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

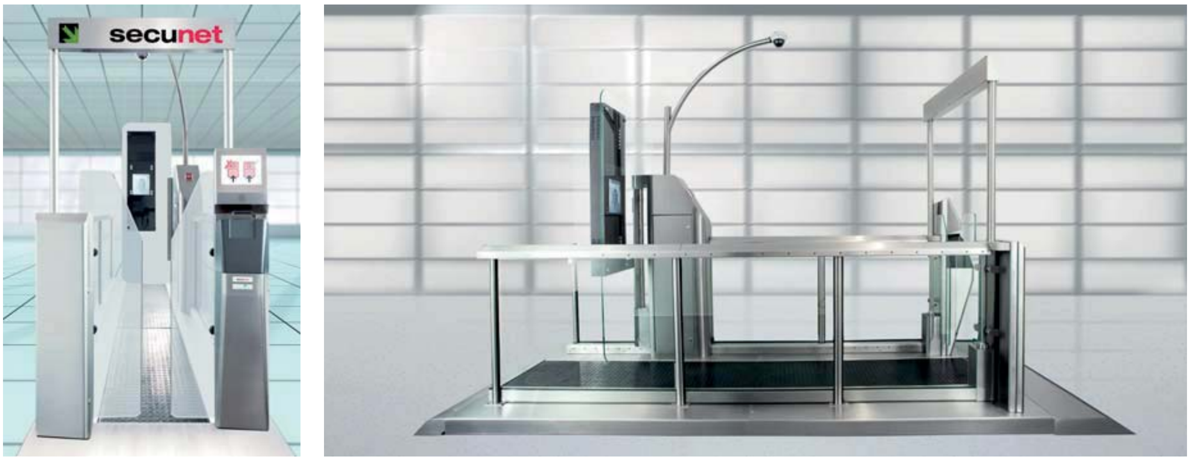

Figure 1.1: Example of an eGate system.
Images: secunet Security Networks AG

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-

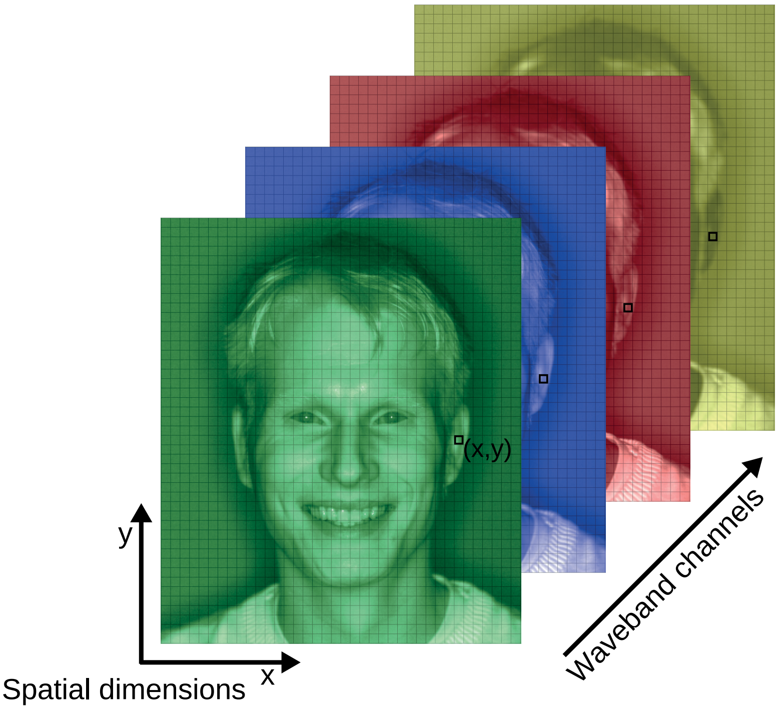

Figure 2.1: Illustration of a multispectral image cube. Waveband channels have been colored similarly to Figure 2.2 for illustration purposes. Based on [30, Fig. 1.1].

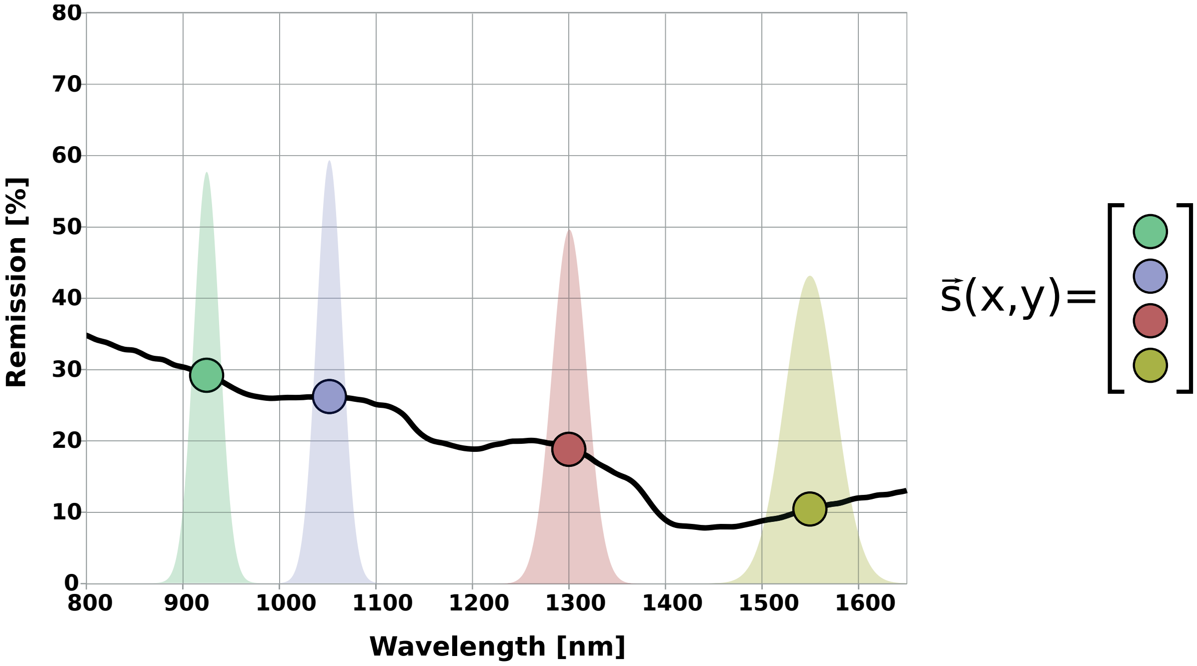

Figure 2.2: Example of a remission spectrum with indicated sensor sensitivity curves and the corresponding 4-dimensional spectral signature of pixel (x,y). Based on [30, Fig. 1.1].

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.

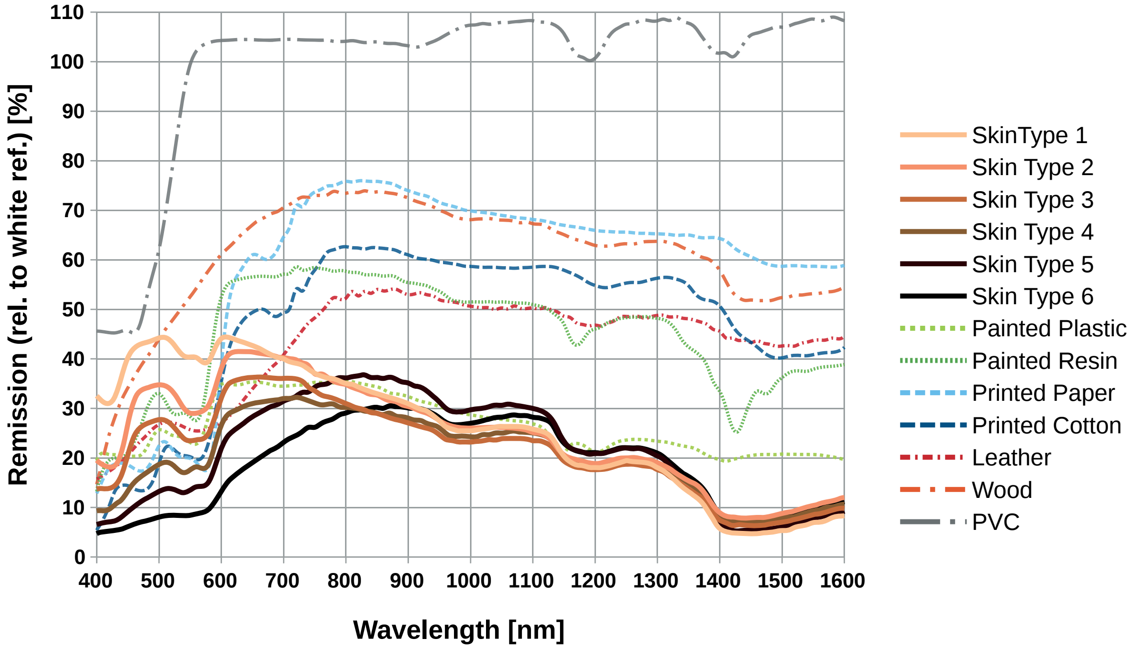

Figure 2.3: Spectral remission intensities of skin and different materials.

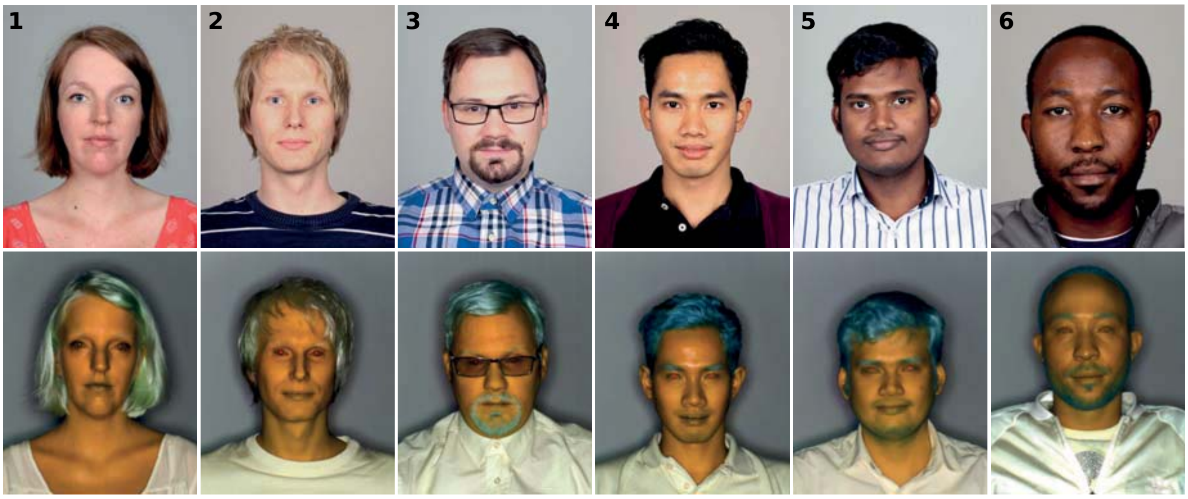

Figure 2.4: Visual spectrum (RGB color) and short wave infrared (false color) portrait images of skin types 1 to 6 according to Fitzpatrick [40].

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

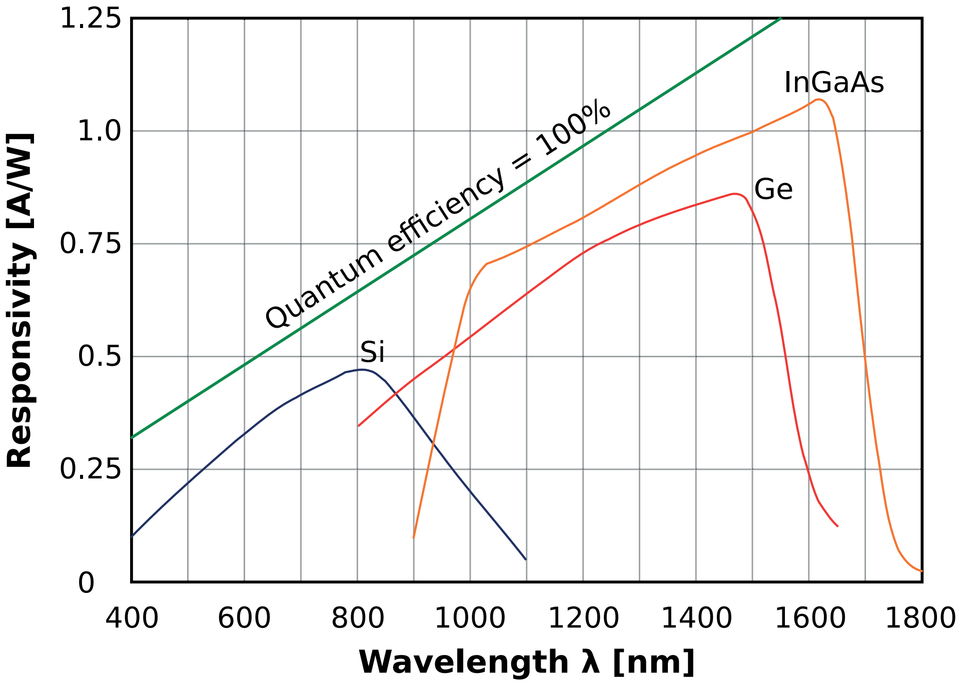

Figure 2.5: Spectral responsivity of photo detectors made of silicon (Si), germanium (Ge) and indium-gallium-arsenide (InGaAs). Adapted from [42, Fig. 1.70].

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:

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