imaging for face recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2008, pp. 1–6.

[87] K. R. Peskosky, "Design of a monocular multi-spectral skin detection, melanin estimation, and false-alarm suppression system," Master's thesis, Air Force Inst Of Tech Wright-patterson AFB OH, 2010.

[88] M. Lindner and A. Kolb, "Compensation of motion artifacts for time-of-flight cameras," in *Dynamic 3D Imaging*, ser. Lecture Notes in Computer Science, A. Kolb and R. Koch, Eds. Springer Berlin Heidelberg, 2009, vol. 5742, pp. 16–27.

[89] D. Lefloch, T. Hoegg, and A. Kolb, "Real-time motion artifacts compensation of ToF sensors data on GPU," in *Proc. SPIE Three-Dimensional Imaging, Visualization, and Display*, vol. 8738, 2013, pp. 87 380U–87 380U–7.

[90] T. Hoegg, D. Lefloch, and A. Kolb, "Real-time motion artifact compensation for PMD-ToF images," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, ser. Lecture Notes in Computer Science, M. Grzegorzek, C. Theobalt, R. Koch, and A. Kolb, Eds. Springer Berlin Heidelberg, 2013, vol. 8200, pp. 273–288.

[91] M. Schmidt and B. Jahne, "Efficient and robust reduction of motion artifacts for 3D time-of-flight cameras," in *Proc. Int. Conf. on 3D Imaging (IC3D)*, 2011, pp. 1–8.

[92] D. Jimenez, D. Pizarro, and M. Mazo, "Single frame correction of motion artifacts in PMD-based time of flight cameras," *Image and Vision Computing*, vol. 32, no. 12, pp. 1127 – 1143, 2014.

[93] J. Galbally and R. Satta, "Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models," *IET Biometrics*, 2015.

[94] K. Kollreider, H. Fronthaler, and J. Bigun, “Non-intrusive liveness detection by face images,” *Image and Vision Computing*, vol. 27, no. 3, pp. 233 – 244, 2009, special Issue on Multimodal Biometrics Multimodal Biometrics Special Issue.

[95] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Li, “Face liveness detection using 3D structure recovered from a single camera,” in *Proc. Int. Conf. on Biometrics (ICB)*, 2013, pp. 1–6.

[96] A. Anjos and S. Marcel, “Counter-measures to photo attacks in face recognition: A public database and a baseline,” in *Proc. Int. Conf. on Biometrics (IJCB)*, 2011, pp. 1–7.

[97] J. Maatta, A. Hadid, and M. Pietikainen, “Face spoofing detection from single images using micro-texture analysis,” in *Proc. Int. Conf. on Biometrics (IJCB)*, 2011, pp. 1–7.

[98] J. Yang, Z. Lei, S. Liao, and S. Li, “Face liveness detection with component dependent descriptor,” in *Proc. Int. Conf. on Biometrics (ICB)*, 2013, pp. 1–6.

[99] L. Mei, D. Yang, Z. Feng, and J. Lai, “WLD-TOP based algorithm against face spoofing attacks,” in *Biometric Recognition*, ser. Lecture Notes in Computer Science, J. Yang, J. Yang, Z. Sun, S. Shan, W. Zheng, and J. Feng, Eds., 2015, vol. 9428, pp. 135–142.

[100] I. Buciu and S. Goldenberg, “Oscillating patterns based face antispoofing approach against video replay,” in *Proc. Int. IEEE Conf. on Automatic Face and Gesture Recognition (FG)*, vol. 02, 2015, pp. 1–6.

[101] J. Komulainen, A. Hadid, M. Pietikainen, A. Anjos, and S. Marcel, “Complementary countermeasures for detecting scenic face spoofing attacks,” in *Proc. Int. Conf. on Biometrics (ICB)*, 2013, pp. 1–7.

[102] J. Yan, Z. Zhang, Z. Lei, D. Yi, and S. Li, “Face liveness detection by exploring multiple scenic clues,” in *Proc. Int. Conf. on Control Automation Robotics Vision (ICARCV)*, 2012, pp. 188–193.

[103] N. Kose and J.-L. Dugelay, “Countermeasure for the protection of face recognition systems against mask attacks,” in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.

[104] T. Dhamecha, A. Nigam, R. Singh, and M. Vatsa, “Disguise detection and face recognition in visible and thermal spectrums,” in *Proc. Int. Conf. on Biometrics (ICB)*, 2013, pp. 1–8.

[105] Y. Wang, X. Hao, Y. Hou, and C. Guo, “A new multispectral method for face liveness detection,” in *Proc. Int. Conf. on Pattern Recognition (ACPR)*, 2013, pp. 922–926.

[106] X. Zhang, X. Zhang, J. Wang, X. Zhou, G. Qiu, L. Shen, and C. Feng, “Systems and method for facial verification,” US Patent US20 130 342 702 A1, 2013.

[107] L. Rakêt, L. Roholm, A. Bruhn, and J. Weickert, “Motion compensated frame interpolation with a symmetric optical flow constraint,” in *Advances in Visual Computing*, ser. Lecture Notes in Computer Science, G. Bebis, R. Boyle, B. Parvin, D. Koracin, C. Fowlkes, S. Wang, M.-H. Choi, S. Mantler, J. Schulze, D. Acevedo, K. Mueller, and M. Papka, Eds. Springer Berlin Heidelberg, 2012, vol. 7431, pp. 447–457.

[108] D. Sage, “Local normalization-filter to reduce the effect on a non-uniform illumination,” Online, 2011, biomedical Image Group, EPFL, Switzerland.

[109] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, “Adaptive histogram equalization and its variations,” *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355 – 368, 1987.

[110] R. Zabih and J. Woodfill, “Non-parametric local transforms for computing visual correspondence,” in *Proc. Europ. Conf. on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science, J.-O. Eklundh, Ed. Springer Berlin Heidelberg, 1994, vol. 801, pp. 151–158.

[111] B. Zitová and J. Flusser, “Image registration methods: a survey,” *Image and Vision Computing*, vol. 21, no. 11, pp. 977 – 1000, 2003.

[112] J. Kern and M. Pattichis, “Robust multispectral image registration using mutual-information models,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 5, pp. 1494–1505, 2007.

[113] R. Shams and N. Barnes, “Speeding up mutual information computation using nvidia cuda hardware,” in *Proc. Conf. on Digital Image Computing Techniques and Applications*, 2007, pp. 555–560.

[114] F. Steinbrücker, T. Pock, and D. Cremers, “Advanced data terms for variational optic flow estimation,” in *Vision, Modeling and Visualization*, vol. 1, 2009, pp. 155–164.

[115] M. Uřičář, V. Franc, D. Thomas, S. Akihiro, and V. Hlaváč, “Real-time multi-view facial landmark detector learned by the structured output SVM,” in *Proc.*

IEEE Int. Conf. on Automatic Face and Gesture Recognition Workshops (BWILD). IEEE, 2015.

[116] M. Velte, “Semantic image segmentation combining visible and near-infrared channels with depth information,” Master’s thesis, Bonn-Rhein-Sieg University of Applied Sciences, 2015.

[117] P. Azad, T. Gockel, and R. Dillmann, *Computer Vision*, 3rd ed. Elektor-Verlag, 2011.

[118] DIN, “Ergonomics - human body dimensions - part 2: Values (DIN 33402-2),” 2005.

[119] E. Parliament, “Richtlinie 2006/25/eg des europäischen parlaments und des rates über mindestvorschriften zum schutz von sicherheit und gesundheit der arbeitnehmer vor der gefährdung durch physikalische einwirkungen (künstliche optische strahlung),” Amtsblatt der Europäischen Union, April 2006.

[120] IEC, “Photobiological safety of lamps and lamp systems,” 2006.

[121] M. Atif, “Optimal depth estimation and extended depth of field from single images by computational imaging using chromatic aberrations,” Ph.D. dissertation, Universität Heidelberg, 2013.

[122] P. Trouvé, F. Champagnat, G. L. Besnerais, J. Sabater, T. Avignon, and J. Idier, “Passive depth estimation using chromatic aberration and a depth from defocus approach,” *Applied Optics*, vol. 52, no. 29, pp. 7152–7164, 2013.

[123] E. Krotkov, “Focusing,” *Int. J. of Computer Vision*, vol. 1, no. 3, pp. 223–237, 1988.

[124] A. P. Pentland, “A new sense for depth of field,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 4, pp. 523–531, 1987.

[125] J. R. Janesick, *Scientific charge-coupled devices*. SPIE press, 2001, vol. 83.

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

[127] D. Yi, Z. Lei, Z. Zhang, and S. Z. Li, *Handbook of Biometric Anti-Spoofing: Trusted Biometrics under Spoofing Attacks*. Springer London, 2014, ch. Face Anti-spoofing: Multi-spectral Approach, pp. 83–102.

[128] ISO, “Safety of machinery – positioning of safeguards with respect to the approach speeds of parts of the human body (ISO 13855),” 2010.

Acronyms & Abbreviations

BM	block matching. *See glossary.*
BRSU	Bonn-Rhein-Sieg University of Applied Sciences
C2C	channel to channel
C2R	channel to reference
CLAHE	contrast limited adaptive histogram equalization
FN	false negative
FNR	false negative rate
FP	false positive
FPR	false positive rate
FPS	frames per second
FRR	false rejection rate
FSWC	field-sequential waveband capturing. *See glossary.*
FWHM	full width at half maximum. *See glossary.*
GPU	graphics processing unit
ICM	inter-channel matching. *See glossary.*
IE	interpolation error
IFI	inter-frame interpolation. *See glossary.*
InGaAs	indium-gallium-arsenide
LBP	local binary pattern
LED	light emitting diode

NIR	near infrared. *See glossary.*
OF	optical flow. *See glossary.*
PSNR	peak signal to noise ratio
ROI	region of interest
SE	spectral error
SNR	signal to noise ratio
SSIM	structural similarity index metric
SVM	support vector machine. *See glossary.*
SWIR	short-wavelength infrared. *See glossary.*
TN	true negative
TNR	true negative rate
TP	true positive
TPR	true positive rate
VIS	visual spectrum. *See glossary.*

Glossary

Anti-spoofing

The term *anti-spoofing* refers to methods used to harden biometric recognition systems against attacks with counterfeit biometric features. 6, 36, 40, 61, 68, 77, 117, 120, 122, 124, 125, 135, *See also* spoofing attack

Bayer pattern

The *Bayer pattern* or Bayer filter is a color filter array that is directly mounted onto the surface of an image sensor with a specific pixelwise spatial distribution of green, blue and red color filters. This distribution mimics the sensitivity of the human eye. Color information of an image is recorded using neighboring pixels and interpolated in a process called *demosaiicing*. 2, 10, 33

Binary decision tree

Binary decision trees are used to classify new observations of input data by a series of binary decisions along the path from the root to a leaf node of a tree structure. Chapter 2.6.1 gives a detailed description of binary decision trees. 20, 59, 122, 134

Block matching

Block matching is a method to estimate motion between two images by dividing one image into so-called macro blocks with a given block size of several pixels and finding the best match for these blocks in the second image. Block matching typically only considers translational movement. 18, 19, 42, 167

Chromatic aberration

Image defects that are caused by dispersion of light in an optical system are called *chromatic aberrations*. As light with different wavelengths is refracted at different angles, a lens is not capable of focusing it on the same point of the image plane, which causes lateral chromatic aberrations which are strongest in the outer areas of the image plane. In addition, the focal length also varies with the wavelength of the light, which causes axial chromatic aberrations all over the image plane. 18, 89, 95

Cooperative user scenario

A *cooperative user scenario* describes a situation in which a subject cooperates with a face recognition system by removing any occlusions of his face such as glasses or headwear and looking into the direction of the camera, because he/she has a personal interest in being recognized correctly. 4, 6, 25, 61, 124, *See also* face recognition

Face recognition

Face recognition denotes the process of detecting a face in an image and matching it against a database of known ("enrolled") faces; see Section 2.7.1. 1, 3, 25, 27, 29, 32, 36, 40, 57, 61, 86, 124, 133, 134, 171

Face verification

Face verification denotes a specific operation mode of face recognition systems in which a presented "query" face image is compared to a known face of a person whose identity is being claimed; see Section 2.7.1. 2, 36, 40, 86, 117, 133, *See also* face recognition

FaceVACS

FaceVACS is a commercial off-the-shelf software for face recognition developed by Cognitec Systems GmbH. 28, 121, 127

FeGeb

The research project with the German title *Fälschungserkennung in der Gesichtsbiometrie (FeGeb)* deals with the detection of spoofing attacks on biometric face recognition systems. It is funded by the German Federal Ministry of Education and Research (BMBF) as part of the program “FHprofUnt” (FKZ: 03FH044PX3) and supported by the German Federal Office for Information Security (BSI). 2, 121

Field-seq. waveband capturing

Following a definition in the field of color imaging, the time-sequential acquisition of images in distinct wavebands that are combined into a multispectral image is denoted as *field-sequential waveband capturing*. 12, 40, 41, 85, 167

Full width at half max. (FWHM)

The *full width at half maximum* denotes the width of a function or spectrum curve measured between those points on the curve at which the function/spectrum reaches half of the maximum amplitude. 10, 167

Hyperspectral imaging

Hyperspectral imaging systems capture the spectral information of a scene with very high detail using a large number (often hundreds) of narrow wavebands through a wide spectral range. 9, 10, 33

Inter-channel matching (ICM)

In this work, *inter-channel matching* denotes a method for motion estimation and compensation for FSWC-based imaging systems that calculates displacement vectors between the different spectral channels of a multispectral image cube. As remission intensities of object surfaces might differ between the spectral channels, this method requires handling of intensity inconsistencies. 42, 49, 167

Inter-frame interpolation (IFI)

In this work, *inter-frame interpolation* denotes a method for motion estimation and compensation for FSWC-based imaging systems that calculates displacement vectors between corresponding spectral channels of two consecutive multispectral image cubes. This method avoids the need to handle intensity inconsistencies due to different remission intensities of object surfaces in different spectral channels. 42, 49, 167

Model tree

Model trees are used to predict continuous output values from new observations of input data using a linear model that is selected by following a series of binary decisions along the path from the root to a leaf node of a tree structure. Chapter 2.6.1 gives a detailed description of model trees. 21, 92, 115, *See also* binary decision tree

Motion compensation

The term *motion compensation* describes the process of estimating motion flow as observed by a camera between successive images and "correcting" the differences between the two images by applying an inverted motion (or, displacement) vector field on the later image. 6, 18, 34, 40, 41, 43, 85, 100, 101, 134

Multispectral imaging

Multispectral imaging systems capture the spectral information of a scene with higher detail than conventional single- or three-channel systems by using several distinct wavebands that are specifically selected for a given application. 2, 6, 9, 10, 12, 40, 133

Near infrared (NIR)

Electromagnetic radiation in the infrared (IR) spectrum within a wavelength range of approximately 750 nm to 1 400 nm is denoted as *near infrared* or IR-A, according to DIN and CIE. This wavelength range is situated in between the visual spectrum (VIS) and the short-wave infrared (SWIR) or IR-B spectrum. 2, 168, *See also* visual spectrum & short-wavelength infrared

Optical flow

The *optical flow* represents the velocity and direction of apparent motion at the image plane of a camera. Using two consecutive images, it can be estimated by finding displacement vectors between corresponding features. 18, 42, 100, 130, 168

Principal component analysis

The *principal component analysis* is a tool that allows to explore high-dimensional data by finding essential patterns that can serve as linear combinations to express the data with reduced dimensionality. 27, 118

Random forest

Random forests are used to classify new observations of input data by evaluating the individual classification results of multiple, randomly created binary decision trees. Chapter 2.6.2 gives a detailed description of random forests. 21, 59, 122, 134, *See also* binary decision tree

Short-wave infrared (SWIR)

Electromagnetic radiation in the infrared (IR) spectrum within a wavelength range of about 1400 nm to 3000 nm is denoted as the *short-wavelength infrared* or IR-B, according to DIN and CIE. This wavelength range is situated directly above the near infrared spectrum (NIR). The spectral signatures discussed in this work are arranged within the wavelength range of 900 nm to 1700 nm, which covers parts of both the NIR and SWIR spectra. As most researchers as well as camera manufacturers use only the term SWIR when describing this wavelength range in order to distinguish their research area or products from those that reach only up to $1\mu m$, this work adopts this simplification and uses only the term SWIR to describe this wavelength range. 2, 57, 168, *See also* near infrared

SkinCam

The active multispectral SWIR camera system developed in the context of this work is denoted as *SkinCam*, which stands for skin detecting camera. 69, 73, 95, 133

SPAI

The research project with the German title *Sichere Personenerkennung im Arbeitsumfeld von Industrierobotern (SPAI)* focuses on safety applications, especially in the field of industrial robotics. It aims on the use of multispectral SWIR imaging for a reliable detection of persons and their limbs in the working range of (possibly dangerous) robots. The project is funded by the Institute for Occupational Safety and Health of the German Social Accident Insurance (IFA). 2, 5, 88, 136

Spectral signature

The *spectral signature* is a vector of multispectral remission intensities. In this work, spectral signatures are used to classify an object's surface material as “skin” or “non-skin”. 2, 6, 10, 20, 31, 57–59, 133, 134

Spoof — In the context of biometric recognition systems, a *spoof* is a counterfeit biometric feature, such as a mask or photo used to attack face recognition systems or a fake finger used to attack fingerprint recognition systems. 1, 3, 36, 61, 120

Spoofing attack — The term *spoofing attack* denotes the attempt to trick a biometric recognition system by presenting a counterfeit biometric feature. 1, 3, 6, 30, 36, 57, 65, 89, 120, 133, 171, *See also* spoof

Support vector machine (SVM) — *Support vector machines* are supervised learning models that are used to classify data by constructing a high dimensional space that allows to separate different classes with a maximized margin between them. Chapter 2.6.3 gives a detailed description of SVMs. 59, 122, 134, 168

Visual (VIS) spectrum — The *visual spectrum* denotes electromagnetic radiation in the spectral range between approximately 380 nm and 780 nm that is visible to the human eye. 1, 13, 68, 168