COLOR-BASED SKIN SEGMENTATION: AN EVALUATION OF THE STATE OF THE ART

Frerk Saxen, Ayoub Al-Hamadi
Otto von Guericke University Magdeburg
Institute for Information Technology and Communications
{Frerk.Saxen, Ayoub.Al-Hamadi}@ovgu.de

ABSTRACT
Skin segmentation is widely used, e.g. in face detection and gesture recognition. In the last years, the number of skin segmentation approaches has grown. However, multiple datasets and varying performance measurements make direct comparison difficult. We address these shortcomings and evaluate 5 threshold-based methods, 5 model-based methods, and 2 region-based state-of-the-art skin segmentation methods. We discuss each algorithm and provide the segmentation performance along with the processing time. All methods are evaluated on the ECU dataset which provides a great amount of training data besides other important attributes.

1. INTRODUCTION
Detecting human skin is rarely the final aim, but it is part of many applications. Among others, in face detection, gesture recognition, finding naked people, and hand sign recognition. The difficulties of skin segmentation have been mentioned early [1] (arbitrary background, unknown lighting conditions and camera calibration, highly variable shape and appearance of skin regions, and different skin tones).

Versatile datasets have been created to cover these attributes and plenty algorithms have been proposed to solve the task. However, basic questions such as “How well do current skin segmentation algorithms work?” , “What is the best approach?”, “Which datasets are available?”, “Which should be preferred?” are not answered easily.

In our study we address these questions. We evaluate 5 threshold-based methods, 5 model-based methods, and 2 region-based methods. Threshold-based methods provide simple decision rules and are easy to implement. This attributes led to a vast use of threshold-based methods in the literature, although most have been evaluated poorly. Model-based methods need training and testing procedures and thus require datasets. Threshold-based methods and model-based methods use single pixel information only. Region-based methods incorporate neighbouring pixels and are often computationally expensive and hence only scarcely used.

Our study extends the work of Vezhnevets et al. [2], Xu and Zhu [3], Schmugge et al. [4], Kelly et al. [5], and Kakumanu et al. [6]. We evaluate the best performing and most often used methods for skin segmentation.

2. SKIN SEGMENTATION METHODS
This section describes skin segmentation methods that are used in the following evaluation. We implemented each algorithm based on the original publication considering the parameter settings provided by the authors. Threshold-based methods provide decision rules for skin segmentation, generally based on thresholds. Model-based methods provide a formula for training and testing the skin segmentation approach. All model-based algorithms provide a threshold \( \tau \), which is determined in the 10-fold cross validation.

2.1. Model-based Methods
The Bayes approach for skin segmentation has been introduced by Jones and Rehg [8]. It assumes that a dataset provides enough samples to model the probability \( P(\text{skin} \mid c) \) that a given color \( c \) belongs to skin by using the relative frequency of skin color observations. The Bayes’ theorem is used to model the probability

\[
P(\text{skin} \mid c) = \frac{P(c \mid \text{skin})P(\text{skin})}{P(c)} = \frac{n(c, X_{\text{skin}})N(X_{\text{skin}})}{n(c)N(X)},
\]

where \( X \) denotes all (skin and non-skin) samples of the training set and \( X_{\text{skin}} \) all skin samples of the training set. \( n(c, X) \) denotes how often the color \( c \) is observed in \( X \) and \( N(X) \) are the total number of samples in \( X \). The color \( c \) is considered as skin pixel if

\[
\frac{n(c, X_{\text{skin}})}{n(c, X)} \geq \tau_{\text{Bayes}}.
\]

The relative frequency can be stored in a lookup table (LUT) for every possible color value \( c \) for fast segmentation. Using the RGB color space, \( 256^3 \) possibilities must be stored in

This work is part of the project done within the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG); Transfer project (T01).
the LUT. Phung et al. [7] have used color quantization to reduce the LUT space. Using only 32^3 bins requires 512 times less memory without noticeable loss in segmentation performance. The effect of color-transformation has been investigated by Shin et al. [9] and Kakumanu et al. [6], who recommended to use the original color space of the dataset because “histogram-based methods are not affected by the color space representation” [6]. Even though Shin et al. have shown that dropping the illuminant component results in a loss in segmentation performance, using 2D color spaces (e.g. the normalized RGB color space) is still prevalent in the literature.

Gaussian models provide the skin distribution using the probability density function (pdf). The mean \( \mu \) and the covariance matrix \( \Sigma \) are computed from all skin samples \( X_{\text{skin}} \) of the training set. A given pixel color \( c \) belongs to skin if the pdf exceeds the threshold \( \tau_{\text{Gauss}} \)

\[
(2\pi)^{-\frac{3}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2} (c - \mu)^T \Sigma^{-1} (c - \mu)} \geq \tau_{\text{Gauss}}.
\]  

Phung et al. [7] have shown that using 3D Gaussian improves the segmentation performance. We evaluate a uni-modal skin pdf because it is dominant in the literature. For an elaborate evaluation of Gaussian models (including mixture of Gaussians) we refer to Phung et al. [7].

DSSD (Dempster-Shafer based Skin Detection) has been proposed and improved by Shoyaib et al. [10, 11]. It handles the uncertainties of threshold-based methods using the Dempster-Shafer (DS) theory of evidence. Six threshold-based methods (which are called source of informations (SoIs)) are weighted by skin and non-skin mass values \( m_{\text{skin}} \) and \( m_{\text{non-skin}} \). Table 1 shows the SoIs with skin and non-skin weights. We use simplifications from Barnett [12] to reduce the computational complexity from \( O(3^n) \) to \( O(n) \) without applying approximations, where \( n = 6 \) denotes the number of SoIs. The uncertainties

\[
\Theta_{\text{skin}} = \prod_{i=1}^{n} \left\{ \begin{array}{ll} 1 - m_{i}^{\text{skin}}, & \text{if } \text{SOI}_i \text{ is true} \\ 1, & \text{otherwise} \end{array} \right. \]

\[
\Theta_{\text{non-skin}} = \prod_{i=1}^{n} \left\{ \begin{array}{ll} 1 - m_{i}^{\text{non-skin}}, & \text{if } \text{SOI}_i \text{ is false} \\ 1, & \text{otherwise} \end{array} \right. \]

could be compared directly (as proposed from Shoyaib et al. [11]). We divide \( \Theta_{\text{non-skin}} \) by \( \Theta_{\text{skin}} \) and map the result to use a threshold \( \tau_{\text{DSSD}} \), which ranges between 0 and 1. Thus, a given pixel belongs to skin if

\[
1 + e^{\frac{\Theta_{\text{non-skin}}}{\Theta_{\text{skin}}}} - 1 \geq \tau_{\text{DSSD}}.
\]

The equation can be rewritten so that only a single comparison \( \tau_{\text{DSSD}} \cdot \Theta_{\text{skin}} \geq \Theta_{\text{non-skin}} \) must be performed. Besides the usage of uncertainties, an advantage of DSSD is that an expensive training procedure is not necessary because the mass values are provided from Shoyaib et al. [11]. Due to its generalization DSSD has shown improved segmentation performance for small training data [11]. Additionally, we trained mass values, which refer to DSSDTrained in our evaluation.

<table>
<thead>
<tr>
<th>Sol</th>
<th>( \mu_{\text{skin}} )</th>
<th>( \mu_{\text{non-skin}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R &gt; 140 )</td>
<td>0.364339536</td>
<td>0.524533342</td>
</tr>
<tr>
<td>( 75 &lt; G &lt; 193 )</td>
<td>0.147315460</td>
<td>0.430099211</td>
</tr>
<tr>
<td>( 43 &lt; B &lt; 161 )</td>
<td>0.222264545</td>
<td>0.44025872</td>
</tr>
<tr>
<td>( 25 &lt;</td>
<td>R - G</td>
<td>&lt; 130 )</td>
</tr>
<tr>
<td>( 45 &lt;</td>
<td>R - B</td>
<td>&lt; 187 )</td>
</tr>
<tr>
<td>( R &gt; G \land R &gt; B ) &amp; ( B &gt; G \land R &gt; B )</td>
<td>0.387888000</td>
<td>0.968090001</td>
</tr>
</tbody>
</table>

Table 1. Source of interests (Sol) with corresponding skin and non-skin mass values from Shoyaib et al. [11]

GDSSD is the generalized DSSD method (provided by Shoyaib et al. [11]). The equations for multiplying uncertainties and segmentation (Eq. 4, 5, 6) remain unchanged, but the threshold \( \tau_{\text{GDSSD}} \) is used. The source of interests (SoIs) are not thresholded anymore. For example, \( \text{SOI}_4 \) becomes \(|R - G|\). Instead, for every particular \(|R - G|\) value a Sol is employed. This \( \text{SOI}_4 \) belongs to skin, i.e. \( \text{SOI}_4 \) is true, if

\[
n(|R - G|, X_{\text{skin}}) > n(|R - G|, X_{\text{non-skin}}),
\]

where \( n(c, X_{\text{skin}}) \) denotes how often \( c \) is observed in the skin training set \( X_{\text{skin}} \). \( X_{\text{non-skin}} \) denotes the training set containing only non-skin samples. Thus, for every \(|R - G|\) value a skin and a non-skin mass value is necessary. The GDSSD method is similar to the Bayes approach but requires less memory. Shoyaib et al. [11] have shown that the GDSSD method can outperform Bayes for a small data set.

2.2. Threshold-based Methods

Threshold-based skin segmentation methods are very popular. About one-third of our reviewed applications use threshold-based methods. The advantages are two-fold. First, it is not necessary to train a threshold-based method. Hence, the efforts in finding datasets and implementing training procedures can be omitted. Second, threshold-based methods are computationally efficient and easy to implement. The lower performance of threshold-based methods and the lack of adjustable thresholds are the disadvantages. We emphasize the first author with the year of publication to provide a unique name in the evaluation.

The Chai1999 method [13] considers only the chrominance components of the \( YC_rC_b \) color space. A given pixel belongs to skin if

\[
133 < C_r < 173 \land 77 < C_b < 127.
\]

Soriano2000 [14] intersects two quadratic functions in the normalized RGB (nRGB) color space. A skin pixel is detected if

\[
g_{\text{down}} < nG < g_{\text{up}} \land W_r > 0.004,
\]
where the lower quadratic equations $g_{\text{down}} = -0.7279 + nR^2 + 0.60667 \cdot nR + 0.1766$ and the upper equation $g_{\text{up}} = -1.8423 + nR^2 + 1.5294 \cdot nR + 0.0422$. The white pixel $(nR = 0.33, nG = 0.33)$ is excluded in Eq. 9 with $W_r = (nR - 0.33)^2 + (nG - 0.33)^2$.

Hsieh2002 [15] provides three empirically obtained zones for human skin segmentation in the HSI color space. The pixel is classified as skin if

$$I > 40 \land [(1 < H < 28 \lor 13 < S < 110) \lor (-28 < H < 0 \lor 13 < S < 110) \lor (-51 < H < -29 \lor 13 < S < 75)].$$

Kovac2003 et al. [16] proposes a method for uniform daylight illumination, which classifies a skin pixel if

$$R > 95 \land G > 40 \land B > 20 \land max(R, G, B) - min(R, G, B) > 15 \land |R - G| > 15 \land R > G \land R > B.$$ (11)

Cheddad2009 [17] calculates a unimodal Gaussian in a one-dimensional error signal $e$ and provides the thresholds $\mu - \sigma$ and $\mu + 3\sigma$ respectively. A skin pixel is detected if

$$6.4031 \leq e \leq 30.0135,$$ (12)

where $e = 0.2989R + 0.5870G + 0.1402B - max(G, B)$.

### 2.3. Region-based Methods

Region-based methods often need seed points, which are commonly taken from pixel-based methods. In our experiments, we use BayesRGB32$^2$ to provide seed points since it provides the best segmentation performance (see Fig. 1).

**GraphCut** has been used for skin segmentation, e.g. by Pierrard and Vetter [19] and Khan et al. [20]. The GraphCut algorithm minimizes the energy function

$$E(L) = \sum_{p \in \mathcal{P}} D_p(L_p) + \sum_{(p, q) \in \mathcal{N}} K(p, q) \cdot T(L_p \neq L_q),$$ (13)

where $L = \{L_p \mid p \in \mathcal{P}\}$ is a labelling of image $\mathcal{P}$, $p$ is a pixel of image $\mathcal{P}$, $L_p \in \{S, T\}$ is the label at pixel $p$, $S$ is the skin label and $T$ the non-skin label. $D_p(\cdot)$ is a data penalty function, $K(p, q)$ is a cost function of neighbourhood pixels $p$ and $q$, $\mathcal{N}$ is the set of neighbour pixels (8-connected neighbourhood) and $T(\cdot)$ is 1 if the statement $\phi$ is true or 0 otherwise. We set $K(p, q) = 1$ because calculating neighbourhood distances did not increment the segmentation performance. The probability $P(\text{skin} \mid c_p)$ that a color $c_p$ of pixel $p$ belongs to skin is provided by the Bayes approach (left-hand side of Eq. 2) and models the data penalty function

$$D_p = \begin{cases} -\lambda \cdot \log(\frac{P(\text{skin} \mid c_p)}{1 - P(\text{skin} \mid c_p)}), & \text{if } L_p = S \\ -\lambda \cdot \log(2\gamma_{\text{Bayes}} - P(\text{skin} \mid c_p)), & \text{if } L_p = T \end{cases},$$ (14)

where $\lambda$, which has been chosen empirically, is set to 1.

**Region Growing** is a common image segmentation method and used for skin segmentation by Al-Wadud et al. [21]. An ambiguous pixel is segmented according to its unambiguous neighbourhood pixels. First, the Bayes approach provides a skin probability map $H(\mathcal{P})$ for image $\mathcal{P}$, by calculating the skin probability of every pixel $p \in \mathcal{P}$ using the left-hand side of Eq. 2. $H(\mathcal{P})$ can be thresholded with $\tau_U = 0.3\gamma_{\text{Bayes}} + 0.7$ to obtain unambiguous skin pixels (seeds) $H(\mathcal{P}) > \tau_U$. Unambiguous non-skin seed pixels are obtained thresholding $H(\mathcal{P}) < \tau_L$, where $\tau_L = 0.15\gamma_{\text{Bayes}}$. Ambiguous pixels satisfy $\tau_L \leq H(\mathcal{P}) \leq \tau_U$ and are determined in the latter steps. Second, the gradient magnitude $M = \sqrt{D_x(\mathcal{P})^2 + D_y(\mathcal{P})^2}$ is calculated using the derivations $D_x(\mathcal{P})$ and $D_y(\mathcal{P})$ that are computed by convolving a $5 \times 5$ Sobel operator with the skin probability map $H(\mathcal{P})$. Third, all ambiguous neighbourhood pixels (8-connected neighbourhood) of the skin and non-skin seeds initialize the queue $Q$. Fourth, the queue $Q$ is sorted according to its containing pixels gradient magnitude value $M$. Fifth, the pixel with the lowest gradient magnitude value $M$ in queue $Q$ belongs to skin if more of its neighbour pixels belong to skin than to non-skin. Otherwise it is classified as non-skin. Sixth, all ambiguous neighbourhood pixels of the classified pixel are added to queue $Q$. Steps four to six are repeated until queue $Q$ is empty and all pixels have been classified. Al-Wadud et al. [21] searched for significant local minima of the histogram of $H(\mathcal{P})$ to determine $\tau_L$ and $\tau_U$. We instead used $\gamma_{\text{Bayes}}$ and varied the weights (i.e. 0.3 for $\tau_U$ and 0.15 for $\tau_L$) to find the optimal thresholds but gained only slight improvements in the segmentation performance. Using the gradient magnitude of the image instead of the gradient magnitude of the probability map decreases the segmentation performance by 10% (using F-Measure and the ECU database [7] with 10-fold cross validation).

### 3. EVALUATION

Section 3.1 provides the chosen dataset along with statistics. The results of the evaluation are given in Section 3.2.

#### 3.1. Datasets

Several datasets for skin segmentation are publicly available. In the following, we discuss the advantages and drawbacks of the most important and frequently used datasets. All datasets provide natural images with arbitrary background in indoor and outdoor scenes. Different lighting conditions are provided and various cameras have been used. Jones and Regh [8] published in 1999 the Compaq dataset containing 13640 images ($80 \times 10^6$ skin, $855 \times 10^6$ non-skin pixels). It is still one of the largest datasets for skin segmentation and widely used in literature. The main drawback is the semi-automatic ground truth generation, which result in some
bad annotations. Another disadvantage is the bad quality of many images.

Sigal et al. [22] proposed in 2004 the PAMI dataset with 21 videos cut from Hollywood films, where every 5th frame is annotated \(8.8 \times 10^6\) skin, \(129 \times 10^6\) non-skin pixels. Ambiguous areas are annotated as "don’t care" areas because a pixel based annotation is often hard to perform on border areas between skin and non-skin. The drawbacks are the small size of the dataset and the rare usage in the literature.

Phung et al. [7] provided in 2005 the ECU database with annotations of 4000 images \((209 \times 10^6\) skin, \(902 \times 10^6\) non-skin pixels\). Different skin types are covered as well as various indoor and outdoor lighting conditions. Accurate pixel annotations are provided. The ECU dataset is the largest database for skin segmentation and well organized but unfortunately not often used in the literature.

For our evaluation we chose the ECU dataset [7] because it is the largest skin segmentation dataset, provides accurate pixel annotations, and covers versatile attributes (e.g. background, lighting conditions, skin tones). The Bayes, Gaussian, and Chai1999 method have already been evaluated by Phung et al. [7] and we obtained the same results.

We use F-Measure and 10-fold cross-validation to evaluate the skin segmentation methods.

3.2. Results

Fig. 1 shows the results. We implemented all methods in C++\(^1\) and evaluated the algorithms on a PC (Intel Core i7 - 4 \(\times\) 3.5 GHz). For DSSDShoyaib we used available parameters from the authors and for DSSDTrained we trained a new model. GraphCut, Region Growing and Bayes provide similar results and outperform other methods. GraphCut performs well for many images compared to Bayes but sometimes fails when big areas of the provided seed points from the Bayes approach hardly exceed the threshold \(\tau_{\text{Bayes}}\). Also Region Growing outperforms Bayes hardly because the skin probability map often provides small edges between skin and non-skin regions, resulting in a large growing procedure.

The Bayes method is one order of magnitude faster than its competitors and is one of the best performing methods. Since every possible color value is stored in a lookup table, it could be expected that Bayes performs best among all pixel-based methods. The number of labeled pixels in the dataset is crucial for the performance of the lookup table approach, since Shoyaib et al. [11] have shown that DSSD and GDSSD can outperform Bayes for smaller datasets.

4. CONCLUSION

Automatically detecting skin in images is used in many applications and often a time critical part of a larger task. This explains the popularity of threshold-based methods and the lack of evaluation studies. We addressed these shortcomings and provided an evaluation of state-of-the-art methods along with detailed descriptions of the algorithms. When possible, we used the parameters provided by the authors. We reviewed publicly available datasets for skin segmentation.

All methods reviewed in this paper are based solely on color information. The methods that are based on texture are applicable for a limited range of scales and will be addressed in future work. Further research might consider superpixels [24] for grouping pixels to reduce the sample size and to allow estimating the scale for texture analysis.

Using a dataset with only a single subject improves the segmentation performance for this individual. Detecting the face to individualize the skin classifier according to each subject can also improve the performance [25]. However, we focus on the basic methods for generic datasets in our evaluation.

---

\(^1\)Please contact the authors to obtain the source code.
5. REFERENCES


