Skin Detection and Melanin Estimation

Michael J. Mendenhall, *Member, IEEE,* Abel S. Nunez, Richard K. Martin, *Member, IEEE*

Abstract—f Skin detection is a well studied area in color imagery and is useful in a number of scenarios to include search and rescue and computer vision. Several approaches exist, but most focus on color imagery due to cost benefit and availability. Although many of thee visible-based approaches do well at detecting skin, they are plagued by relatively high false alarm rates in the urban environment (around 15%). We present novel algorithms for the remote skin detection and the estimation of its melanin content are presented. Our approach is derived from our previous efforts in modeling the interaction of light with human tissue, giving the advantage of allowing us to consider a nearly infinite number of subjects. Our approach results in 0.4% P_{FA} compared to 6.7-7.8% P_{FA} with color-based approaches at a fixed 95% P_D. Furthermore, at a 0.05% P_{FA}, we achieve an 82% P_D compared to 2.1-10.2% P_D for the color-based approaches. Our novel melanin estimation algorithm produces an absolute average error of 2.9% across a wide range of subjects.

Index Terms—skin detection, dismount detection, hyperspectral.

1 Introduction

Hyperspectral sensors provide a great deal of spectral granularity and a potential means for improved detection and classification of select materials [**?**]. Exploitation of these images has proven useful in identifying materials of interest from airborne platforms over a large geographic area to include geologic and biologic surface cover as well as locate anomalous materials such as air craft debris or in search and rescue (SAR) operations [**?**] [**?**] [**?**]. It is noted by experts that any hyperspectral system developed for use in SAR must be simple enough to operate for a non hyperspectral-exploitation expert [**?**], the system must be able to discriminate small targets in a large scene [**?**], and real-time processing is essential [**?**]. A hyperspectral/multispectral system designed to automatically detect skin and classify its pigmentation level could be a components of such a critical system. Beyond search and rescue, a system that detects skin and estimates its pigmentation level is useful in providing the requisite spatial discriminant and skin color information for facial and hand gesture recognition systems [**?**], [**?**], [**?**] and help address the difficulties in their automation caused by varying illumination levels [**?**].

This work presents efficient algorithms for detecting skin in an urban environment and suppressing typical false alarm sources. For pixels detected as skin, we further estimate the melanin content. Although the detection of skin is possible by observing tissue spectra, we build our detectors by exploiting an engineering model of human skin developed in [**?**]. Using a human skin model allows us to *quantitatively* describe the color of skin based on its primary chromophore, melanin. This gives our work a correspondence based on a physical model providing a sound theoretical mapping from image acquisition to skin detection and melanin estimation not described in other literature.

- *M.M., A.N., and R.M. are with the Department of Electrical and Computer Engineering, Air Force Institute of Technology.*

- *The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.*

2 Literature Review

Detection of human skin in color imagery is challenging as many materials have a color similar to one of the many shades of skin. Skin detection methods in color imagery vary from a ratio of color-space channels to more sophisticated machine learning methods such as the self-organizing map (SOM) [**?**]. Regardless of the methodology used, the end result is often a probability of detection (P_D) above 90% and a probability of false alarm (P_{FA})on the order of 15% [**?**].

Color-space channel methods typically use two channels. For example, the full range of skin colors has a red to green, red to blue, and green to blue ratio greater than one [**?**]. These methods produce a significant number of false alarms, and some have attempted to reduce them using rules approaches that combine ratios, color-space channel thresholds and differences [**?**] [**?**]. These methods essentially define a volume of the 3-dimensional color-space which encompasses all possible skin colors. Other methods reduce false alarms by examining how skin pixels cluster spatially and then attempt determine if the spatial clustering resembles a body part [**?**] [**?**].

Other skin detection methods use three-channel color-space examples as a training set to train a binary classification system. Some project pixels onto a plane within the color-space that provides the furthest separation between skin and non-skin pixels [**?**]. This technique in particular provides a slight improvement over the ratio-based methods.

The red-green-blue (RGB) color space is most common in the literature, but may not be ideal for the skin detection task where the primary disadvantage is the lack

of separation of luminance from chrominance and the strong correlation between the channels [**?**]. Color-spaces that separate luminance and chrominance often result in a better clustering of skin-colored pixels [**?**]. Although, the selection of a color-space may allow for a simpler or more intuitive algorithm, others have shown that no optimum three channel color-space for skin detection exists if the optimal skin detector for that color-space is used [**?**].

Some statistical-based approaches analyze training images for the probability of skin occurring given a pixel in the image, the probability of a skin color occurring given a pixel is skin, and the probability of skin color occurring in an image overall. From these quantities, the probability a pixel is skin given a specific skin color is calculated from the training images [**?**]. This general approach is susceptible to the "data problem" – too few training and testing images under various operating conditions to estimate the statistics accurately.

The same optical parameters that affect the color of human skin in the visible (VIS) affect its appearance in the near-infrared (NIR). A useful observation in skin reflectance is that it is high between 800-1100nm and low beyond 1400nm, which has been noted and exploited by others. Skin detection using two NIR channels is used for the purpose of counting occupants in a vehicle [**?**] and for face detection [**?**]. Both works use bands that are several hundred nanometers wide in the NIR.

The patent described in [**?**] exploits the absorptive and reflective properties of skin operating in the range of 800-1400nm for the lower wavelengths and 1400-2500nm for the upper wavelengths. Although not explicitly shown, the authors describe a scaled distance between the upper and lower wavelengths and threshold that scaled distance to declare the presence or absence of skin. Our work uses a similar detection scheme we call the normalized difference skin index (NDSI) [**?**] that carefully chooses *narrow* spectral bands of interest due both to the spectral properties of skin as well as the illumination source. We further incorporate additional spectral information to help reduce false alarm sources in the natural and urban environments yielding robust detectors.

3 Reflectance Measurements

3.1 Skin reflectance

An engineering model of human skin reflectance in the VIS/NIR based on the optical parameters of skin constituents is presented in [**?**] [**?**]. Constituents include: blood volume, oxygenated hemoglobin, epidermal and dermal thickness, collagen and water makeup in the epidermis, bilirubin, beta-carotene, and melanin. A comparison of measured skin reflectance of Type I/II and Type III/VI skin with model results are shown in Fig. 1(a) where the parameter values used in the model were constrained to ranges documented in the literature and adjusted to provide the best ℓ_2-norm fit [**?**]. Skin types are labeled by the Fitzpatrick scale where Type I/II/III/IV/V/VI are always/usually/sometimes/rarely/very rarely/never burns [**?**].

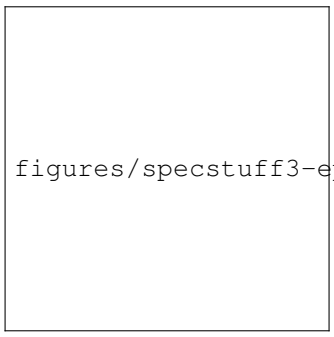
Although the model is not described in detail in this work, we introduce some notation to aid discussion later. Define measured skin reflectance as $\rho(\lambda)$ and modeled reflectance defined by the function $\tilde{\rho}_\lambda(M, Z)$, where λ indicates the dependency on wavelength, M is melanin percentage (by volume) in the epidermis, and Z is the set of remaining model parameters. We often use a "standard" model defined by Z equal to: a blood volume of 0.51% of the epidermis, oxygenated hemoglobin levels of 75% (based on the assumption that hemoglobin is 100% oxygenated leaving the lungs and 50% oxygenated entering the lungs), an epidermal thickness of 60μm, a dermal thickness of 1500μm, the epidermis containing 30% collagen (70% water), and subcutaneous fat with a maximum reflectance of 70% in the NIR. Melanin percentage of the epidermis is the parameter we often vary when generating reflectance spectra for various skin colors.

Several important features are noted in the spectra. First, as pigmentation level increases, the reflectance of skin decreases over the VIS and NIR. Second, as wavelength increases, the difference between the reflectance of skin with different pigmentation levels decreases and is due to melanin absorption decreasing as wavelength increases [**?**]. Beyond 1300nm, melanin absorption is not significant and skin reflectance is approximately the same [**?**]. In the VIS, hemoglobin significantly affects the spectrum [**?**] accounting for the w-shaped absorption feature around 570nm and the decreased reflectance in the VIS up to 600nm. Water absorption becomes significant in the NIR accounting for the reduced reflectance beyond 1150nm, the local maxima at 1080nm and 1250nm, and the local minima at 1200nm and 1400nm [**?**].

3.2 Reflectance measurements of false alarm sources

Several materials have colors similar to one of the wide varieties of skin-tones, some by design such as mannequins and dolls [**?**], others by coincidence such as brown cardboard, wood, leather, and some metals [**?**] [**?**] [**?**]. In some cases, the natural environment is rich in colors similar to skin such as a desert with various shades of brown, red, and yellow [**?**].

A comparison of Type I/II skin with the reflectance of a plastic flesh-colored doll is shown in Fig. 1(b) (dashed and solid lines respectively). Like skin, the reflectance of the flesh-colored doll rapidly increases as wavelength increases in the VIS. Beyond 1200nm, the reflectance of the flesh-colored doll is significantly greater than skin since the surface of the doll does not contain water and therefore lacks the water absorption characteristics in the NIR. A comparison of the reflectance of Type III/IV skin to brown cardboard is demonstrated in Fig. 1(b)



(a) Skin spectra (b) Skin and confuser spectra

Fig. 1. (a) Measured and modeled skin reflectance measurements. (b) Spectra of Type I/II and Type III/IV skin (dashed and dotted respectively) and spectra of a plastic doll and brown cardboard (solid and dash-dotted respectively).

(dotted line/dash-dotted respectively). Cardboard and Type III/IV skin exhibit an increase in reflectance as wavelength increases in the VIS. The reflectance of cardboard remains relatively high while the reflectance of the skin is much lower due to water absorption. (We further measured wet cardboard, and it does not exhibit the same absorption characteristics of skin.)

4 Feature Definitions

The efficacy of any detection algorithm is based on the quality of the received signal; since we are interested in detecting human skin under solar illumination, consideration must be given to the irradiance of the sun through the earth's atmosphere. The irradiance on a sunny day in Dayton, Ohio (scaled so its maximum value is one) is shown in Fig. 2(a) (solid line). The water vapor absorption bands, nominally at 1400nm and 1900nm (not shown), need to be avoided as there is no solar energy reaching the surface of the earth. The object of interest further imposes constraints on the spectra used in detection. As an example, a measurement of the reflected radiance of Type I/II skin under solar illumination, scaled by the same factor as the solar irradiance, is shown in Fig. 2(a) (dashed line). The local minima of the skin reflectance measurement corresponds to water absorption at approximately 950nm, 1150nm, and 1400nm, and 1600nm and beyond is dominated by water absorption. As such, the location of the local minima and maxima in the NIR of the reflected radiance of skin corresponds to the locations of the local minima and maxima of skin reflectance.

4.1 Normalized difference skin index (NDSI)

The NDSI is a function of reflectance at 1080nm and 1580nm. The reflectance at 1080nm is the location of a local maxima of the reflectance of skin in the NIR where melanin absorption dominates. Beyond 1080nm, water absorption in the skin becomes more significant until a local minima at approximately 1400nm (a known atmospheric water vapor absorption region). A stable, yet low valued, reflectance feature in skin spectra is noted at 1580nm (beyond atmospheric water vapor band).

Model-generated skin reflectance spectra in [**?**] shows that the difference in reflectance for the darkest to lightest skin types is fairly large at 1080nm versus 1580nm, which is consistent with reflectance measurements in the literature and in this article. Furthermore, according to the measured (and known theoretical) solar irradiance curves in Fig. 2(a), a significant amount of solar illumination power reaches the surface of the earth ensuring a strong signal-to-noise ratio at that longer wavelength. This ensures that the derivative is large between a melanin-dominated and water-dominated portion of the spectra. We define the normalized difference skin index (NDSI) as:

$$\gamma^i = \frac{\hat{\rho}^i_{\lambda_1=1080\text{nm}} - \hat{\rho}^i_{\lambda_2=1580\text{nm}}}{\hat{\rho}^i_{\lambda_1=1080\text{nm}} + \hat{\rho}^i_{\lambda_2=1580\text{nm}}} \quad (1)$$

where γ^i is the NDSI value for the i^{th} pixel. The normalized difference is immune to multiplicative power affects and is used frequently in the remote sensing community.

4.2 Normalized difference green-red index (NDGRI)

One can observe that human skin is more red then green, as indicated by Fig. 1, which shows the extremes of skin based on melanin content. To eliminate common false alarm sources, such as heavy water bearing vegetation (conifers) and water-bearing objects that are highly forward-scattering (e.g., snow and murky water), we invert the red-green relationship of skin with the normalized difference green-red index (NDGRI). The bands are chosen to correspond the red-green channels in the RGB color space and are 660nm and 540nm respectively. The NDGRI is defined as:

$$\beta^i = \frac{\hat{\rho}^i_{\lambda_1=540\text{nm}} - \hat{\rho}^i_{\lambda_2=660\text{nm}}}{\hat{\rho}^i_{\lambda_1=540\text{nm}} + \hat{\rho}^i_{\lambda_2=660\text{nm}}} \quad (2)$$

where β^i is the NDGRI value for the i^{th} pixel.

4.3 Features for melanin estimation

Estimating melanin levels of pixels detected as skin is based on a couple key skin reflectance features. As seen in Fig. 1, when melanin levels increases, the reflectance at 650nm decreases significantly while the reflectance at 1080nm decreases much less. The *near-infrared melanin index* (NIMI) in Eqn. (3) takes advantage of this relationship providing a mechanism to estimate the melanin content of skin. Work by Jablonski & Chaplin [**?**] identify melanin as the dominate chromophore at $\lambda = 685$nm and uses it to define skin color for indigenous people from different regions of the world. As such, we use the reflectance at 685nm and the reflectance at 1080nm since it is dominated by melanin absorption, but does not vary significantly compared to other wavelengths.

(a) Solar illumination (b) NIMI values

Fig. 2. (a) Solar irradiance scaled by the maximum irradiance (solid) and the radiance spectra of Type I/II skin illuminated by sunlight scaled by the same maximum irradiance (dashed). (b) Distribution of N (gray dots) using the sensor-plus-noise model. Each regression line is a fifth order polynomial: median (dashed),and "standard" model (solid).

We do not reuse 1580nm as our earlier experimentation indicates sensitivity to noise encountered in that region of the spectrum. The NIMI (N) is defined as:

$$N = \frac{\hat{\rho}_{\lambda=685\text{nm}}}{\hat{\rho}_{\lambda=1080\text{nm}}}. \tag{3}$$

The lines in Fig. 2(b) are based on the regression of the median (dashed, Eqn. 4) and "standard" person (solid, Eqn. 5). The estimated melanin level (in %) from the NIMI (N) is denoted as $\hat{M}$ for both median (md) and standard person (sp) as:

$$\begin{aligned}\hat{M}_{\text{md}} = &-106.72N^5 + 492.73N^4 - 912.71N^3 \\ &+880.39N^2 - 489.85N + 139.83,\end{aligned} \tag{4}$$

$$\begin{aligned}\hat{M}_{\text{sp}} = &-178.86N^5 + 737.05N^4 - 123.49N^3 \\ &+108.42N^2 - 54.96N + 144.49.\end{aligned} \tag{5}$$

NIMI values computed using model spectra versus the melanin content used as the model parameter are shown in Fig. 2(b). Model parameters (M, Z) are varied within their biologically feasible values [**?**]. Due to the variation of the parameter space, there are multiple NIMI values that map to a single melanin level. The natural distribution of the parameter set is unknown, only their ranges. As such, we assume they are uniformly distributed.

4.4 Extending features to an arbitrary imager

The algorithms described previously are based on having perfect knowledge of the reflectance of human skin and were generated using diffuse modeled and measured reflectance, which do not account for specular reflection. The uncertainty in atmospheric correction, sensor noise, and specular reflection affects the estimated reflectance from image data. A signal-plus-noise model is used to generate estimated reflectance according to:

$$\hat{\rho}_\lambda = \tilde{\rho}_\lambda\left(M, Z\right) + s_\lambda + n_\lambda \tag{6}$$

where $\hat{\rho}_\lambda$ is the estimated image spectra at wavelength λ, $\tilde{\rho}\left(\cdot\right)$ is the diffuse model-generated spectra, s_λ is a specular reflection term where $4\% \leq s_\lambda \leq 14\%$, and n_λ is a noise term distributed as $N\left(0, \sigma_\lambda^2\right)$. The diffuse skin spectra and the noise components of the sensor-plus-noise model are relatively easy to acquire. However, the specular reflection component is much more difficult. This is largely due to a lack of available data in the NIR portion of the spectrum. Existing works characterize specular reflection in the VIS [**?**], but often for the entire visible spectrum (monochromatically) due to the difficulties in obtaining the specular component at multiple (hundreds) of wavelengths. Although we know specular reflection is wavelength-dependant, we treat it as wavelength *independent*. A second issue with specular reflection is that it is often measured in sensor-reaching radiance and not transformed to reflectance space, which is where the current work exists. As such, we use observation of the hyperspectral data from the sensor to estimate *reasonable* specular components where we assume that specular component is not wavelength dependant. The sensor noise component is spectrometer-dependant and is assumed to be the noise term in estimated reflectance (that is, after atmospheric correction). We compute this term for each portion of the spectra used in this work.

Given wavelength-dependant noise and wavelength independent specular reflection, the NDSI in Eqn. 1 is rewritten as:

$$\gamma^i = \frac{\left(\hat{\rho}^i_{\lambda_1} + s^i_{\lambda_1} + n^i_{\lambda_1}\right) - \left(\hat{\rho}^i_{\lambda_2} + s^i_{\lambda_2} + n^i_{\lambda_2}\right)}{\left(\hat{\rho}^i_{\lambda_1} + s^i_{\lambda_1} + n^i_{\lambda_1}\right) + \left(\hat{\rho}^i_{\lambda_2} + s^i_{\lambda_2} + n^i_{\lambda_2}\right)}. \tag{7}$$

By collecting diffuse reflectance, noise, and specular reflection terms together, Eqn. 7, is rewritten as:

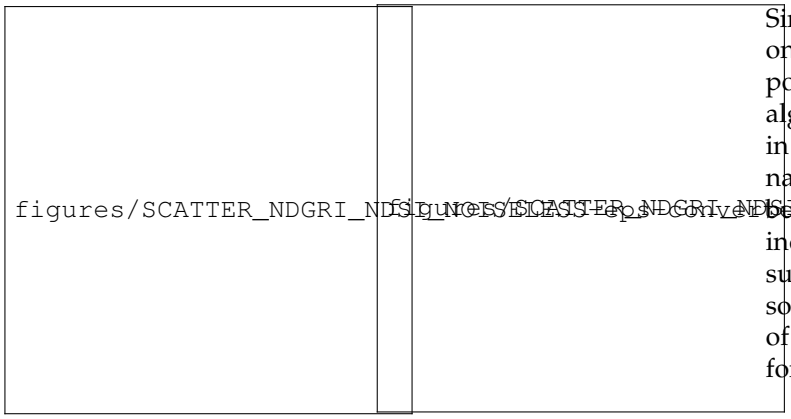
$$\gamma^i = \frac{\left(\hat{\rho}^i_{\lambda_1} - \hat{\rho}^i_{\lambda_2}\right) + \left(n^i\left(\sigma^2_{\lambda_1}\right) - n^i\left(\sigma^2_{\lambda_2}\right)\right)}{\left(\hat{\rho}^i_{\lambda_1} + \hat{\rho}^i_{\lambda_2}\right) + \left(n^i\left(\sigma^2_{\lambda_1}\right) + n^i\left(\sigma^2_{\lambda_2}\right)\right) + 2c}. \tag{8}$$

If $n^i\left(\sigma^2_\lambda\right)$ is distributed as $N\left(0, \sigma^2_\lambda\right)$, then:

$$E[\gamma^i] = \frac{\left(\hat{\rho}^i_{\lambda_1} - \hat{\rho}^i_{\lambda_2}\right)}{\left(\hat{\rho}^i_{\lambda_1} + \hat{\rho}^i_{\lambda_2}\right) + 2c}. \tag{9}$$

As indicated in Eqn. 9, a significant amount of specular reflection can significantly lower a pixel's (NDSI,NDGRI) values.

Fig. 3(a) shows that (NDSI,NDGRI) values for modeled and measured skin cluster in the same small area in the top left quadrant where $0.6 \leq NDSI \leq 0.8$ and $-0.4 \leq NDGRI \leq -0.05$. However, specular reflection is an issue and affects detection negatively if not accounted for. Adding uniformly distributed specular reflection of $4\% \leq s_\lambda \leq 14\%$ and sensor noise to the skin samples and recomputing the (NDGRI,NDSI) pairs is shown Fig. 3(b). Included in Fig. 3(b) are (NDGRI,NDSI) pairs



(a) Noiseless data (b) Noisy data

Fig. 3. (a) Distribution of (NDSI,NDGRI) features from model-generated and measured skin spectra, and common urban background materials. (b) Distribution of skin samples using the signal-plus-noise model applied to model-generated and measured skin spectra compared to that of imager obtained spectra. In both (a) and (b), features from model-generated data are black dots and spectrometer measurements are light gray '+'s. In (a), features from false alarm sources are dark gray circles and in (b) features from imager obtained spectra are dark gray circles.

computed from human skin measured with a hyperspectral imager. The results demonstrate two important points. First, the effects of sensor noise and specular reflection dramatically spread the distribution of the features. Second, the sensor-plus-noise model reasonably approximates the distribution of measured data. The disparity of the distributions can be attributed primarily to two causes. First is that the "truthing" of the imaged subjects includes boundary pixels that are polluted by non skin material such as the following pairs: (forehead,hat), (cheeck,background), and (face,eyes). A second source of error is the specular reflection component, which we model as wavelength independent, but in fact is wavelength dependant. A visual comparison of the distributions of the signal-plus-noise model (Eqn. 8) and the (NDGRI,NDSI) pairs from a hyperspectral imager shows a good match.

4.5 Features In Color Imagery

Skin detection methods for color imagery rely on the fact that skin is more red than green or blue and that skin retains its unique color regardless of brightness [**?**]. Many skin detection methods exploit this feature by projecting RGB values onto a set of components that separate brightness from chrominance. One such color space is normalized RGB, in which the red, green, and blue components are normalized as in Eqn. 10 to represent their percentage contribution to the pixel color:

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B}. \tag{10}$$

Since the three normalized components (r, g, b) sum to one, the pixels can be represented by the (r, g) components to reduce dimensionality. The skin detection algorithm in [**?**] observes that skin pixels are distributed in a crescent-shaped locus in rg-space. As the illumination source varies, the cluster of skin pixels shifts, but stays within the locus. Also, because of intensity independence, skin pixels appear relatively invariant to surface orientation of the skin relative to the illumination source in color imagery. This well-defined concentration of skin pixels makes the rg-space a good representation for skin pixel detection.

Another popular color space for representing skin features is HSV. The components of HSV are derived from RGB by

$$\begin{aligned} H =& \arccos \frac{\frac{1}{2}\left((R-G)+(R-B)\right)}{\sqrt{\left((R-G)^2+(R-B)(G-B)\right)}} \\ S =& 1 - 3 \times \frac{\min\{R,G,B\}}{R+G+B} \\ V =& \frac{1}{3}(R+G+B) \end{aligned} \tag{11}$$

where hue represents the dominant color and saturation is the proportion of dominant color to value, which corresponds to brightness. Dropping V, the HS-space represents skin well, as skin has a characteristic hue near red, a low saturation level, and is invariant to V.

The YC_bC_r color space is similar to HSV in that it projects RGB pixels into luminosity and chromaticity components [**?**]. The luminosity component (Y) is a weighted sum of RGB values, and the blue (C_b) and red (C_r) chrominances are the differences between luminosity and RGB blue and red components.

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ C_b &= B - Y \\ C_r &= R - Y \end{aligned} \tag{12}$$

In addition to being easy to compute, the C_bC_r components are a robust representation of skin color under different lighting conditions [**?**] [**?**]. Skin remains clustered in an ellipse in C_bC_r-space when taken from both shadowy and well-lit areas of an image. The face detection method in [**?**] identifies skin pixels by using an ellipse-shaped boundary in C_bC_r-space.

4.6 Skin Seperability From False-Alarm Sources in the VIS and NIR Feature Spaces

The VIS feature spaces described in Section 4.5 provide tight skin clustering, but not very good separability between skin and false-alarm sources. Fig. 4 demonstrates the distribution of skin in the above feature spaces. These points are taken from HST3 imager data and RGB data of the same scene that have been manually separated into skin and non skin classes. The skin distribution in NDGRI,NDSI feature space appears similar to the

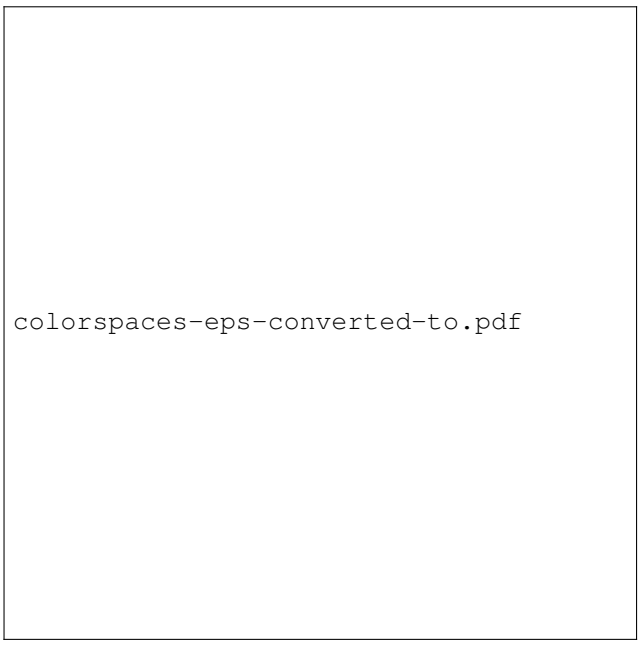


Fig. 4. Distribution of skin (black) and false-alarm sources (gray) in the feature spaces from Sections 4.1/ 4.2 and 4.5 from HST3 imager data and RGB data of the same scene. The NDSI,NDGRI feature space (top left) has better separability between skin and false-alarm sources than normalized rg-space (top right), C_bC_r-space (bottom left), and HS-space (bottom right, shown in cartesian coordinates).

signal-plus-noise model in Fig. 3b, and the false-alarm sources are clustered like the common urban background materials in Fig. 3a. There is little crossover between the two distributions. In the VIS feature spaces, however, the distinction between skin and false-alarm sources is less discernable. Skin is well represented in dense distributions in rg-space, C_bC_r-space, and HS-space, but the non skin distributions have a large amount of overlap with skin. This causes high false alarm rates for rules-based detectors in VIS-based feature spaces.

5 Skin detection algorithms

5.1 Rules-based skin detection algorithm

Given an understanding of human skin reflectance through modeling theory and measurements, one can define a detector by a set of rules:

$$S_i = \begin{cases} 1 & \text{if } b_1 \leq \beta^i \leq b_2 \text{ and } c_1 \leq \gamma^i \leq c_2 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

The rules then define a rectangle that bounds the 2D (NDGRI,NDSI) space.

The advantage of the detector described here is the dependence solely on the extremes in skin spectra (measured or modeled). Given the availability of the model in [**?**], the diffuse spectra are generated with a high degree of confidence. Two primary limitations of this approach are that it does not take into account information on potential false alarm sources beyond the design of the normalized difference indices, and it ignores the distribution of the target and false alarm sources and therefore lacks optimality in terms of minimizing the Bayes risk.

The detector described in Eqn. 13 produces a rectangular decision region. In order to generate a receiver operating characteristic curve (ROC), one would sweep over (β, γ) yielding a 2D ROC surface. In the evaluation of the detector, we fix one parameter at one of several operating points and vary the second to generate several ROC curves to demonstrate the affect the two parameters have on performance.

5.2 Skin detection using the likelihood ratio test

An optimal detector, one that minimizes the Bayes Risk, is used as a comparison to the rules detector. We choose the likelihood ratio test defined as:

$$\Lambda_\Theta(\theta) = \frac{\hat{f}_1(\theta)}{\hat{f}_0(\theta)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad (14)$$

where H_0 is the hypothesis that the sample is not skin, H_1 is the hypothesis that the sample is skin, $\hat{f}_0(\theta) = P[\Theta = \theta|\text{not skin}]$, $\hat{f}_1(\theta) = P[\Theta = \theta|\text{skin}]$, $\Theta = \{B, \Gamma\}$, $\theta = \{\beta, \gamma\}$ are sets of parameters based on the (NDGRI,NDSI)-based detector, $\hat{f}_1(\theta)$ is the estimated probability density function (*pdf*) of human skin, and $\hat{f}_0(\theta)$ is the estimated *pdf* of the false alarm sources.

The functional forms of $\hat{f}_1(\theta)$ and $\hat{f}_0(\theta)$ are estimated by Gaussian mixture models (GMMs) parameterized using Expectation Maximization [**?**] such that

$$\hat{f}_j(\theta) = \sum_{k=1}^{K_j} \pi_{j,k} N\left(\underline{\mu}_{j,k}, \underline{\Sigma}_{j,k}\right), j \in \{0, 1\} \quad (15)$$

where K_j is the number of Gaussians utilized to estimate $\hat{f}_j(\theta)$, $\pi_{j,k}$ is the weighted value of each Gaussian such that $\pi_{j,k} \in [0, 1]$ and $\sum_{k=1}^{K_j} \pi_{j,k} = 1$. The parameters of each Gaussian are represented by mean vector $\underline{\mu}_{j,k}$ and covariance matrix $\underline{\Sigma}_{j,k}$. The likelihood ratio represents a 2D decision surface.

The skin model described in Section 3.1 is used to generate samples to compute $\hat{f}_1(\theta)$. This makes the implicit assumption that all normal skin types are equally probable and that the specular reflection component is distributed uniformly on $[4\%, 14\%]$. The USGS spectral library [**?**] augmented with measurements with a handheld spectrometer are used to generate $\hat{f}_0(\theta)$.

6 Results

6.1 NIR skin detection on model and laboratory spectra

We first present the results of the rules and LRT-based detectors on the combination of modeled human skin data and data from the USGS spectral library [**?**] data

augmented with field samples collected by the authors using a handheld spectrometer. Modeled skin data is modified as described earlier using the signal-plus-noise model described in Eqn. 6 with sensor noise parameters: INSERT SENSOR NOISE VARIANCES. USGS spectral library and field sample data are modified with the estimated sensor noise only.

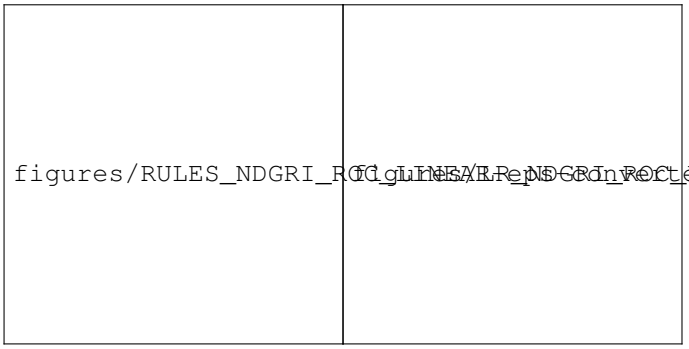
The results presented in Fig. 5 and summarized in Table 1 are an aggregate of 20 noise realizations where each noise realization is further subject to K-Fold cross validation (for K=5 [**?**]). The average performing ROC curve is the mean of the 100 simulations (5 cross validation runs $\times$ 20 noise realizations).

Results of the detectors are presented as ROC curves in Fig. 5. The rules detectors for the (NDGRI,NDSI) pair are presented in Fig. 5(a) where the values for the NDGRI are $\beta = \{-0.02, -0.05, 0.1, 1\}$, and the NDSI threshold varies as $-1 \leq \gamma \leq 0.93$ (where 0.93 is an experimentally determined upper bound). For $\beta = 1$, the detector becomes an NDSI-based detector only and provides a relative comparison between skin detection only and skin detection with false alarm suppression. Results of the LRT-based detector for the (NDGRI,NDSI) pairs is presented in Fig. 5(b).

Neither the rules nor the LRT detector ROC curves are strictly concave down. In the rules detector case, this is likely due to the fact that it is not optimal for minimizing the Bayes risk. In the LRT detector case, this is likely due to our assumption that a GMM adequately represents the true distribution of target and non-target samples when in fact this assumption does not likely hold true.

The error bars depicted in Fig. 5 represent $\mu \pm \sigma$ in the P_D and P_{FA} directions respectively. This is done at arbitrary points along each ROC curve to illustrate the performance envelope. In general, variance in the P_{FA} direction is worse than in the P_D direction. This is intuitive since there is more variation in the non-skin class than the skin class. Furthermore, the P_D and P_{FA} variance is greater for the LRT detector than for the rules detector as we use K-fold cross validation for the LRT detector. Conversely, the rules detector does not change between folds, only the test set applied to it.

Specific operating points (OPs) drawn from the ROC curves in Fig. 5 for a constant $P_{FA} = 0.05\%$ and constant $P_D = 95\%$ are shown in Table 1. Complimentary OPs (C-OPs) are the minimum, average, and maximum values for the best average performing ROC curve where a C-OP is the corresponding P_D (P_{FA}) for a P_{FA} (P_D) OP. In the case we are using the rules detector, we consider the best average performing curve over one of four detector regions ($\beta \in [b_1, b_2]$). The NDSI threshold, γ, is varied over the range [-1,0.93] (where 0.93 is an experimentally determined upper bound). In the case of the LRT detector, we use the average of all 100 results where models are recomputed for each fold in the cross validation for each of the noise realizations. The summary in Table 1 indicates that for a $P_D = 95\%$, the rules and LRT detectors perform in a similar manner with the exception of the maximum error where the rules has a lower P_{FA}.



(a) Rules-based detector (b) LRT detector

Fig. 5. The ROC curves are for the modeled skin data and spectral library false alarm source data. The vertical dashed line represents a constant $P_{FA} = 0.05\%$ while the horizontal dashed line represents a constant $P_D = 95\%$. (a) ROC curve using the rules detector where $-1 \leq \gamma \leq 0.93$ and $-1 \leq \beta \leq \{-0.02, -0.05, -0.1, 1.0\}$ yielding four detector regions ({solid, dashed, dashed-dotted, dotted} respectively). (b) ROC curve using the LRT detector varying $0 \leq \eta \leq 5 \times 10^6$ (dashed lines are typical best and worst case ROC curves while the solid line is the mean ROC).

TABLE 1
NIR detector results for model and laboratory data.

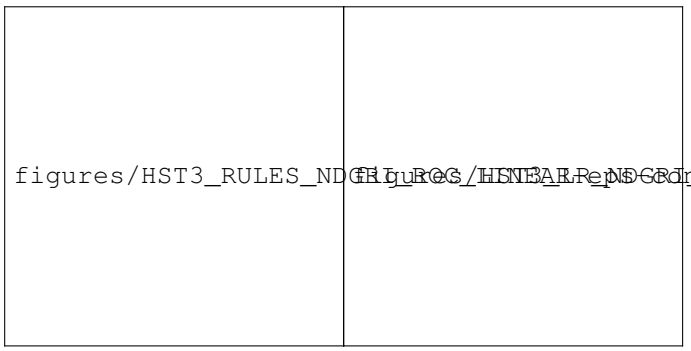
Det	OP (%)	C-OP (%)	P	Val{upr,lwr}
Rules	$P_D = 95$	0.7/0.8/0.9	β	$\{-1, -0.05\}$
			γ	$\{0.38, 0.93\}$
Rules	$P_{FA} = 0.05$	2.2/4.6/11.9	β	$\{1, 0.05\}$
			γ	$\{0.86, 0.93\}$
LRT	$P_D = 95$	0.8/0.9/1.4	η	$\{4, 8\}$
LRT	$P_{FA} = 0.05$	0/0.001/29.7	η	$\{1.05 \times 10^{-5}\}$,40 }

6.2 NIR skin detection on hyperspectral imagery

6.2.1 *Hyperspectral test imagery*

Data for this test were collected with the SpecTIR Hyper-SpecTIR Version 3 (HST3) Hyperspectral Imager [**?**]. The HST3 collects data in the range of 400nm – 2500nm. The spectral bandwidth is nominally 12nm in the VIS and 8nm in the NIR. Images are transformed into estimated reflectance using empirical line correction (ELC).

To test the skin detection algorithms, four images are collected with skin color confusers and skin with various levels of pigmentation as in Fig. 8(top). The image contains typical color-based skin detection confusers including flesh-colored doll, cardboard, red brick, leather boot, and pieces of wood. A branch from a conifer (from the yew family) is included in the scene as it tends to have a high NDSI value. The scene is a suburban environment with houses, streets, sidewalks, trees, grass, bushes, and other assorted materials. Skin truth pixels are identified as white pixels in Fig. 8(middle).



(a) Rules-based detector (b) LRT detector

Fig. 6. The ROC curves for a set of hyperspectral images similar to that of Fig. 8(top). (a) ROC curve for the rules detector varying $-1 \leq \gamma \leq 0.93$ and fixing the upper bound on NDGRI $-1 \leq \beta \leq \{-0.02, -0.05, -0.1, 1.0\}$ yielding four detector regions ({solid, dashed, dashed-dotted, dotted} respectively). (b) ROC curve for the LRT detector varying $0 \leq \eta \leq 5 \times 10^6$ (dashed lines are typical best and worst case ROC curves while the solid line is the mean).

Due to the noise inherent in the system/environment and the fact that the bands selected for our algorithms do not line up with the HST3 band centers, the NDSI and NDGRI algorithms are modified to accommodate the available spectra. The algorithms are implemented with the mean of the estimated reflectance of the three HST3 bands closest to the algorithms' band centers. For example, the estimated reflectance at 540nm used for the NDGRI algorithm is implemented using the mean of the estimated reflectance at 531.37nm, 542.74nm, and 554.06nm.

6.2.2 NIR detector results

The ROC curves for the rules and LRT-based detectors on the hyperspectral image data are presented in Fig. 6. (In the case of the image data, ROC curves are concave down.) For the rules detector, the same four detector regions used in Section 6.1 are used to generate the detection results on the hyperspectral image data. Similarly, the 100 detectors used to generate the detector results for the LRT detector described in Section 6.1 are used on the hyperspectral image data.

Overall, the rules detector outperforms the LRT detector for the image data. This may be attributed to one of several reasons: fewer false alarm types exist in the image data versus the spectral library data; a bias may exist in the skin reflectance model that works favorably on the image data; the rules method is better tuned to the hyperspectral image data.

Consistent with the previous analysis, specific OPs are drawn from the ROC curves in Fig. 6 for a constant $P_{FA} = 0.05\%$ and constant $P_D = 95\%$ and are shown in Table 2. Complimentary OPs are provided for the minimum, average, and maximum values attained for the best average performing ROC curve. For the rules detector, we consider the best average performing curve over one of four detector regions where each detector region is specified by an upper and lower bounds the NDGRI thresholds ($\beta \in [b_1, b_2]$). The NDSI threshold, γ, is varied over the range [-1,0.93] (where 0.93 is an experimentally determined upper bound). For the LRT detector, we use the average of all 100 results where models are recomputed for each fold in the cross validation for each noise realization.

TABLE 2
NIR detector results for image data.

Det	OP (%)	C-OP (%)	P	Val{upr,lwr}
Rules	$P_D = 95$	0.4/0.4/0.4	β	$\{-1, -0.02\}$
			γ	$\{0.26, 0.93\}$
Rules	$P_{FA} = 0.05$	82.0/82.0/82.0	β	$\{-1, -0.20\}$
			γ	$\{0.41, 0.93\}$
LRT	$P_D = 95$	0.4/0.4/0.5	η	$\{0.034, 0.022\}$
LRT	$P_{FA} = 0.05$	77.2/77.6/78.8	η	$\{3, 4\}$

The summary in Table 2 indicates that for a $P_D = 95\%$, The rules and LRT detectors perform in a similar manner with the exception of the maximum error where the rules has a lower P_{FA}. For a $P_{FA} = 0.05\%$, the rules detector consistently produces a higher P_D.

6.2.3 Comparison of NIR and VIS Skin Detection Performance

Figure 7 shows ROC curves for the LRT-based detector in NDGRI,NDSI feature space compared to those of LRT-based detectors in rg-space, RGB-space, HS-space, and C_bC_r-space. The ROC curves are generated from the HST3 image data and RGB data that is plotted in Fig. 4. Estimates of distributions $\hat{f}_1(\theta)$ and $\hat{f}_0(\theta)$ for skin and non skin are determined by fitting GMMs to both classes with Expectation Maximization as in Section 5.2. The likelihood ratio in Eqn. 14 is then calculated for each pixel in the image and compared to a threshold η. The threshold η starts at zero, where all pixels are classified as not skin, and is increased in small increments until all pixels are classified as skin. The P_D and P_{FA} for each threshold are plotted to form the ROC curves in Fig 7.

The LRT-based detector in the NDGRI,NDSI feature space outperforms the LRT-based detectors in all four of the tested VIS feature spaces. Specific OPs and areas under the ROC curves in Fig. 7 are listed in Table 3. The NDGRI,NDSI feature space has greater separation between skin and false alarm sources, leading to a much lower P_{FA} for $P_D = 95\%$ and a much higher P_D for $P_{FA} = 0.05\%$.

6.3 Skin color estimation

We use Fig. 8(top) as the test scene for skin color estimation. Subjects $\{1, 2, 4\}$ are Type V/VI skin, subjects $\{3, 6\}$ are Type III/IV skin, and subjects $\{5, 7\}$ are Type I/II skin [**?**].

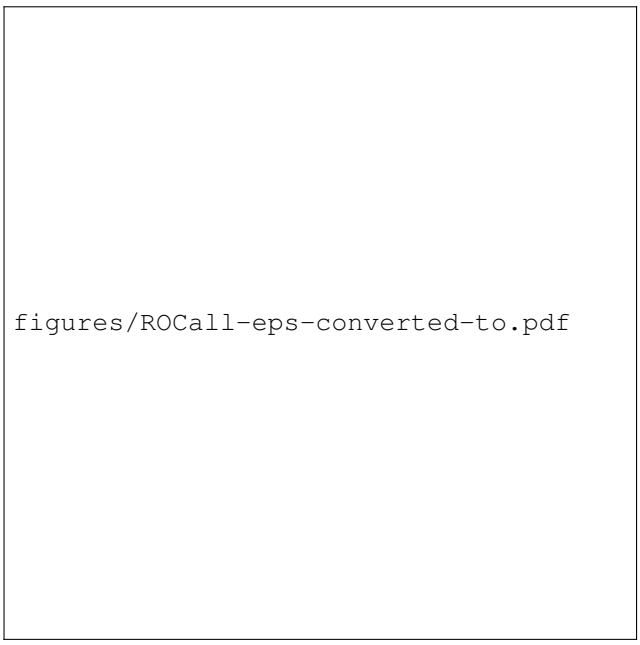


Fig. 7. Left: ROC curves for LRT-based detectors in NDGRI,NDSI feature space (thick solid line), rg-space (thin solid line), RGB-space (dashed line), C_bC_r-space (dotted line), and HS-space (dashed-dotted line) skin detection performance on imager data plotted in Fig. 4. Right: The same ROC curves plotted along log(P_{FA}) to show the improved skin detection performance in the NIR feature space over VIS feature spaces.

Feature space	OP (%)	C-OP (%)	AUC
NDGRI,NDSI	$P_D = 95$ $P_{FA} = 0.05$	$P_{FA} = 0.22$ $P_D = 88.05$	0.9984
RGB-space	$P_D = 95$ $P_{FA} = 0.05$	$P_{FA} = 7.27$ $P_D = 10.21$	0.9758
rg-space	$P_D = 95$ $P_{FA} = 0.05$	$P_{FA} = 7.47$ $P_D = 4.71$	0.9759
C_bC_r-space	$P_D = 95$ $P_{FA} = 0.05$	$P_{FA} = 6.74$ $P_D = 1.04$	0.9759
HS-space	$P_D = 95$ $P_{FA} = 0.05$	$P_{FA} = 7.76$ $P_D = 2.08$	0.9706

TABLE 3
LRT-based detector results for image data compared to VIS-based skin detectors. The NIR feature space skin detector has a higher AUC and better OPs than the VIS feature space skin detectors.

Qualitative results are shown for melanin estimates for the "standard" person regression of Eqn. 5 ($\hat{M}_{\text{sp}}$) are shown in Fig. 8(bottom). From a qualitative perspective, the estimate of the melanin from the $\hat{M}_{\text{sp}}$ is reasonable.

The challenge with this estimation problem is the difficultly in assessing true melanin volume in living tissue (based on our collaboration with pathologists, it is not clear to the authors that this is feasible with extracted tissue). Since we cannot readily acquire truth from living or deceased subjects, we assume the model is accurate and able to provide us with a reasonable estimate truth.

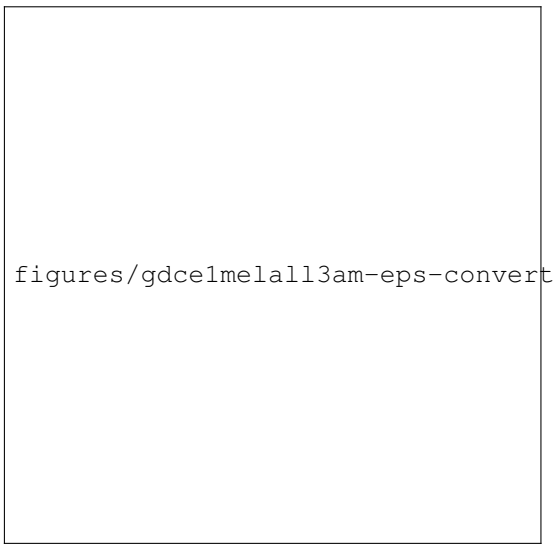


Fig. 8. (Top) Test scene with test subjects covering the range of Fitzpatrick skin types. (Middle) Skin mask. (Bottom) Estimated melanin level based on the regression for "standard" person.

This is accomplished using a diffuse measurement with a handheld reflectometer and adjusting the model parameters for the best ℓ_2-norm fit (this is demonstrated in Fig. 1(a)). We assume that the estimate from the model fitting is approximately correct [**?**] and use this value to compare with both melanin estimation results.

Overall, the observations of the error between the $\hat{M}_{\text{sp}}$ and $\hat{M}_{\text{md}}$ suggests that the subject are near the median and standard person NIMI values. This is indicative of the observed error values compared to the spread of the NIMI values when computed based on the signal-plus-noise model of Eqn. 8 for the two respective NIMI feature points (this spread is shown as gray $\cdot$'s in Fig. 2 in comparison to the median and standard person regression lines of that same figure.) General observations from Table 4 concludes that the $\hat{M}_{\text{sp}}$ performs better than $\hat{M}_{\text{md}}$, with the exception of two of the three Type V/VI skin types (subjects 1 and 2). Interestingly, $\hat{M}_{\text{sp}}$ performs better than $\hat{M}_{\text{md}}$ for the darkest of the Type V/VI skin types and both of the Type III/IV skin types, but not the two Type V/VI skin types between them.

The sources of error are likely due to addition of sensor noise and specular reflection, but are also inherent in the variability amongst the reflectance of human tissue due to variations in the components that affect the reflectance. Furthermore, the human skin reflectance model in [**?**] is not perfect. There are likely biases in the error that appear both in the regression coefficient for the median and standard person regressions and the estimated melanin level extracted from the best ℓ_2-norm fit of the model to the handheld spectrometer reflectance of the test subjects. Even with these sources of error, the melanin estimate is reasonable where the $\hat{M}_{\text{sp}}$ produces an average error 2.93% and the $\hat{M}_{\text{md}}$ a slightly smaller

TABLE 4
Estimated melanin content ($\hat{M}$ reported in % Melanin).

Type	I/II	III/IV	V/VI
Subject	{5,7}	{3,6}	{1,2,4}
$\hat{M}_{\text{mdl}}$	{1.60,2.89}	{15.70,13.50}	{21.10,29.60,31.20}
$\hat{M}_{\text{sp}}$	{1.66,4.60}	{17.01,19.49}	{17.89,25.73,37.47}
$\sigma^2 \times 10^{-3}$	{0.02,4.73}	{5.17,2.97}	{2.34,3.81,2.98}
$\hat{M}_{\text{md}}$	{1.60,1.60}	{14.43,18.46}	{17.07,25.52,38.38}
$\sigma^2 \times 10^{-3}$	{0.02,4.73}	{5.17,2.97}	{2.34,3.81,2.98}

average error of 2.90%. Although one may draw the conclusion from the tables that the lightest skin types have the least amount of error in the melanin estimates, it is the darker skin that has a braoder definition and thus truley shows less error in their estimates.

7 Conclusions and Future Work

Algorithms and results for detection and color estimation of human skin in hyperspectral images are presented in this article. The algorithms are based on skin reflectance measurements and results from a diffuse reflectance model developed by the authors in [**?**]. Images used to test the algorithm contain skin with a wide range of pigmentation levels and a variety of skin-color confusers. The detection of skin is conducted with four bands of data in the VIS and NIR. With the proper selection of thresholds and bands, the skin detection algorithm has a probability of detection (P_D) of 95% with a corresponding probability of false alarm (P_{FA}) of 0.7% on modeled data and a $P_D = 95\%$ with a corresponding $P_{FA} = 0.4\%$ on image data. This is a markable improvement over that reported in the literature for RGB-based skin detection with P_D's reported in the low 90% range with large P_{FA}'s around 15%.The impact of this average error is less with darker skin compared to lighter skin because they have a categorization that has a larger melanin span, Once pixels are identified as skin, the skin's melanin level is estimated with an average error of 2.9%.

The likelihood ratio test (LRT) performs marginally better than the rules detector for the experiments accomplished in this article. The distinct advantage of the LRT is the optimality of the detector in terms of minimizing the Bayes Risk. The distinct advantage of the rules detector is its ease of implementation – no training is required for the rules detector where one has to compute the Gaussian Mixture Model parameters for the LRT detector.

There are distinct challenges in melanin estimation. First is the ability to obtain truth to test the accuracy of the estimation methods. Second is the variation in the feature computed from the data, especially for the signal-plus-noise case, where the near-infrared melanin index varies dramatically for a fixed melanin level. Despite these variations, the estimate of the melanin using both the median person and standard person regressions produce reasonable results. If classification is the goal, then one would anticipate a larger classification error for the boundary cases. It is possible that the estimation process can be improved by computing a different set of features, or using a portions of the spectra around 685nm to estimate the melanin level by way of signature matching. This is an area of future research that deserves attention.

Finally, due to the nature of SAR, the ability to do skin detection in real-time is important. The rules detector proposed in this work is computationally efficient. We are able to achieve video rates of 15 frames-per-second for the skin detection, false alarm suppression, and melanine estimation with a software processing solution with our custom monocular system designed after the work presented here.

Acknowledgments

The authors would like to thank Christina Schutte and Dr. Devert Wicker of the Air Force Research Laboratory Sensors Directorate for sponsoring this work. Thanks to Dr. Heidi Bertram, 88th Medical Group, and Dr. Frank Nagy of the Wright State University Anatomical Gift Program for their support and consultation. The authors thank Tracey Hong, Sherry Jaio, Amber Hanson, and Richard Durbin for their assistance with data collection. Finally, we would like to thank Adam Brooks and Andrew Beisley for their assistance with running various simulations.

PLACE PHOTO HERE

Michael J. Mendenhall received the B.S. degree in Computer Engineering from Oregon State University, Corvallis, OR in 1996, the M.S. degree in Computer Engineering from the Air Force Institute of Technology (AFIT), Wright-Patterson AFB, OH, in 2001, and the Ph.D. degree in Electrical Engineering from Rice University, Houston, TX, in 2006. Currently, he is an Assistant Professor of Electrical Engineering at AFIT. He recieved the Dr. Leslie M. Norton teaching award by the AFIT student association and an honorable mention for the John L. McLucas basic research award at the Air Force level, both in 2010. His research interests are in hyperspectral signal/image processing, hyperspectral signature modeling, and computational intelligence.

PLACE PHOTO HERE

Abel S. Nunez received the B.S degree in Engineering from Baylor University, Waco, TX in 1995 and the M.S. degree in Electrical Engineering from the Air Force Institute of Technology (AFIT), Wright-Patterson AFB, OH in 2004, and the PhD degree in Electrical Engineering from the AFIT in 2009. His research interests include hyperspectral signature modeling, automatic target recognition, and waveform diversity for communication systems.

PLACE PHOTO HERE

Richard K. Martin Richard K. Martin received dual B.S. degrees (summa cum laude) in physics and electrical engineering from the University of Maryland, College Park in 1999, and the M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY, in 2001 and 2004, respectively. Since August 2004, he has been an Assistant Professor at the Air Force Institute of Technology (AFIT), Dayton, OH. Dr. Martin has been elected ”ECE Instructor of the Quarter” three times and ”HKN Instructor of the Year” twice, by the AFIT students. His research interests include equalization for multicarrier and single-carrier cyclic-prefixed systems; blind, adaptive filters; sparse adaptive filters; navigation and source localization; and cognitive radio. He has authored eighteen journal papers, thirty-seven conference papers, and four patents.