Statistical Color Models with Application to Skin Detection

Michael J. Jones and James M. Rehg
Cambridge Research Laboratory
Compaq Computer Corporation
One Cambridge Center
Cambridge, MA 02142
{michael.jones, jim.rehg}@compaq.com

Abstract

The existence of large image datasets such as the set of photos on the World Wide Web make it possible to build powerful generic models for low-level image attributes like color using simple histogram learning techniques. We describe the construction of color models for skin and non-skin classes from a dataset of nearly 1 billion labelled pixels. These classes exhibit a surprising degree of separability which we exploit by building a skin pixel detector achieving a detection rate of 80% with 8.5% false positives. We compare the performance of histogram and mixture models in skin detection and find histogram models to be superior in accuracy and computational cost. Using aggregate features computed from the skin pixel detector we build a surprisingly effective detector for naked people. Our results suggest that color can be a more powerful cue for detecting people in unconstrained imagery than was previously suspected. We believe this work is the most comprehensive and detailed exploration of skin color models to date.

1 Introduction

A central task in visual learning is the construction of statistical models of image appearance from pixel data. When the amount of available training data is small, sophisticated learning algorithms may be required to interpolate between samples. However, as a result of the World Wide Web and the proliferation of on-line image collections, the vision community today has access to image libraries of unprecendented size. These large data sets can support simple, computationally efficient learning algorithms.

This paper describes the construction of statistical color models from a data set of unprecedented size: Our model includes nearly 1 billion labeled training pixels obtained from random crawls of the World Wide Web. From this data we construct a generic color model as well as separate skin and non-skin color models. We use visualization techniques to examine the shape of these distributions. We show empirically that the preponderance of skin pixels in Web images introduces a systematic bias in the generic distribution of color.

We use skin and non-skin color models to design a skin pixel classifier with an equal error rate of 88%. This is surprisingly good performance given the unconstrained nature of Web images. Our visualization studies demonstrate the separation between skin and non-skin color distributions that make this performance possible. Using our skin classifier, which operates on the color of a single pixel, we construct a system for detecting images containing naked people. This second classifier is based on simple aggregate properties of the skin pixel classifier output. Our naked people detector compares favorably to recent systems by Forsyth *et al.* [4] and Wang *et al.* [17], which are based on complex image features. Because it is based on

pixel-wise classification, our detector is extremely fast. These experiments suggest that skin color can be a more powerful cue for detecting people in unconstrained imagery than was previously suspected.

Given a large amount of training data, even simple learning rules can yield good performance. We explore this point by comparing histogram and Gaussian mixture models learned from our dataset. We show that histogram models slightly outperform mixture densities in this context.

We believe this work is the most comprehensive and detailed exploration of skin color models to date. We are making our labeled dataset of 13,640 photos freely available to the academic community. See the Appendix for details.

Section 2 describes the construction and visualization of histogram color models. These models are applied to skin classification in Section 3, where they are also contrasted to mixture densities. Section 4 explores the application of the skin detector to image classification. We review previous work in Section 5 and discuss our conclusions and future plans in Section 6. The Appendix gives more details about our dataset.

2 Histogram Color Models

We first learn a general histogram density using all the photos in our dataset. The dataset was obtained from a large crawl of the Web, which returned around 3 million images (including icons and graphics).[1] A smaller set of images was randomly sampled from this large set to produce a manageable dataset which is representative of the web as a whole. All icons and graphics were removed by hand (see [1] for an automatic approach), resulting in a final set of 18,696 photographs. This dataset contains nearly 2 billion pixels. In comparison, an RGB histogram model with 256 bins per channel has around 16.7 million degrees of freedom (256^3 bins), which is two orders of magnitude less. Details on how to obtain this dataset for academic research purposes can be found in the Appendix.

We organize this dataset in two different ways. In Section 2.1 we use all 18,696 images to build a general color model. We refer to this set of images as the "generic training set". Then, in Section 2.2 we use a subset containing 13,640 photos to build specialized skin and non-skin color models. We refer to this set as the "classifier training set". The images in the classifier training set have been manually separated into those containing skin and those not containing skin. Skin pixels have been manually labeled in the set of skin images. This labelling process is described in more detail in the Appendix. The number of manually labelled pixels in the classifier training set totals nearly 1 billion.

2.1 General Color Model

We first construct a general color model from the generic training set using a histogram with 256 bins per channel in the RGB color space.[2] The histogram counts are converted into a discrete probability distribution $P(\cdot)$ in the usual manner:

$$P(rgb) = \frac{c[rgb]}{T_c}, \tag{1}$$

[1] The breadth-first crawl was initiated from multiple locations, including portals such as yahoo.com and netscape.com. All images embedded in web pages and contained in on-line directories were returned by the crawl, which terminated when a sufficient number of images had been acquired.

[2] Each of the three histogram dimensions is divided into 256 bins, and each bin stores an integer counting the number of times that color value occurred in the entire database of images.

where $c[rgb]$ gives the count in the histogram bin associated with the RGB color triple rgb and T_c is the total count obtained by summing the counts in all of the bins.

To visualize the probability distribution, we developed a software tool for viewing the histogram as a 3-D model in which each bin is rendered as a cube whose size is proportional to the number of counts it contains. The color of each cube corresponds to the smallest RGB triple which is mapped to that bin in the histogram. Figure 1 (a) shows a sample view of the histogram, produced by our tool. This rendering uses a perspective projection model with a viewing direction along the green-magenta axis which joins corners $(0, 255, 0)$ and $(255, 0, 255)$ in color space. The viewpoint was chosen to orient the gray line horizontally. The gray line is the projection of the gray axis which connects the black $(0, 0, 0)$ and white $(255, 255, 255)$ corners of the cube. The histogram in Figure 1 (a) is of size 8 and only shows bins with counts greater than $336,818$. Down-sampling and thresholding the full size model makes the global structure of the distribution more visible.

By examining the 3-D histogram from several angles its overall shape can be inferred. Another visualization of the model can be obtained by computing its marginal distribution along a viewing direction and plotting the resulting 2-D density function as a surface. Figure 1 (b) shows the marginal distribution that results from integrating the 3-D histogram along the same green-magenta axis used in Figure 1 (a). The positions of the black-red and black-green axes under projection are also shown. The density is concentrated along a ridge which follows the gray line from black to white. White has the highest likelihood, followed closely by black.

Additional information about the shape of the surface in Figure 1 (b) can obtained by plotting its equiprobability contours. These are shown in Figure 1 (c). They were obtained with the `contour` function in Matlab 5.0. It is useful to compare Figure 1 (c) with Figure 1 (a) as they are drawn from the same viewpoint. This plot reinforces the conclusion that the density is concentrated around the gray line and is more sharply peaked at white than black. An intriguing feature of this plot is the bias in the distribution towards red.

This bias is clearly visible in Figure 1 (d), which shows the contours produced by a different marginal density, obtained by integrating along the gray axis. The distribution shows a marked asymmetry with respect to the axis of projection that is oriented at approxmiately 30 degrees to the red line in the figure. In the next section, we will demonstrate empirically that this bias is due largely to the presence of skin in Web images.

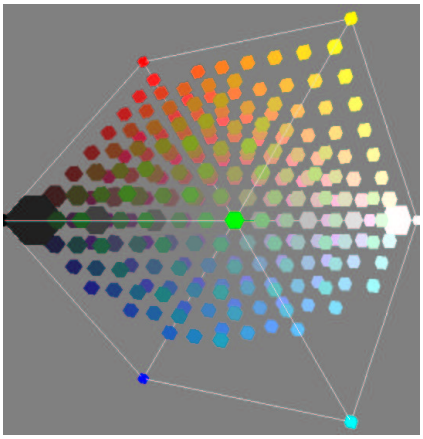
In summary, the generic color model built from Web images has three properties:

1. Most colors fall on or near the gray line.

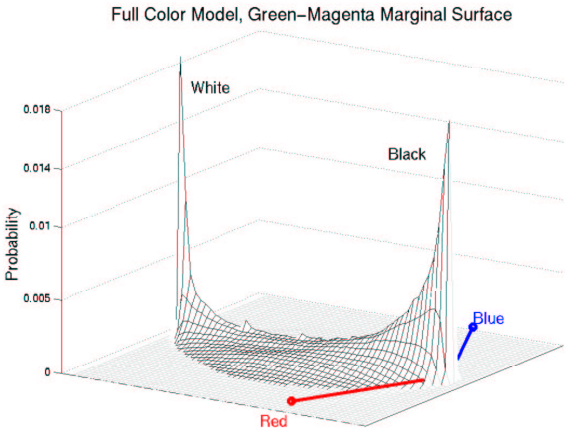
2. Black and white are by far the most frequent colors, with white occuring slightly more frequently.

3. There is a marked skew in the distribution toward the red corner of the color cube.

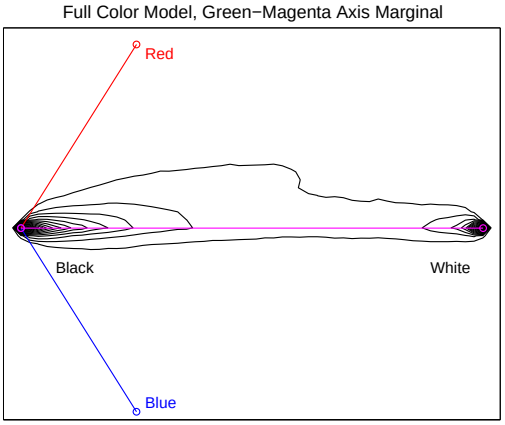
In gathering our dataset we made two additional observations about images on the Web. First, 77% of the possible 24 bit RGB colors are never encountered (i.e. the histogram is mostly empty). Second, about 52% of our Web images have people in them. Table 1 contains a summary of facts about our dataset and color models.



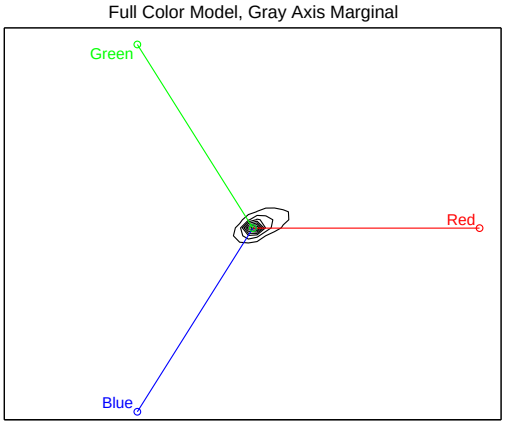
(a) 2-D rendering of 3-D histogram model viewed along the green-magenta axis.



(b) Surface plot of the marginal density formed by integrating along the viewing direction in (a).



(c) Equiprobability contours from the surface plot in (b).



(d) Contour plot for an integration of (a) along the gray axis.

Figure 1: Four visualizations of a full color RGB histogram model constructed from nearly 2 billion Web image pixels.

2.2 Skin and Non-skin Color Models

Next we use the labeled classifier training set to construct skin and nonskin color models for skin detection. The color of skin in the visible spectrum depends primarily on the concentration of melanin and hemoglobin [15]. The distribution of skin color across different ethnic groups under controlled conditions of illumination has been shown to be quite compact, with variations expressable in terms of the concentration of skin pigments (see [3] for a recent study). However, under arbitrary conditions of illumination the variation in skin color will be less constrained. This is particularly true for web images captured under a wide variety of imaging conditions. However, given a sufficiently large collection of labeled training pixels we can still model the distribution of skin and non-skin colors accurately.

We constructed skin and non-skin histogram models using our classifier training set of images. The skin pixels in the 4675 images containing skin were labelled manually and placed into the skin histogram. The 8965 images that did not contain skin were placed into the non-skin histogram. Given skin and non-skin histograms we can compute the probability that a given color value belongs to the skin and non-skin classes:

$$P(rgb|skin) = \frac{s[rgb]}{T_s}, \qquad P(rgb|\neg skin) = \frac{n[rgb]}{T_n} \tag{2}$$

where $s[rgb]$ is the pixel count contained in bin rgb of the skin histogram, $n[rgb]$ is the equivalent count from the non-skin histogram, and T_s and T_n are the total counts contained in the skin and non-skin histograms, respectively.

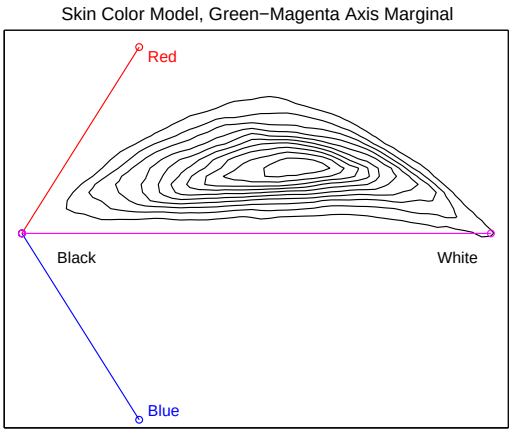
The skin and non-skin color models can be examined using the same techniques we employed with the full color model. Contour plots for marginalizations of the skin and non-skin models are shown in Figure 2. The marginalizations are formed by integrating the distribution along two orthogonal viewing axes. These plots show that a significant degree of separation exists between the skin and non-skin models. The non-skin model, is concentrated along the gray axis, while the majority of the probability mass in the skin model lies off this axis. This separation between the two classes is the basis for the good performance of our skin classifier, which will be described in Section 3.

It is interesting to compare the non-skin color model illustrated in Figure 2 (c) and (d) with the full color model shown in Figure 1 (c) and (d). The only difference in the construction of these two models is the absence of skin pixels in the non-skin case. Note that the result of omitting skin pixels is a marked increase in the symmetry of the distribution around the gray axis. This observation suggests that although skin pixels constitute only about 10% of the total pixels in the dataset, they exhert a disproportionately large effect on the shape of the generic color distribution for Web images, biasing it strongly in the red direction. We suspect that this effect results from the fact that the skin class occurs more frequently than other classes of object colors (52 % of our images contained skin).

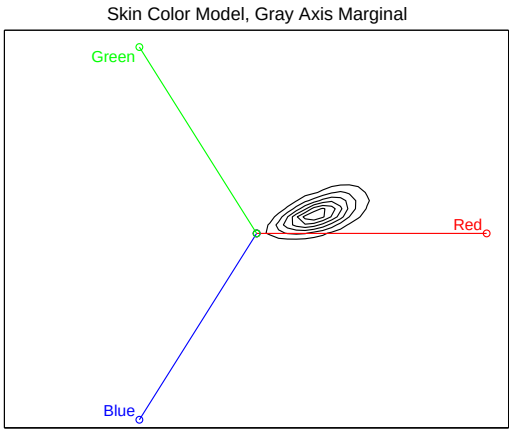
2.3 Discussion

A number of statistics about the general, skin, and non-skin histogram color models are summarized in Table 1. Total counts gives the total number of pixels used to form each of the three models.[3] Note that the skin model was formed from more

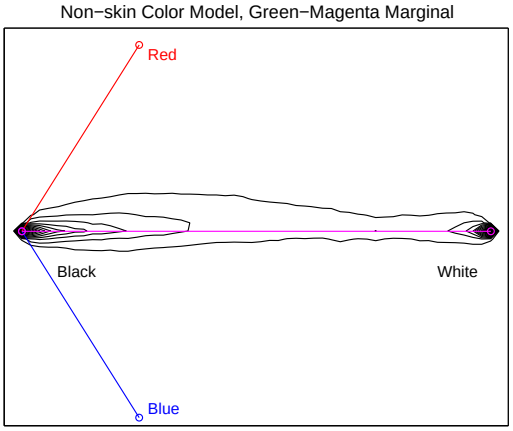
[3]The general model was constructed from 18,696 photos, while the skin and non-skin models were constructed from 13,640 photos. See the Appendix for details.



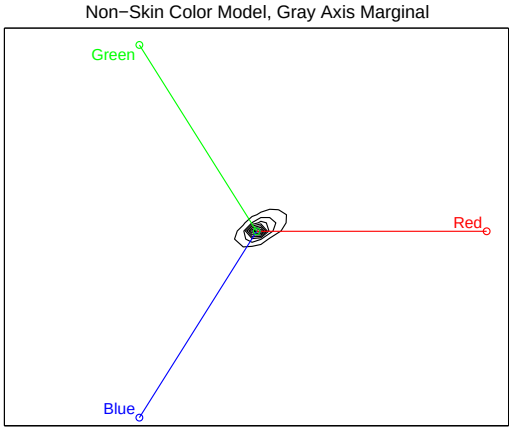
(a) Contour plot for skin model, marginalized along the green-magenta axis.



(b) Contour plot for skin model, marginalized along the gray axis.



(c) Contour plot for non-skin model, marginalized along the green-magenta axis.



(d) Contour plot for non-skin model, marginalized along the gray axis.

Figure 2: Contour plots for marginalizations of the skin and non-skin color models. The top row shows the skin model, the bottom row shows the non-skin model. The left column uses the viewing direction from Figure 1 (c) while the right column uses the view from Figure 1 (d).

	Total Counts	Total Occupied Bins	Percent Unoccupied
General Model	1,949,659,888	3,925,498	76.6
Skin Model	80,377,671	959,955	94.3
Non-skin Model	854,744,181	3,273,160	80.5
Overlapping skin and non-skin bins:			933,275
Skin pixels as a percentage of total pixels:			10 %
Total photos in labeled dataset:			13,640
Percentage of photos containing skin:			52 %

Table 1: Facts about photo image data set and the general, skin, and non-skin color models that were constructed from it.

than 80.3 million hand labelled skin pixels! Total occupied bins refers to the number of bins in each model with nonzero counts. This is also expressed as the percentage of the bins in each model that were unoccupied. Overlapping bins gives the number of bins which are non-empty in both skin and non-skin histogram models.

We make a few observations about these statistics. First, 76.6% of the 16.7 million possible RGB values were not encountered in any of the training images. Second, of the 959,955 colors that occurred as skin, 933,275 (97.2%) also occurred as non-skin. This suggests that the skin detection problem could be difficult since there is significant overlap between the skin and non-skin models. However, overlap is only a significant problem if the counts in the shared bins are comparable in the skin and non-skin cases. The plots in Figure 2 demonstrate that there is in fact reasonable separation between the skin and non-skin classes.

3 Skin Detection Using Color Models

Given skin and non-skin histogram models we can construct a skin pixel classifier. Such a classifier could be extremely useful in two contexts. First, for applications such as the detection and recognition of faces and figures, skin is a useful low-level cue that can be used to focus attention on the most relevant portions of an image. This approach is used in many systems, see [2, 12, 4]. A second role for skin pixel detection is in image indexing and retrieval, where the presence of skin pixels in a photo is an attribute that could support queries or categorization. We give two examples of this application in Section 4.

We derive a skin pixel classifier through the standard likelihood ratio approach [5]. A particular RGB value is labeled skin if

$$\frac{P(rgb|skin)}{P(rgb|\neg skin)} \geq \Theta, \tag{3}$$

where $0 \leq \Theta \leq 1$ is a threshold which can be adjusted to trade-off between correct detections and false positives. We can also write Θ as a function of the priors and the costs of false positives and false negatives: [5]:

$$\Theta = \frac{c_p P(\neg skin)}{c_n P(skin)} \tag{4}$$

where c_p and c_n are the application-dependent costs of false positives and false negatives, respectively. One reasonable choice of priors is $P(skin) = T_s/(T_s + T_n)$. The most important property of equation 3 is the receiver operating char-

acteristic (ROC) curve [16], which shows the relationship between correct detections and false detections as a function of the detection threshold Θ. We make extensive use of ROC curves in this paper to quantify the effect of design choices on classifier performance.

3.1 Histogram-based Skin Classifier

We conducted a series of experiments with histogram color models using the skin classifier defined by equation 3. For these experiments, we divided our classifier training set into separate training and testing sets. Skin and non-skin color models were constructed from a 6822 photo training set using the procedure described in Section 2.2. In this case there were 4483 training photos which formed the non-skin color model and 2339 training photos which formed the skin color model. From our 6818 photo testing set (4482 non-skin and 2336 skin photos) we obtained two populations of labelled skin and non-skin pixels which were used to test the classifier performance.

Figure 3 shows some examples of skin detection in test images for $\Theta = 0.4$. The classifier does a good job of detecting skin in most of these examples. In particular, the skin labels form dense sets whose shape often resembles that of the true skin pixels. The detector tends to fail on highly saturated or shadowed skin. An example of the former type of failure can be seen on the forehead of the woman in the middle of the top row. An example of the latter failure is visible in the neck of the athlete in the middle of the bottom row.

The example photos also show the performance of the detector on non-skin pixels. In photos such as the house (lower right) or flowers (upper right) the false detections are sparse and scattered. More problematic are images with wood or copper-colored metal such as the kitchen scene (upper left) or railroad tracks (lower left). These photos contain colors which often occur in the skin model and are difficult to discriminate reliably. This results in fairly dense sets of false postives.

Classifier performance can be quantified by computing the ROC curve [16] which measures the threshold-dependent trade-off between misses and false detections. In addition to the threshold setting, classifier performance is also a function of the size of the histogram (number of bins) in the color models. Too few bins results in poor accuracy while too many bins lead to over-fitting.

Figure 4 shows the family of ROC curves produced as the size of the histogram varies from 256 bins/channel to 16. The axis labelled “Probability of correct detection” gives the fraction of pixels labelled as skin that were classified correctly, while “Probability of false detection” gives the fraction of non-skin pixels which are mistakenly classified as skin. These curves were computed from the test data. Histogram size 32 gave the best performance, superior to the size 256 model at the larger false detection rates and slightly better than the size 16 model in two places.

The performance of the skin classifier is surprisingly good considering the unconstrained nature of Web images. The best classifier (size 32) can detect roughly 80% of skin pixels with a false positive rate of 8.5%, or 90% correct detections with 14.2% false positives. Its equal error rate is 88%. This corresponds to the point on the ROC curve where the probability of false rejection (which is one minus the probability of correct detection) equals the probability of false detection. Another scalar measure of classifier performance is the area under the ROC curve. Our best skin classifier has an area of 0.942 (it

Figure 3: **Examples of skin detections. For each pair, the original image is shown above and the detected skin pixels are shown below.**

would be 1.0 for a perfect detector).

We tested the performance of the skin classifier as the amount of training data was increased, using the 256^3 histogram model. We divided the skin and non-skin images in the training set into chunks containing approximately 2.5 million skin pixels and 28 million non-skin pixels. On each iteration we added one such chunk of new skin and non-skin pixels to the evolving training set. A ROC curve was computed at each iteration showing the classifier performance on the partial training set as well as on the full test set. The results are shown in Figure 5. As more data is added, performance on the training set decreases because the overlap between skin and non-skin data increases. Performance on the test set improves because the test and training distributions become more similar as the amount of training data increases. Performance on both training and test sets converges relatively quickly. There is little change in either after about 8 iterations.

This ROC curve convergence guided our data collection process. During this research, we added photos selected at

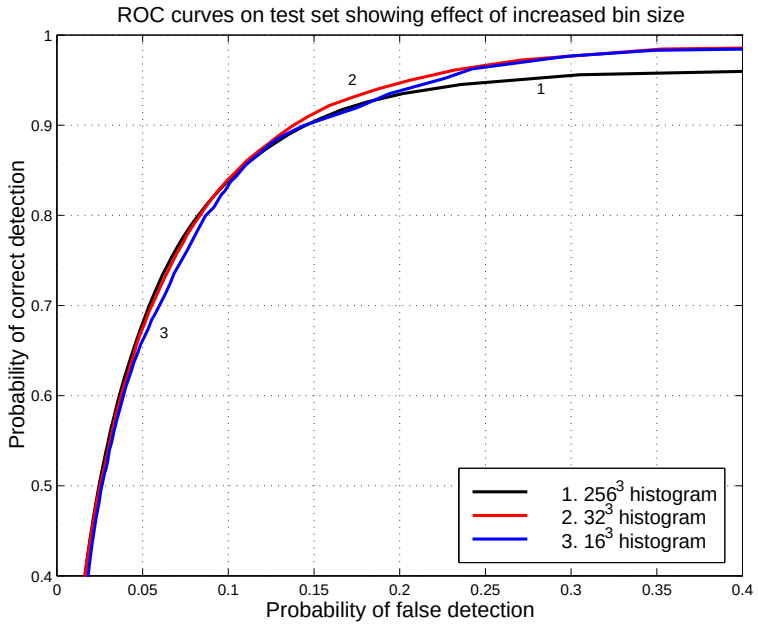


Figure 4: ROC curves for the skin detector as a function of histogram size.

random from a larger set to our model until we judged that the ROC curves had converged. Our final total of 13,640 photos corresponds to this stopping point.

3.2 Comparison to Mixture of Gaussian Classifier

Much of the previous work on skin classification has used a mixture of gaussian model of skin color (some representative examples are [7, 13]). One advantage of mixture models is that they can be made to generalize well on small amounts of training data. One possible benefit of a large dataset is the ability to use density models such as histograms which are computationally simpler to learn and apply. We trained mixture models for our dataset and compared their classification performance to the histogram models of Section 3.1.

A mixture density function is expressed as the sum of gaussian kernels:

$$P(\mathbf{x}) = \sum_{i=1}^{N} w_i \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)}, \tag{5}$$

where $\mathbf{x}$ is an RGB color vector and the contribution of the i^{th} gaussian is determined by a scalar weight w_i, mean vector μ_i, and diagonal covariance matrix $\mathbf{\Sigma}_i$.

We trained two separate mixture models for the skin and non-skin classes. We used 16 gaussians in each model. The models were trained using a parallel implementation of the standard EM algorithm [11]. The non-skin model was trained using the same data as the histogram model in Section 3.1. The skin model was trained using a subset of approximately 74% of the histogram training data. This was simply because that was all the skin training data we had at the time that we performed the mixture experiments. A complete listing of the learned parameters for the skin and nonskin mixture models

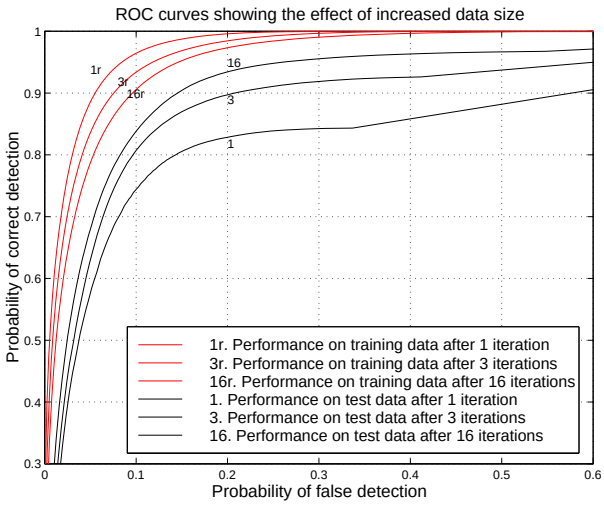


Figure 5: ROC curves for skin classifier training and testing as a function of the amount of training data.

can be found in the Appendix.

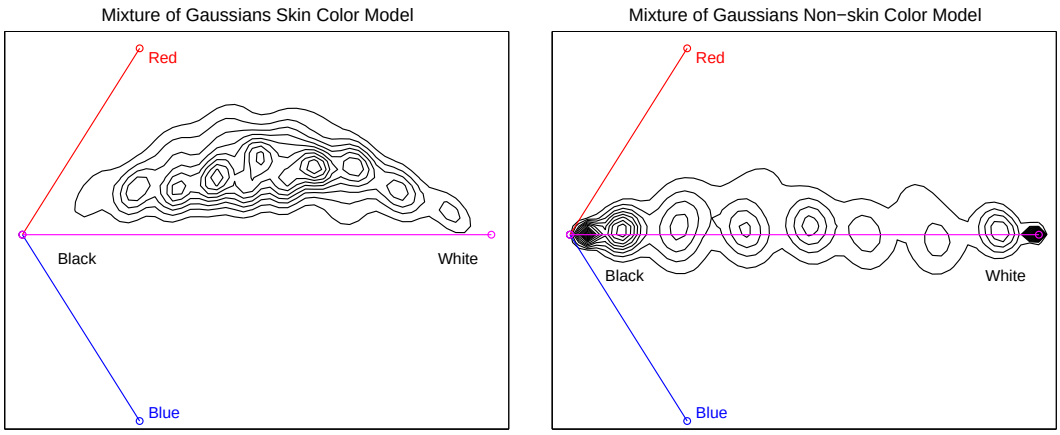
Contour plots for the mixture of gaussian skin and non-skin models are shown in Figure 6. In both plots the 3-D density is integrated along the green-magenta axis. These plots correspond to the marginalizations of the related histogram models shown in Figures 2(a) and (c). The positions of individual gaussian kernels can be observed in the level sets.

Figure 7 (a) shows the ROC curve for the skin pixel classifier based on the mixture of gaussian color models. It is shown in comparison to the *best* histogram model ROC curve, which uses a histogram of size 32. We can see that the histogram model gives slightly better performance in this case. The area under the ROC curve for the mixture model is 0.932 as compared with 0.942 for the histogram model. The fact that the mixture density performance is slightly below the histogram performance may be due in part to the influence of the non-skin model, which has a significant impact on classifier performance and is less likely to form a compact distribution.

It is interesting to compare the mixture and histogram models from the standpoint of computational and storage costs. The mixture of gaussian model is significantly more expensive to train than the histogram models. It took about 24 hours to train both skin and non-skin mixture models using 10 Alpha workstations in parallel. In contrast, the histogram models could be constructed in a matter of minutes on a single workstation. The mixture model is also slower to use during classification since all of the gaussians must be evaluated in computing the probablity of a single color value. In contrast, use of the histogram model results in a fast classifier since only two table lookups are required to compute the probability of skin.

From the standpoint of storage space, however, the mixture model is a much more compact representation of the data. There are a total of 224 floating point parameters (896 bytes assuming 4 byte floats) in the skin and non-skin mixture densities that we used. In contrast, the size 32 histogram model requires 262 Kbytes of storage, assuming one 4 byte integer per bin.

We conducted an additional experiment to verify the importance of having a large data set in obtaining good classifier



(a) Contour plot for skin model.

(b) Contour plot for non-skin model.

Figure 6: Contour plots for marginalizations of the mixture of gaussian skin and non-skin color models.

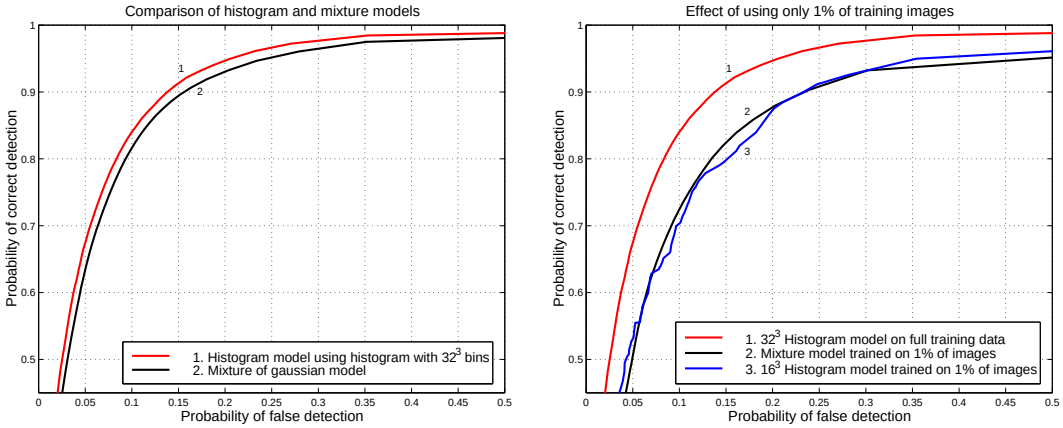
performance. Since the ROC curves in Figure 5 (a) used a histogram of size 256, there remained the possibility that a model with better generalization, such as a mixture density, might require far less data. To test this hypothesis, we built histogram and mixture models from a much smaller set of images. We picked 30 skin images and 58 non-skin images, which make up approximately 1% of the training set. This sample yielded 406,135 skin pixels and 4,017,896 non-skin pixels for training the models. The ROC curves for the best histogram and mixture models are shown in Figure 7 (b). The area under the histogram model's ROC curve is 0.890 and the area under the mixture model's ROC curve is 0.895. They both perform much worse than models using the full training set.

3.3 Discussion

We have demonstrated that a surprisingly effective skin detector for Web images can be constructed from histogram color models. An equal error rate of 88% was obtained from a histogram of size 32, which gave the best generalization. The histogram model compared favorably to mixture models trained on similar data (see Figure 7(a)). This is presumably due to its increased degrees of freedom.

We also explored the sensitivity of our detector to the amount of training data. As demonstrated in Figure 5, the size of our dataset was determined empirically by monitoring the convergence of the skin detector ROC curve as data was added to the model. This graph suggests that the use of additional training data beyond our current dataset is unlikely to improve the skin detector's performance. We demonstrated in Figure 7(b) that using a smaller amount of photos leads to decreased performance even with color models that perform significant generalization.

We conclude that achieving higher detection rates for skin is likely to require analysis at a greater spatial scale than the color values of individual pixels. In many applications, however, purely color-based techniques play an important role because they provide extremely useful information at very low computational cost. For example, our histogram-based



(a) ROC curves comparing histogram and mixture models on full training data.

(b) ROC curves for 1 % of images, histogram and mixture models.

Figure 7: ROC curves comparing mixture model and histogram models under varying training data.

classifier can analyze an image in less time than it takes to read it in from disk storage. Nearly every real-time system for face analysis uses some type of color-based skin detector as a focus-of-attention mechanism.

It is possible that color spaces other than RGB could result in improved detection performance [6, 14]. Different color spaces would result in different decision boundaries, assuming that histogram bin sizes greater than one are used for good generalization. This is an interesting question for future research. The primary advantage of RGB is its simplicity and speed in dealing with web images. In many cases classification can be done directly on pixel values without a color space conversion.

4 Image Classification by Skin Detection

One interesting application of skin detection is as part of a larger system for detecting people in photos. A person detector that worked reliably on Web images could be a valuable tool for image search services on the web and in digital libraries, as well as for image categorization. We examine the problem of person detection in Section 4.1 and the easier problem of naked person detection in Section 4.2. While complete solutions to these problems will undoubtedly require more complex analysis, it is interesting to see what performance is possible based on skin color alone. We find that skin color in the absence of strong shape or texture cues is surprisingly effective, particularly for naked person detection.

4.1 Person Detection

Our goal is to determine whether or not an input image contains one or more people by aggregating the pixel-wise output of the skin detector. The baseline detection rate for this problem is 52%, which is the percentage of images in our dataset containing people. We computed a simple feature vector from the output of the skin detector and then trained a classifier on

	% correctly classified person images	*% correctly classified non-person images*	*Overall % correctly classified images*
Training data	83.0% (2488/2999)	70.6% (1412/2000)	78.0% (3900/4999)
Test data	83.2% (835/1004)	71.3% (645/905)	77.5% (1480/1909)

Table 2: Performance of person detector on training and test data.

these features to determine whether a person is present or not. The features we used are:

- Percentage of pixels detected as skin
- Average probability of the skin pixels
- Size in pixels of the largest connected component of skin
- Number of connected components of skin
- Percent of colors with no entries in the skin and non-skin histograms

These features can all be computed in a single pass over the input image, making the resulting person detector extremely fast. No effort was spent tuning or adjusting the feature set, so it is possible that other choices would yield better performance. We used 4999 images which were manually classified into person and non-person sets to train a decision tree classifier using C4.5 [10]. The resulting classifier was then tested on a set of 1909 test images. Table 2 summarizes the results.

The results show that simply analyzing color values allows reasonably good classification of images into those containing people and those not, but this cue alone is not sufficient to fully solve the problem of person detection. One obvious problem is that people will expose varying amounts of skin in a given image. Using other cues such as texture and shape would probably lead to greater accuracy, see [9] for a recent example.

4.2 Adult Image Detection

By taking advantage of the fact that there is a strong correlation between images with large patches of skin and adult or pornographic images, the skin detector can also be used as the basis for an adult image detector. There is a growing industry aimed at filtering and blocking adult content from Web indexes and browsers. Some representative companies are *www.surfcontrol.com* and *www.netnanny.com*. All of these services currently operate by maintaining lists of objectionable URL's and newsgroups and require constant manual updating. An image-based scheme has the potential advantage of applying equally to all images without the need for updating (see [4] for additional discussion).

To detect adult images, we followed the same approach as with person detection. A feature vector based on the output of the skin detector was computed for each training image. The feature vectors included the same five features used for person detection, plus two additional elements corresponding to the height and width of the image. These two were added based on informal observations that adult images are often sized to frame a standing or reclining figure.

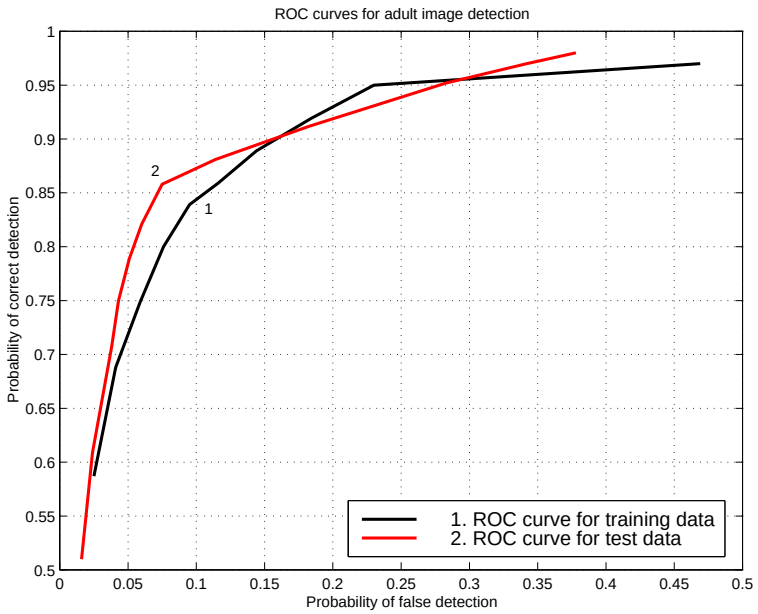


Figure 8: ROC curves for the adult image detector on both training and testing images

We used 10679 images which were manually classified into adult and non-adult sets to train a neural network classifier. There were 5453 adult images and 5226 non-adult images. The neural network outputs a number between 0 and 1, with 1 indicating an adult image. We can threshold this value to make a binary decision. By varying the threshold, we get the ROC curve shown in Figure 8 for the training data.

To test the adult image detector, we gathered images from two new crawls of the Web. Crawl A used adult sites as starting points and gathered many adult images. Crawl B used non-adult sites as starting points and gathered very few adult images. Crawl A consisted of 2365 html pages containing 5241 adult images and 6082 non-adult images (including icons and other graphics). Crawl B consisted of 2692 html pages containing 3 adult images and 13970 non-adult images. We used the adult images from Crawl A and the non-adult images from Crawl B to test the classifier. [4]

The ROC curve for the adult image detector on the test set is shown in Figure 8. The detector achieved, for example, 85.8% correct detections on adult images from Crawl A with 7.5% false positives on the non-adult images from Crawl B. This performance is suprisingly good considering the simplicity of color-based features that were used. In previous systems for adult image detection, skin color is used as a prefilter to guide a detection process which is based wholely on shape, texture, etc. Our results suggest that skin color deserves a more prominent role. A direct comparison between our adult detector and others can be found in Section 5. Some example images for which the adult detector is successful are shown in Figure 9 as well as an example of a typical false positive.

The adult image detector is essentially looking for images with connected regions of skin of the right size. As a consequence, the most common false positive for our system is a close-up image of a face. Use of a face detector in conjunction

[4]Of course the classification of an image in the training set as "adult" is entirely subjective. In our experiment, we labeled any image containing exposed genitals or female breasts as adult, and all others as non-adult.

(a) Examples of images correctly classified by our detector. Both images were classified as adult images.

(b) Example of an image misclassified as adult by our detector.

Figure 9: Examples of correctly and incorrectly classified images.

with the skin detector could alleviate this particular problem. However, major improvements in performance are likely to require the use of other cues such as text, as well as more detailed analysis of image structures.

4.3 Incorporating Text Features into Adult Image Detection

In our final experiment, we explored combining the adult image detector just described with a text-based classifier (obtained from AltaVista) which uses the text occurring on a Web page to determine if it is pornographic. The text-based detector uses the standard approach of matching words on the page against a manually-specified list of objectional words, and classifying based on the match frequencies. To apply the text classfier to individual images occurring on a page, we simply use the label for the page to classify each image that it contains. The text-based classifier on its own achieves 84.9% correct detections of adult images with 1.1% false positives. There is no threshold associated with the text-based classifier we used, so only one point on the ROC curve is realized.

We combined the color-based detector (using a threshold that yielded 85.8% correct detections and 7.5% false positives) with the text-based detector by using an "OR" of the two classifiers, i.e. an image is labelled adult if either classifier labels it adult. The combined detector correctly labels 93.9% of the adult images from crawl A and obtains 8% false positives on the non-adult images from crawl B. Table 3 summarizes these results. One interesting observation is that the color and texture features appear to be complementary, since the combined detector exhibits an increase in both the detection rate and the false alarm rate. This suggests that more sophisticated combination schemes than the "OR" operator could yield even better combined performance.

A significant advantage of our detector is its speed. The skin detector is very fast, as it requires only two table look-ups and an integer division per pixel. The seven element feature vector can be computed in the same pass through the image as the skin detector. Thus the overall computational cost is linear in the number of pixels, with a small constant factor.

	% correctly detected adult images	*% false alarms*
Color-based Detector	85.8%	7.5%
Text-based Detector	84.9%	1.1%
Combined Detector	93.9%	8.0%

Table 3: Comparison of adult image detector using color-based, text-based and combined classifiers on the test data

The average image size in our test set was 269 rows by 301 columns. This resulted in an average classfication time of 43 milliseconds on a 400 MHz Alpha workstation. This is a bit less than the average time it takes to read an image in from disk.

5 Previous work

While there has been much previous work on skin color modeling, we know of no previous effort based on such a large corpus of training and testing data and no comparably detailed study of skin classification in Web images.

Many systems for tracking or detecting people in user-interface or video-conferencing applications have employed skin color models. Histogram models are employed by Schiele and Waibel [13] and Kjeldsen and Kender [8]. Yang *et al.* [18] model skin color as a single gaussian, while Jebara *et al.* [7] employ a mixture density. In all of these systems, the color model is trained on a small number of example images taken under a representative set of illumination conditions. Most, with the exception of [8, 7], do not use non-skin models. These color models are effective in the context of a larger system, but they do not address the question of building a global skin model which can be applied to a large set of images.

The closest works to ours are two systems for detecting images containing naked people developed by Forsyth and Fleck [4] and Wang *et al.* [17]. Both of these systems use a skin color model as a preprocessing step and have been tested on a corpus of Web images. The skin color model used by Forsyth *et al.* consists of a manually specified region in a log-opponent color space. Detected regions of skin pixels form the input to a geometric filter based on body plans. The WIPE system developed by Wang *et al.* uses a manually-specified color histogram model as a prefilter in an analysis pipeline. Input images whose average probability of skin is low are rejected as non-offensive. Images that contain skin pass on to a final stage of analysis where they are classified using wavelet features. Since neither of these works report the performance of their skin detector in isolation, a direct comparison with Figure 4 is not possible.

Forsyth reports two sets of experimental results: the skin filter alone, and used in conjunction with the geometric filter. Their skin filter is not directly comparable to ours, as it uses texture analysis and groups pixels into skin regions. However, they also report strong performance when images that contain one or more detected skin regions are labelled as containing naked people. The detection rate is 79.3 % with a false alarm rate of 11.3 %. When combined with the geometry filter the false positives fall to 4.2 % while the detection rate falls to 42.7 % for the “primary” configuration of the system. Wang *et al.* report the overall the results of the WIPE system on objectionable images: 96% detection rate with 9% false positives. The Forsyth test set contained 4,854 images, the WIPE test set contained 11,885 images, and our test contained 19,211 images.

System	*Detection Rate*	*False Alarm Rate*
Forsyth (Skin Only)	79.3	11.3
Jones-Rehg	88.0	11.3
Forsyth (Skin+Geom)	42.7	4.2
Jones-Rehg	75.0	4.2
WIPE	96	9.0
Jones-Rehg	86.7	9.0

Table 4: Performance comparison for three adult image detection systems.

Table 4 gives a summary of the performance of these two systems in comparison to ours. Since these papers did not report ROC curves, we simply compare our detection rate to theirs under identical false positive rates. In contrast to this previous work, our detector uses very weak global attributes of the detected skin pixels to classify the image. Both body plans and wavelet coefficients have more descriptive power than our seven element feature vector. Perhaps surprisingly, we find that our detection performance is comparable to theirs.

The conclusions that can be drawn from Table 4 are limited by the fact that all three systems use different test sets and exploit different image cues. Of the three, our system provides the strongest test of the value of color alone, since it is the weakest in exploiting shape or geometry cues. Our results suggest that adult detection systems can get more mileage out of skin color than has been previously expected.

6 Conclusions

Color distributions for skin and nonskin pixel classes learned from web images can be used to fashion a surprisingly accurate pixel-wise skin detector with an equal error rate of 88%. The key is the use of a very large labelled dataset to capture the effects of the unconstrained imaging environment represented by web photos. Visualization studies show a surprising degree of separability in the skin and nonskin color distributions. They also reveal that the general distribution of color in web images is strongly biased by the presence of skin pixels. Our dataset of nearly 1 billion labelled pixels is one of the largest ever used in a computer vision task. We are making this dataset freely available to the academic research community. See the Appendix for details on how to obtain it.

One possible advantage of using a large dataset is that simple learning rules, such as histogram density estimators, may give good performance. This can result in computationally simple algorithms for learning and classification. We show that in our context histogram classifiers compare favorably to the more expensive but widely-used Gaussian mixture densities.

A pixel-wise skin detector can be used to detect images containing naked people, which tend to produce large connected regions of skin. We show that a detection rate of 88% can be achieved with a false alarm rate of 11.3%, using a seven element feature vector and a neural network classifier. This performance is comparable to systems which use more elaborate spatial image analysis. In comparison, our classifier is much faster. It operates in less time than it takes to read in the image from disk storage.

Our results suggest that skin color is a more powerful cue for detecting people in unconstrained imagery than was

previously suspected.

Acknowledgments

The authors would like to thank Michael Swain and Henry Schneiderman for some valuable discussions and Pedro Moreno for his help in fitting the mixture models using a parallel implementation of the EM algorithm. We would also like to thank Nick Whyte of AltaVista for providing the image dataset. The reviewers provided numerous helpful comments which improved the presentation of the paper.

Appendix

Our dataset of labelled skin and non-skin pixels is freely available for academic research purposes. Contact the first author (Michael.Jones@compaq.com) for instructions on how to obtain it. Readers who are interested in using our color models should refer to Table 5, which contains all of the parameters for the mixture of gaussian skin and non-skin models described in Section 3.2.

Each photo in our dataset was processed in the following manner: The photo was examined to determine if it contained skin. If no skin was present, it was placed in the non-skin group. If it contained skin, regions of skin pixels were manually labeled using a software tool, whose interface is shown in Figure 10. This tool allows a user to interactively segment regions of skin by controlling a connected-components algorithm. Clicking on a pixel establishs it as a seed for region growing. The threshold slider controls the Euclidean distance in RGB space around the seed that defines the skin region. By clicking on different points in the photo and adjusting the slider, regions of skin with fairly complex shapes can be segmented quickly. In labelling skin we attempted to exclude the eyes, hair, and mouth opening. The result is a binary mask identifying the skin pixels, which is stored along with each photo.

Figure 10: Snap shot of the tool for segmenting the skin region of an image. The left image shows the completed manual segmentation with the skin pixels highlighted in red. The right image shows the original image.

Non-skin pixels that appeared within a photo containing skin were not included in either color model. This was necessary because of the difficulty in getting a perfect segmentation of the skin in any given image. Some photos contained skin patches of such a small size (e.g. crowd scenes) that segmentation was problematic. Even in photos with large regions of skin it was often hard to precisely define their boundaries (e.g. on the forehead where skin is obscured by hair). We chose the conservative strategy of segmenting the easily identifiable skin pixels and discarding the remainder to avoid contaminating the non-skin model.

One of the issues that arises in a dataset taken from the Web is the question of color quantization. Digital images obtained from different sources such as scanners, capture cards, and digital cameras will have different color resolutions.

Mixture of Gaussian Skin Color Model

Kernel	*Mean*	*Covariance*	*Weight*
1	(73.53, 29.94, 17.76)	(765.40, 121.44, 112.80)	0.0294
2	(249.71, 233.94, 217.49)	(39.94, 154.44, 396.05)	0.0331
3	(161.68, 116.25, 96.95)	(291.03, 60.48, 162.85)	0.0654
4	(186.07, 136.62, 114.40)	(274.95, 64.60, 198.27)	0.0756
5	(189.26, 98.37, 51.18)	(633.18, 222.40, 250.69)	0.0554
6	(247.00, 152.20, 90.84)	(65.23, 691.53, 609.92)	0.0314
7	(150.10, 72.66, 37.76)	(408.63, 200.77, 257.57)	0.0454
8	(206.85, 171.09, 156.34)	(530.08, 155.08, 572.79)	0.0469
9	(212.78, 152.82, 120.04)	(160.57, 84.52, 243.90)	0.0956
10	(234.87, 175.43, 138.94)	(163.80, 121.57, 279.22)	0.0763
11	(151.19, 97.74, 74.59)	(425.40, 73.56, 175.11)	0.1100
12	(120.52, 77.55, 59.82)	(330.45, 70.34, 151.82)	0.0676
13	(192.20, 119.62, 82.32)	(152.76, 92.14, 259.15)	0.0755
14	(214.29, 136.08, 87.24)	(204.90, 140.17, 270.19)	0.0500
15	(99.57, 54.33, 38.06)	(448.13, 90.18, 151.29)	0.0667
16	(238.88, 203.08, 176.91)	(178.38, 156.27, 404.99)	0.0749

Mixture of Gaussian Non-skin Color Model

Kernel	*Mean*	*Covariance*	*Weight*
1	(254.37, 254.41, 253.82)	(2.77, 2.81, 5.46)	0.0637
2	(9.39, 8.09, 8.52)	(46.84, 33.59, 32.48)	0.0516
3	(96.57, 96.95, 91.53)	(280.69, 156.79, 436.58)	0.0864
4	(160.44, 162.49, 159.06)	(355.98, 115.89, 591.24)	0.0636
5	(74.98, 63.23, 46.33)	(414.84, 245.95, 361.27)	0.0747
6	(121.83, 60.88, 18.31)	(2502.24, 1383.53, 237.18)	0.0365
7	(202.18, 154.88, 91.04)	(957.42, 1766.94, 1582.52)	0.0349
8	(193.06, 201.93, 206.55)	(562.88, 190.23, 447.28)	0.0649
9	(51.88, 57.14, 61.55)	(344.11, 191.77, 433.40)	0.0656
10	(30.88, 26.84, 25.32)	(222.07, 118.65, 182.41)	0.1189
11	(44.97, 85.96, 131.95)	(651.32, 840.52, 963.67)	0.0362
12	(236.02, 236.27, 230.70)	(225.03, 117.29, 331.95)	0.0849
13	(207.86, 191.20, 164.12)	(494.04, 237.69, 533.52)	0.0368
14	(99.83, 148.11, 188.17)	(955.88, 654.95, 916.70)	0.0389
15	(135.06, 131.92, 123.10)	(350.35, 130.30, 388.43)	0.0943
16	(135.96, 103.89, 66.88)	(806.44, 642.20, 350.36)	0.0477

Table 5: Means, covariances and weights for mixture of gaussian skin and non-skin color models described in Section 3.2.

Unfortunately, most of the information about color resolution is lost once an image has been stored in one of the file formats that are in wide-spread use on the Web.

References

[1] Vassilis Athitsos, Michael J. Swain, and Charles Frankel. Distinguishing photographs and graphics on the world wide web. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 10–17, San Juan, Puerto Rico, June 20 1997.

[2] Qian Chen, Haiyuan Wu, and Masahiko Yachida. Face detection by fuzzy pattern matching. In *Proc. of Fifth Intl. Conf. on Computer Vision*, pages 591–596, Cambridge, MA, June 1995.

[3] Symon D'Oyly Cotton and Ela Claridge. Do all human skin colors lie on a defined surface within LMS space? Technical Report CSR-96-01, School of Computer Science, Univ. of Birmingham, UK, Jan 1996.

[4] David A. Forsyth and Margaret M. Fleck. Automatic detection of human nudes. *International Journal of Computer Vision*, 32(1):63–77, August 1999.

[5] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.

[6] Yihong Gong and Masao Sakauchi. Detection of regions matching specified chromatic features. *Computer Vision and Image Understanding*, 61(2):263–269, March 1995.

[7] Tony S. Jebara and Alex Pentland. Parameterized structure from motion for 3d adaptive feedback tracking of faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–150, San Juan, Puerto Rico, June 17-19 1997.

[8] Rick Kjeldsen and John Kender. Finding skin in color images. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 312–317, Killington, VT, October 14-16 1996.

[9] Michael Oren, Constantine Papageorgiou, Pawan Sinha, Edgar Osuna, and Tomaso Poggio. Pedestrian detection using wavelet templates. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 193–199, San Juan, Puerto Rico, June 17-19 1997.

[10] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kauffman Publishers, Inc., 1993.

[11] R. Redner and H. Walker. Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Revew*, 26:195–239, 1994.

[12] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[13] Bernt Schiele and Alex Waibel. Gaze tracking based on face-color. In *Proceedings of the International Workshop on Automatic Face- and Gesture-Recognition*, pages 344–349, Zurich, Switzerland, June 26-28 1995.

[14] T. Syeda-Mahmood and Y. Q. Cheng. Indexing colored surfaces in images. In *Proc. of Intl. Conf. on Pattern Recognition*, Vienna, Austria, 1996.

[15] M. J. C. Van Gemert, Steven L. Jacques, H. J. C. M. Sterenborg, and W. M. Star. Skin optics. *IEEE Trans. on Biomedical Engineering*, 36(12):1146–1154, December 1989.

[16] Harry L. Van Trees. *Detection, Estimation, and Modulation Theory*, volume I. Wiley, 1968.

[17] James Ze Wang, Jia Li, Gio Wiederhold, and Oscar Firschein. System for screening objectionable images using daubechies' wavelets and color histograms. In *Proc. of the International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services*, pages 20–30, 1997.

[18] Jie Yang, Weier Lu, and Alex Waibel. Skin-color modeling and adaptation. In *Proceedings of the 3rd Asian Conference on Computer Vision*, pages 687–694, Hong Kong, China, January 8-10 1998.