Statistical Color Models with Application to Skin Detection

Michael J. Jones  James M. Rehg

Cambridge Research Laboratory
Compaq Computer Corporation
One Kendall Square, Bldg 700
Cambridge, MA 02138
{rehg,mjones}@crl.dec.com

Abstract

The existence of large image datasets such as 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 labeled pixels. These classes exhibit a surprising degree of separability which we exploit by building a skin pixel detector that achieves an equal error rate of 88%. 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 detector we build a remarkably effective detector for naked people. 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 digital imagery, the vision community today has access to image libraries of unprecedented size and richness. These large data sets can support simple, computationally efficient learning algorithms.

However, a data set such as the web constitutes a biased sample from the space of possible imagery. Thus, the process of building image models from web data must be accompanied by a process of visualizing these models and investigating the statistical characteristics of on-line image data sets. Color is the simplest visual attribute to model, and it is a natural starting point when working with large data sets. Three dimensional color space results in computationally inexpensive algorithms and models that can be visualized easily.

Recently a number of authors have addressed the problem of constructing “generic prior models” [15] of images using multi-scale statistical modeling techniques [4, 15, 1, 11]. In this work, texture models are constructed from the outputs of multi-scale spatial filters, such as wavelets or steerable pyramids. Applications of these models include texture synthesis and classification, as well as noise removal and image coding. In most cases, models are built from a single example image, or a few examples in the case of [1]. A color histogram model can be viewed as the 0th order version of these spatial models in which the neighborhood structure is limited to a single pixel.

We describe 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 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 on the web. We learn both histogram and mixture densities from this data, and show that histogram models slightly outperform mixture models in this domain.

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. Using our skin classifier, we construct a system for detecting images containing naked people, based on simple aggregate properties of the classifier output. This system compares favorably to recent systems by Forsyth et al. [2] and Wang et al. [13]. This suggests that skin 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. We are making our labeled dataset of 13,640 photos available to the academic research community. Contact the first author for instructions.
2 Generic Histogram Color Model

There are two issues that must be addressed in building a color histogram model: the choice of color space and the size of the histogram, which is measured by the number of bins per color channel. The 24-bit RGB color space is a natural representation for color images found on the web. High quality color images require 24 bit resolution and images with coarser color quantizations can be mapped into it. In contrast, the size of the histogram depends upon the task. Our starting point is the construction of a histogram model in 24 bit RGB color space. Such a model has a size of 256 bins per color channel, which corresponds to more than 16.7 million \(256^3\) bins, each mapped to a specific R,G,B color triple. In Section 3.2 we will show that skin classification requires a smaller histogram size for good generalization.

The dataset for the experiments described in this report were obtained by a large crawl of the web which produced about 3 million images (including icons and graphics). A smaller set of images was randomly sampled from this large set and cleared of all icons and graphics by hand. This produced a set of 18,696 photographs. This is a dataset of nearly 2 billion pixels, which is two orders of magnitude more data than the number of degrees of freedom in a histogram model of size 256. We used this data to construct a generic histogram color model.

The counts in the histogram bins are converted to a discrete probability distribution \(P(\cdot)\) in the usual manner:

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

(1)

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 over all of the bins in the histogram.

To visualize this probability distribution, we display 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 enclosed RGB triple. Figure 1 shows a sample view of the histogram which was produced by our visualization 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 histogram in Figure 1 has \(8^3\) bins 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 apparent.

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

Additional information about the shape of the surface in Figure 2 can be obtained by plotting its equiprobability contours. These are shown in Figure 3. 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. In the next section, we will demonstrate empirically that this bias is due largely to the presence of skin in web images.

It is also interesting to note that in gathering our data set we have found that 77% of the possible colors are never encountered (i.e. the histogram is mostly empty), and about 52% of web images have people in them.

3 Skin and Non-skin Color Models

Our next step is to specialize the generic color model into separate skin and non-skin models using labeled training pixels. These models can be used for skin detection in color images. The color of skin in images depends primarily on the amount of hemoglobin and melanin in the dermis and on the conditions of illumination. It is well-known that the hue of skin is roughly invariant across different ethnic

\(^1\)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.
groups after the illuminant has been discounted. This is because differences in skin color result primarily from differences in the concentration of melanin, which affects the intensity of skin color but not its hue.

Unfortunately we do not know the illumination conditions in an arbitrary image\(^2\) and so the variation in skin colors is much less constrained in practice. This is particularly true for web images captured under a wide variety of conditions. However, given a large collection of labeled training pixels we can still model the distribution of skin and non-skin colors in un-normalized color space.

We constructed skin and non-skin histogram models using a subset of 13,640 photos sampled from the total set of 18,696 photographs described in Section 2. In the 4675 photos containing skin, the skin pixels were segmented by hand using a tool which allowed a person to carefully “paint” the skin region. Areas of the face such as eyes, teeth and hair were not labeled as skin. Pixels not labeled as skin in these photos were discarded to reduce the chance that segmentation errors would contaminate the model. These labeled skin pixels were placed into the skin histogram model. The remaining 8965 photos which did not contain any skin were placed into the non-skin model. These two models together contain almost one billion labeled pixels, which includes more than 38.7 million hand labeled skin pixels! In passing we note that skin pixels make up about 10% of the total pixels in our dataset.

Given these two histograms, we can compute the probability that a given pixel belongs to the skin and non-skin

\[^2^\text{The illuminant could be discounted, however, if a solution to the color constancy problem were available.}\]

\[P(\text{rgb}|\text{skin}) = \frac{s[\text{rgb}]}{T_s}, \quad P(\text{rgb}|\sim\text{skin}) = \frac{n[\text{rgb}]}{T_n} \quad (2)\]

where \(s[\text{rgb}]\) is the pixel count contained in bin \(\text{rgb}\) of the skin histogram, \(n[\text{rgb}]\) 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 generic color model. Contour plots for marginalizations of the skin and non-skin models are shown in Figures 4 and 5. They are formed by integrating along the same green-magenta axis used in Figure 3. 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, described in Section 3.1.

It is interesting to compare the non-skin color model illustrated in Figure 5 with the full color model shown in Figure 3. 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 remarkable 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 exhibit 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). See [6] for more details, including additional visualizations.

3.1 Skin Pixel Detection

Given the skin and non-skin distributions, we can obtain a pixel classifier through the standard likelihood ratio approach [3]. A particular RGB value is labeled skin if

\[
\frac{p(\text{rgb}|\text{skin})}{p(\text{rgb}|\neg\text{skin})} \geq \Theta,
\]

where \(0 \leq \Theta \leq 1\) is a threshold. \(\Theta\) depends upon the application-specific costs of classification errors, as well as on the prior probabilities of skin and non-skin, \(P(\text{skin})\) and \(P(\neg\text{skin})\). One reasonable choice of prior is \(P(\text{skin}) = T_s/(T_s + T_n)\).

An important property of equation 3 is the receiver operating characteristic (ROC) curve, which shows the relationship between correct detections and false detections as a function of the detection threshold \(\Theta\) [12]. The ROC curve provides a global measure of classifier performance which can be used to compare classifier designs. It is also a useful tool when setting detection thresholds for a particular application.

In order to test our classifier, we divided our labeled pixel data into separate training and testing sets. The test set consisted of 2336 skin images and 4482 non-skin images taken from the set of 13,640 labeled photos. The ROC curve for the skin classifier on this test data is shown as plot 5 (the topmost curve) in figure 5. The axis "probability of correct detections" gives the fraction of pixels labeled as skin that were classified correctly, while "probability of false detections" gives the fraction of non-skin pixels which are mistakenly classified as skin.

The performance of the classifier, as measured by the ROC curve, is surprisingly good given the unconstrained nature of web imagery. The classifier has an equal error rate of 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. The area under the ROC curve is 0.942 (it would be 1.0 for a perfect detector). Both the equal error rate and the area under the ROC curve provide scalar measures of overall classifier performance.

Figure 7 shows some representative examples of the skin classifier's performance (with \(\Theta = 0.4\)) on images from our test set. Directly below each image is a mask image in which detected skin pixels are drawn in black. The classifier does a good job of detecting skin in most of these examples, but tends to fail on either highly saturated or shadowed skin. In many of the non-skin images the false detections are sparse and scattered (e.g. the flowers image). More problematic are images with wood or copper-colored metal which are hard to discriminate from skin (e.g. the railroad tracks image).

Note that the use of color spaces other than RGB (such as YUV or HSV) will not improve the performance of the skin detector. Detector performance depends entirely on the amount of overlap between the skin and non-skin samples. Colors which occur in both the skin and non-skin classes with comparable frequencies cannot be classified reliably. No fixed global transformation between color spaces can affect the degree of overlap.

3.2 Analysis of Skin Model

It is interesting to examine the effect of modeling decisions on the classifier performance. Here we focus on three factors: The amount of training data, the number of
Figure 6: ROC curves for a family of skin detectors based on different histogram and mixture modeling choices. The best ROC curve (number 5) is the result of using a $32^3$ bin histogram model.

histogram bins, and a comparison to mixture of Gaussian models. We can summarize our findings as follows: 1) The large size of our data set is crucial, 2) Histograms with 32 bins/channel give the best performance, suggesting that significant generalization is required, 3) Histogram models are superior to mixture of Gaussian models, 4) Using a small amount of data sampled from a large dataset also produces good results. Each of these points is discussed below, see [6] for the complete details.

We constructed a series of histogram color models with varying amounts of training data. This resulted in a family of ROC curves indexed by the number of training pixels. Our dataset of 13,640 labeled photos represents the empirical limiting point of this progression, at which adding additional pixels did not significantly improve performance. The importance of a large dataset is underscored by ROC curve no. 1 in Figure 6. This classifier was constructed from 1% of the available training images and exhibits relatively poor performance (the ROC curve area is 0.890, compared to 0.942 for the full data model).

Generalization in the histogram model is controlled by the number of bins. We found that a histogram with $32^3$ bins (for both skin and non-skin) performed the best when using our full dataset. The ROC curve for this classifier is shown as plot 5 (the topmost curve) in Figure 6. Increasing or decreasing the number of bins reduced the performance (see [6] for details.)

An alternative form of generalization is provided by mixture of Gaussian models. Mixture models have been popular in earlier skin color modeling work [5, 14] and we examined their performance on our dataset. Using the EM algorithm [8], we fit separate mixture models with 16 Gaussian each to the full set of skin and non-skin pixel data. The performance of the resulting classifier was slightly worse than the 32 bin histogram model. Its ROC curve area was 0.932 in comparison to 0.942.

Mixture models might be expected to do better as the size of the state space increases or as the amount of training data decreases. We tested this hypothesis by fitting mixtures to the reduced dataset containing 1% of the total images. The ROC curve for the resulting mixture model is number 2 in Figure 6. Its area is 0.895, compared to 0.890 for the histogram model on this dataset, a slight improvement. However, since mixture models are computationally more expensive than histograms during both learning and evaluation, these results suggest that histograms are the
best choice for skin color modeling.

Finally, we tested the performance of the models trained on a small set of data sampled uniformly from the large training set. We sampled 387,172 skin pixels and 4,261,703 non-skin pixels (1% of the training data) and built both histogram and mixture models from this data. The performance of these models is also shown in figure 6. The area under the ROC curve for the histogram model trained on 1% of the data is 0.9405 and for the mixture model it is 0.9378. They are almost as good as the histogram model using the full training set. This demonstrates that while a large data set is necessary to capture the underlying distribution of skin and non-skin colors, it is sufficient to train models on a smaller set of samples.

4 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 be used as the basis for an adult image detector. The ability to filter out adult images is important for image search engines on the web that wish to avoid offensive content.

To detect adult images, a feature vector is formed based on the output of the skin detector and then a neural network classifier is trained on a set of labeled feature vectors. The features we used are:

- Percentage of pixels detected as skin
- Average probability of the skin pixels
- Size of the largest connected component of skin
- Number of connected components of skin
- Percent of novel pixels (those with zero counts in both the skin and non-skin histograms)
- height of the image
- width of the image

We used 10681 images which were manually classified into adult and non-adult sets to train a neural network classifier. The neural network outputs a number between 0 and 1 with 1 signifying 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 for the crawl and so 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. The classification for each image was again determined manually.

![ROC curves for adult image detection](image)

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

The important statistics for the popular detector on these test sets is the percentage of correct detections for the set of adult images from crawl A and the percentage of false positives for the set of non-adult images from crawl B. The ROC curve for the adult image detector for this training data is also shown in figure 8. The performance is very good considering color is the only feature used. For example, the classifier attains about 85.8% correct detections with about 7.4% false positives.

We have also explored combining the adult image detector just described with a text-based classifier which uses the text around an image on an HTML page to determine if an image is pornographic. 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. The text-based detector by itself correctly labels 84.9% of the adult images with 1.1% false positives.

The results show that simply analyzing color values allows very good detection of adult images. Not surprisingly, adding information from the surrounding text can boost performance significantly.

5 Previous Work

There have been a number of researchers who have looked at using color information to detect skin. We believe we are the first to build a general statistical model of skin color from a large data set. Forsyth et al. [2] and Rowley et. al. [9] have employed ad hoc skin color models as a preprocessor in analyzing large image databases. Other researchers have built small scale skin models using single Gaussians [14], mixtures of Gaussians [5] or histograms [10, 7]. Most of these models are based on skin data ac-
quired from a limited number of people under a limited range of lighting conditions. Our work demonstrates the superiority of histogram models: They are equivalent to mixture models in accuracy and are more efficient computationally.

Forsyth et. al. [2] and Wang et. al. [13] have also looked at the problem of detecting adult images. Both used a simple color model and emphasized shape and texture cues. In contrast we have used a more accurate color model to construct simple spatial features. It is interesting that our detection results are comparable to theirs. This suggests that color is a more powerful cue than might have been expected.

6 Conclusions

The existence of large image datasets such as the 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 have demonstrated this point empirically by constructing generic color models, as well as specialized skin and non-skin color models, from nearly 1 billion labeled pixels. We are making our labeled dataset of 13,640 photos available to the academic research community. Contact the first author for instructions.

We demonstrate that a significant degree of separability exists between the skin and non-skin distributions, which we exploit in building a skin pixel detector with an equal error rate of 88%. Furthermore, we show empirically that the preponderance of skin pixels in Web images leads to a systematic bias in the generic distribution of color on the Web. We explore the performance of both histogram and mixture of Gaussian models in classification, and find histogram models to be superior in accuracy and speed. We believe this work to be the most comprehensive and detailed exploration of skin color models to date.

We demonstrate a surprisingly effective detector for images containing naked people which is based on the output of our skin pixel classifier. This suggests that skin color can be 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. We would also like to thank Pedro Moreno for his help in fitting the mixture models using a parallel implementation of the EM algorithm. Thanks to Nick Whyte of AltaVista for providing the image dataset.

References