Skin Color Detection in Images and Video

Color segmentation is an important aspect of digital image analysis. Skin tone classification is a critical subset of this field, useful when it is necessary to extract parts of an image that could possibly contain parts of a person. Upon recognition of a skin region, one could perform simple detection for a security application, or apply gesture analysis for the sake of human-computer interaction (such as for control or gaming). Machine learning can be applied to both fields, but only skin detection is covered here.

Electronic skin tone detection is made possible by analyzing the different properties of pixels, which will be the attributes learned by the system. These attributes simply refer to the location of a pixel in a color space that defines that pixel based on its Red, Green, and Blue contents, or its Hue, Saturation, and Value, or any parameter from any imaginable color space. This project aims to compare the efficacy of different combinations of attributes from color spaces with different neural networks, and compare them with a baseline to see how well these sophisticated machine learning tools perform against simpler approaches.

Neural networks are useful in this scenario because they can be used to classify nonlinear regions of a space. When more perceptrons are added, the function represented can become increasingly specific (and susceptible to overfitting). However, a function of certain complexity is required to represent the same information encoded in the baseline case, which is simply a histogram tracking the presence of any desired pair of attributes. For this study, the RGB and HSV color spaces were used, and pairs of attributes from within each were applied to neural networks of varying sizes. Training data came from over 200,000 pixels of positively-classified skin tones covering all ethnicities. (Red, green), (red, blue), (green, blue), (hue, saturation), (hue, value), and (saturation, value) data pairs from each pixel were stored into two-dimensional histograms. Neural networks also require negatively classified training data. Given 8 bits per attribute, there can be 65,536 possible attribute pairs per histogram (and per neural network), so it still remained practical to feed the neural network training algorithm every possible example, whether positive or negative.

All six collections of data were fed into four classifiers: one histogram, and three neural networks with either two, three, or five perceptrons in the hidden layer. The neural networks were constructed using the "Multi-Layer Perceptron" model in Weka, which allows for automatic generation of an appropriate neural network, as well as graphical manipulation of the network before learning commences. As a possible trigger to stop overfitting, the error per epoch was monitored, and the training was stopped once this rate of error began to increase. It was observed that once it began to increase, it would continue to increase. As the neural networks featured more perceptrons, it became less likely that the error per epoch would ever increase before the hard limit of 500 iterations; also more perceptrons meant more time was spent finding good weights for the network. For each image, there were 24 outputs covering all six color spaces and four classification methods. Image analysis was performed using the OpenCV C++ image processing library. In order to measure the performance of each algorithm and color space, the precision (true positives / (true positives + false positives)) and recall (true positives / (true positives + false negatives)) was calculated and added together, and ranked according to the sum. To help calculate these figures and analyze the validity of the prediction, an image mask was made by hand such that regions believed to be skin by a human were white, and non-skin regions would be set to black. The classification program would simply leave skin pixels untouched. The edges of the image mask often contain colors that are somewhat difficult to assign to a classification by hand, but compared to the area that is unquestionably skin, this is typically small.

Skin Color Detection in Images and Video

Given the training data and almost half a million pixels of testing data, the task of skin detection seems quite feasible. Over all implementations tested, the average precision was 52.2% and the average recall was 90.7%. The results look extremely promising for neural networks; at least one neural network outperformed the baseline case for each color space, except for RB (red, blue). Many of the neural nets were able to attain recall of 98% or greater, while keeping a precision of at least 50%, sometimes greater than 60%. The top-performing classifiers were neural networks trained on the HV and HS color space attribute pairs, whereas the four classifiers trained on RB took the four lowest slots in performance rankings. The precision could perhaps be improved by tweaking the training data to eliminate information that results in a significant number of false positives.

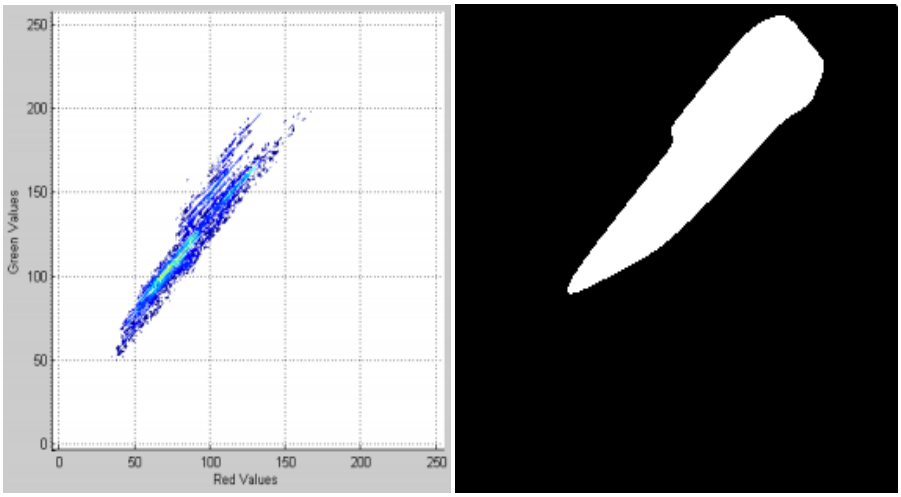
The numbers are nice, but what does this actually look like? Most skin tones across all ethnicities are detected, given they are in relatively even, standard lighting conditions, and photographed with a decent camera. Spots where light reflects directly back into the camera are often missed because they are too white, and extreme shadows are usually missed too. Typically, eyes, teeth, and lips do not get classified with skin, but it is possible that hair and other background elements of the picture will get classified as skin. This is an especially common phenomenon when using a webcam that tends to reduce saturation of the entire picture in general. Because analysis in the HSV space discovers that skin tones have a low saturation level across many different hues, many more things will be classified as skin on a webcam than what would typically be classified when seen through a nicer camera. This is problematic for the HS classifier, which is one of the best-performing spaces otherwise.

Based on correlation data, the choice of color space is far more important than the number of perceptrons present in the hidden layer of the neural network. Given a vector containing performance for all algorithms in sorted order by performance, the correlation between that and another vector containing the number of perceptrons (0, 2, 3, or 5, and also in the same order) came out to only 5.97%. However, correlation between attributes chosen and performance was much more intense. Given the same performance vector, its correlation with a vector representing the attribute pair by the average performance of that attribute pair was 89.2%. It is fairly evident upon coarse examination that certain attribute pairs appear toward the top, and others appear toward the bottom of the performance list.

Specifically, the HV neural network classifier with 3 hidden perceptrons worked the best over the validation data, featuring a precision of 62.0% and a recall of 99.97%. The best baseline case came from HV as well, sporting a precision of 54.2% and a recall of 100%. The worst performer was RB with 3 hidden perceptrons; even though precision came in at 74.0%, this mostly had to do with the fact that it classified almost everything as "not skin" and had very little false positives to speak of to increase the denominator. The recall, on the other hand, was a dismal 5.6%, and these scores were low enough to put this classifier in last place overall.

Neural networks would be useful to apply in systems with limited memory, such as a microcontroller deployed to perform skin detection. Instead of having to remember thousands of values pertaining to whether skin tone data was seen at a particular pair of values, a relatively short mathematical function involving sums and products could be implemented in order to perform a quicker (and more accurate) classification. Some future work on this classifier might involve incorporating all of the attributes of a given color space into the neural network, and comparing this with a suitable baseline. Performance can be further improved by adding morphological operations to the output, which can cleverly enhance salient regions and erode irrelevant ones based on the presence of a subset of the region.

Skin Color Detection in Images and Video



Left: The baseline histogram for the RG attribute pair.
Right: The 5-hidden-perceptron neural network output classifications for the RG attribute pair.

Left: A female's face.
Middle: Skin detection on the face with baseline RG histogram.
Right: Skin detection with 5-hidden-perceptron RG neural network. Hair is detected as skin in both cases.

Left: A hand. Then, the skin detected by the best three classifiers:
(Left Center) HV, 3 hidden perceptrons; (Right Center) HV, 2 hidden perceptrons;
(Right) HS, 5 hidden perceptrons

In the table below, the attribute pair is listed, along with either "Baseline" to indicate the histogram method or a number describing how many perceptrons existed in the hidden layer.

Specific performance results, sorted from best to worst:

	Precision	Recall	Average
HV3	0.620313	0.999668	0.809991
HV2	0.627653	0.9917	0.809677
HS5	0.619597	0.999336	0.809467
HV5	0.619391	0.999004	0.809198
HS3	0.617689	0.999336	0.808513
HS2	0.621033	0.994024	0.807529
HV Baseline	0.542703	1	0.771352
GB5	0.553081	0.989376	0.771229
HS Baseline	0.525297	0.999668	0.762483
SV2	0.567311	0.954183	0.760747
GB3	0.534762	0.985724	0.760243
GB2	0.471942	0.994024	0.732983
GB Baseline	0.451025	0.986056	0.718541
SV3	0.487402	0.931275	0.709339
RG3	0.405002	1	0.702501
SV Baseline	0.397587	0.995684	0.696636
SV5	0.541862	0.842297	0.69208
RG5	0.370006	0.999336	0.684671
RG2	0.343561	1	0.671781
RG Baseline	0.302522	0.999668	0.651095
RB Baseline	0.487868	0.807769	0.647819
RB5	0.592101	0.607238	0.59967
RB2	0.493086	0.62749	0.560288
RB3	0.74026	0.056773	0.398517
Average	0.522211	0.906651	