CONSISTENT IMAGE-BASED MEASUREMENT AND CLASSIFICATION OF SKIN COLOR

Michael Harville, Harlyn Baker, Nina Bhatti
Hewlett-Packard Laboratories
Palo Alto, CA 94304 United States

Sabine Süsstrunk
Ecole Polytechnique Fédérale de Lausanne
CH-1015 Switzerland

ABSTRACT

Little prior image processing work has addressed estimation and classification of skin color in a manner that is independent of camera and illuminant. To this end, we first present new methods for 1) fast, easy-to-use image color correction, with specialization toward skin tones, and 2) fully automated estimation of facial skin color, with robustness to shadows, specularities, and blemishes. Each of these is validated independently against ground truth, and then combined with a classification method that successfully discriminates skin color across a population of people imaged with several different cameras. We also evaluate the effects of image quality and various algorithmic choices on our classification performance. We believe our methods are practical for relatively untrained operators, using inexpensive consumer equipment.

1. INTRODUCTION

Prior image processing work has well addressed skin color *detection*, in which image pixels are classified into skin and non-skin categories. Recent efforts show good results across imagery of people of all races, for a wide variety of cameras and illuminants [6, 7]. Much less work, however, has investigated objective measurement of human skin tone, to enable its color *classification*. An understanding of a person's skin color that is independent of the illuminant and the imaging system would have use in many domains, including:

- Medicine: for quantification of skin erythema, lesions, ultra-violet radiation effects, and other phenomena
- Computer graphics: for more accurate rendering of people in video-conferencing, or for improving or altering their appearance
- Fashion: for automated suggestion of personal appearance products, such as clothing and eyeglasses, that complement skin tone
- Biometrics: to aid in person recognition within small groups, or for systems in which determination of skin coloring is useful

Ideally, this analysis would be produced with easily obtained equipment and minimum effort. However, prior work in the medical domain requires sophisticated instrumentation and controlled lighting, or is not designed to discriminate skin color across people [8, 10, 12]. In computer graphics and interfaces, Tsumura et.al.[5, 11] and Angelopoulou et.al.[1] emphasize representation and synthesis rather than classification, and require multispectral data beyond what cameras normally provide. Methods for dynamic construction of personalized skin models to aid in face and hand tracking in video have been proposed (e.g.[3, 9]), but the models are dependent on the illuminant and imager, and are therefore unsuited for application to skin colors obtained under other conditions.

In this paper, we present and validate consistent, fast techniques for measuring and classifying facial skin color from a single, casually posed digital camera image. The techniques are designed to allow comparison of skin tones across a wide variety of cameras and environments, for all races of people. We also investigate the effects of varying image quality on our methods. Our aim is to develop methods that are practical for application in many computer vision and image processing tasks, beyond what prior art has enabled.

An overview of the image processing pipeline for our methods is shown in Figure 1. The pipeline may be divided into three primary segments: 1) image color calibration and correction, 2) automated sampling of facial skin pixels to produce a skin tone estimate, and 3) classification of skin tone. We present algorithms and results for each segment in turn in Sections 2, 3, and 4.

2. IMAGE COLOR CORRECTION

To classify skin color accurately across a wide variety of imaging devices and environments, we must account for the effects of the camera system and scene illuminants in each image. This could be done by pre-calibrating the camera, and controlling or measuring the illuminant at capture time. However, it is unreasonably difficult, expensive, or time-consuming for end users to do this in many application contexts, particularly in the home, outdoor, and mobile domains. We believe that, for many applications, a better solution is to require the presence in the image of a detectable pattern containing a known reference color set. This pattern might take the form of a paper color chart, available at a doctor's office or store counter, that would be held by the user when the picture is taken. Alternatively, for less conspicuous human-computer interface and monitoring applications, known colors may be embedded naturally in the furniture, walls, or other parts of the background.

Hence, for this paper, we assume the presence of a "color calibration target" in each image, containing colors of known spectral reflectance arranged in a distinctive pattern. We use the color target to transform raw image pixel colors, for all devices and illuminants, into a single "corrected" space in which we analyze skin color.

2.1. Color correction method

Much prior work exists on color calibration of imaging systems; for a summary, see [2]. We present a technique that is designed for accurate correction of skin colors, and that is robust enough for use with a single, casually-posed image from most any consumer camera.

Our color calibration target, visible in the images of Figure 2, was designed for robust automatic detection, as well as for exploration of the space of reference colors best suited for our task. It contains 3 rows of 8 color patches set against a black background, wrapped by a white then a black frame. The top row contains primary and secondary colors for general scene tone balancing, and two shades of gray for white balance. The other two rows contain 16 patches representative of the range of human skin color. We printed the target on photopaper medium, measured the spectral reflectances of each patch, and use a simple image formation model to approximate them as 3-component, sRGB encoded digital values.

Automatic detection of the target is based on segmentation of the image into regions whose contours are located at zero crossings of the Laplacian of a smoothed luminance version of the image. A valid detected target has a black rectangle boundary contained within a

0-7803-9134-9/05/$20.00 ©2005 IEEE

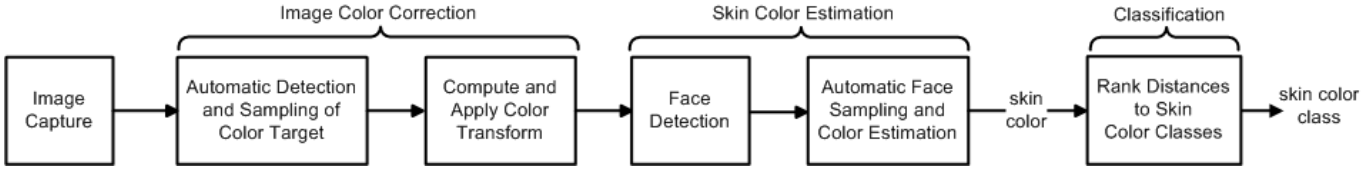


Fig. 1. Overview of image processing pipeline for skin color measurement and classification.

white rectangle boundary. Homographies relate the flat physical surface within each rectangle to the camera image coordinate system, and can be computed from the corner locations of each rectangle. We require the homographies to be in sufficient agreement for valid target detection. If valid, they are used to predict patch locations and shapes within the target, which are verified by search for luminance edges. If a sufficient number (currently 75%) of the patches are found, multiple samples from each patch's interior are extracted and averaged to produce patch mean colors. The configuration of these colors are used to determine the target's orientation.

An optimal color transform is determined by least-squares estimation of the 3x4 transform $\mathcal{A}$ that best maps measured patch means $\vec{M}$ to the reference sRGB patch colors $\vec{R}$:

$$\begin{bmatrix} R_{red} \\ R_{green} \\ R_{blue} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \end{bmatrix} \begin{bmatrix} M_{red} \\ M_{green} \\ M_{blue} \\ 1 \end{bmatrix} \quad (1)$$

This provides a 3x3 color transformation plus a per-component offset. Prior to least-squares, patch means with at least one saturated component are excluded, and the sRGB component function is inverted for both $\vec{M}$ and $\vec{R}$. For each image, we computed four "types" of color transforms, using different subsets of color target patches:

- Base: black, white, gray, blue, green, red, cyan, magenta, yellow
- Skin: the 16 patches in the lower two rows of the target
- Skin+gray: skin patches plus black, white, and gray
- All: all patches on the chart

While transforms based only on skin-colored patches may produce poor correction for the overall image, they may perform best for the skin pixels that are our focus. Alternatively, failure to sample the raw image color space more broadly, using the "base" patches, may destabilize and skew the color transforms even with the skin region.

Fig. 2. Images before (top row) and after (bottom row) color correction, for all four cameras. Left to right: HP850, Nikon D1, iPAQ, web-cam. Transforms based on skin patches only.

2.2. Validation of color correction

Ninety people were each captured with four different cameras while holding a copy of our color calibration target. The cameras were an HP 850 (3.9 MPixel), a Nikon D1 (2.7 MPixel), an HP iPAQ camera (1.3 MPixel), and a commodity web-cam (0.1 MPixel). Illumination also varied across a subject's images. Our color target detection scheme worked well across all cameras, although the success rate declined from 100% for the Nikon and HP850 to 78% and 93% for the iPAQ and web-cam, with failures due largely to motion blur.

A comparison of images before and after color correction, for the same person captured by different cameras, is shown in Figure 2. In this case, color transforms were based only on skin-colored target patches. This provided good agreement among facial color across images, at the expense of less consistent backgrounds. Transforms based on all target patches produced more visually pleasing results.

To quantitatively evaluate the effects of the color correction on skin pixels of subject faces, we compared statistics of manually selected skin pixels for pairs of images of each of the 90 subjects captured with different cameras. Two different people (hereafter designated "expert samplers") each collected at least 20 color samples in each image, avoiding selection of pixels in shadows, specularities, and blemishes. They also avoided nose, upper cheeks, and other areas known to often be sun-affected, and thus less indicative of true skin tone. For two images of the same subject captured by different cameras, manually collected sample sets can be compared via multivariate analysis of variance (MANOVA) to test the hypothesis that both sets were drawn from the same normal distribution. For a given pair of cameras, this MANOVA score (in $[0..1]$ range, 0 being best) was averaged across all subjects. If color correction is working well, the average cross-device MANOVA score should be similar to that obtained when comparing color sample distributions obtained by different expert samplers on the same image. We therefore normalize each average cross-device MANOVA score by the average cross-expert MANOVA score, to obtain our cross-device color correction error measure. This measure should, ideally, be near 1.

Table 1 shows the cross-device color correction error measure obtained for each pairing of cameras we used. We computed these measures for each of the four types of color transforms (e.g. "base", "skin", etc.) listed in Section 2.1. In general, color transforms based only on skin patches performed best, with post-correction, cross-device color differences that were about 15-35% greater than typical differences in corrected colors collected by different expert samplers on the same image. Transforms based on "skin+gray" and "all" patches performed slightly worse, while those computed from the "base" set alone were significantly worse. Hence, it appears that dense sampling of the color region of interest (skin tones) was more important than broad sampling of the color space. We re-examine this question when evaluating skin tone classification in Section 4.2.

3. SKIN COLOR ESTIMATION

Little study has been made on how best to extract accurate skin tone information from images of faces. Prior work typically either computes the average color near the location of a face pattern detection

Table 1: Cross-Device Color Correction
Numbers indicate error measure ratios (cross-device vs. cross-expert-sampler) described in text; 1 is ideal value.

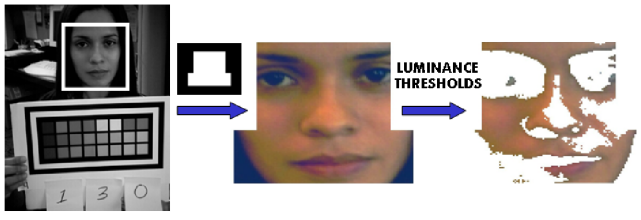
Comparison	*Skin*	*Skin+Gray*	*All*	*Base*
HP850 v. Nikon	1.15	1.23	1.42	1.67
HP850 v. IPAQ	1.35	1.59	1.59	2.17
HP850 v. Web-cam	1.18	1.27	1.37	1.63
Nikon v. IPAQ	1.36	1.44	1.43	1.98
Nikon v. Web-cam	1.14	1.28	1.40	1.74
IPAQ v. Web-cam	1.33	1.46	1.60	2.16

[3], or applies heuristics after attempting to parse facial structure [4]. We have not found prior work that evaluates the accuracy of such approaches, via comparison to ground truth.

We divide the skin color estimation problem into two parts: 1) finding the face, and then 2) sampling the face and analyzing color statistics to extract a skin tone estimate that is relatively unaffected by blemishes, shadows, and specularities. For the first part, we rely on face pattern detection, for which many methods have achieved very high success rates [14]. For most domains in which we might apply our method, we can either expect the user to present his face to the camera, or we can repeatedly sample a video stream until a suitable image is found. We use a C++ implementation of the Viola-Jones face detector [13], applied at 24 resolutions, with lenient detection thresholds and, if necessary, image rotation. When multiple detections occur in an image, we select that with largest area. We found this to have high reliability across all quality of imagery: 95-98% detection of the subject's face, for each of our four cameras. Most misses were due to faces extending past the edge of the image.

For the second part (face color extraction), we begin with the face bounding box provided by the Viola-Jones detector. The location of facial features within this box is not constant across all detections, and the box typically includes non-face background. Lighting and hair can cause shadows to fall on any part of the face, and skin texture is well known to produce large regions of specularities. Despite all this, we found that accurate estimates of skin tone could usually be obtained without detailed parsing of facial features and without segmentation of the face from the background. Specifically, we employed the following steps:

1. Apply a binary "face mask" template within the bounding box, excluding pixels that lie in the zero portions of the template. Our template excludes the outer region, to avoid background contamination, but extends somewhat widely in the lower portion where cheeks - good face color sampling locations - are likely to be.
2. Sort remaining pixels by luminance. Luminance (Y) may be computed in many ways, but we found $Y = R + G + B$ to give best cross-device consistency of results.
3. Compute the mean color of pixels ranked in the $[L..U]$ percentile in luminance, where L and U are lower and upper bounds. This excludes high-end outliers (e.g. specularities), as well as hair, nostrils, mouths, and shadowed parts of the face, all of which tend to have lower luminance than the person's true skin tone.



(a) **(b)** **(c)**

Fig. 3. Automatic face color estimation. (a) Face detection (b) Apply face mask template (c) Exclude pixels with luminance ranked outside specified bounds. Note removal of hair, eyebrows, nose and cheek shadows and specularities, and most of eyes and lips.

Table 2: Automatic Skin Color Estimation Performance
Percent differences between manual and automatic estimates, in chromaticity space of Y=R+G+B, R/Y, G/Y.

Luminance Percentile Range	*Color Component*		
	R/Y	G/Y	Y
33-98%	0.92	0.40	2.37
33-90%	0.98	0.39	2.45
50-98%	1.08	0.37	2.98

The above procedure can be performed independently of color correction; hence, applications that do not need such correction may use it. Figure 3 shows an example of face pixels selected by the method.

To test this method, we computed, over all 90 of our subjects, the average absolute difference between the automatically extracted skin tone and the mean of the manually sampled distributions collected by our expert samplers. We repeated this for three different choices of luminance bounds $[L..U]$. The results are summarized in Table 2. The differences between automatic and manually generated face color estimates were less than 3% per color component, with the luminance having the greatest error. All three choices of $[L..U]$ performed well, but using pixels with the broadest range of luminances (33-98th percentile) best matched our manual estimates, with a 50-98% range being worst. Hence, it appears more critical to include lower luminance pixels than high ones. The effect of this luminance range on skin tone classification will be examined in Section 4.2.

4. CLASSIFICATION OF SKIN COLOR

By transforming our automatically generated skin tone estimates of Section 3 into the corrected color space obtained through the calibration methods of Section 2, we hope to relate, cluster, and classify skin tones in a consistent manner, across multiple people, capture devices, and illuminants. An example classification scheme and its resulting performance are discussed below.

4.1. Ground truth estimation

To train and test methods for skin tone classification, ground truth is needed. Because there is no standard set of skin color classes, and no standard for labeling people by skin tone, we need to create our own. In effect, we need to partition our corrected color space into regions (one per class) that cover and sensibly divide the full gamut of human skin color. We cannot arbitrarily select such a partitioning, e.g. by quantizing each color space dimension into thirds, as this will trivialize classification. Definition of these regions through physical measurements of skin color, for example with a spectrophotometer, is also problematic. Although a spectrophotometer may provide more spectral information than a camera, it is still not clear where to sample skin color on the face or body, and we will need to relate the measurement space of these samples, in the end, to the corrected color space of our cameras. It is difficult to make reasonable choices in both matters that are not somehow correlated with our methods in Sections 2 and 3, and hence we would expect an artificial bias toward better success in our classification testing.

We believe that a better, more unbiased approach is to develop skin color classes based on the calibration target colors present in our images. Specifically, we label each subject according to the skin-colored patch on our target to which her manually sampled skin color

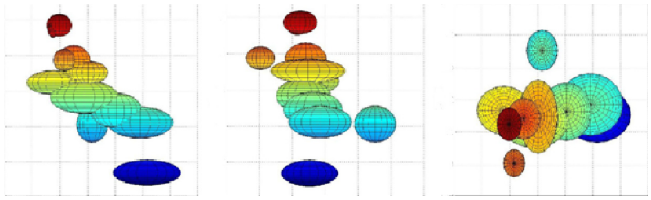


Fig. 4. 2D projections of color class data along each pair of axes in chromaticity space. Ellipses encompass one standard deviation.

distribution is closest, in the uncorrected color space of the original image, as measured by the Mahalanobis distance:

$$(\vec{p} - \vec{m}_s)^T \, V_s^{-1} \, (\vec{p} - \vec{m}_s) \qquad (2)$$

$\vec{p}$ is the patch mean, and $\vec{m}_s$ and V_s are the mean and covariance for the distribution of skin color data manually collected for a given image. This method completely avoids contamination from the color correction of Section 2, and is correlated with the automated skin sampling methods of Section 3 only to the extent that those methods adhere to the guidelines used by our human expert samplers, which we hope to be true.

We thus have 16 skin color classes, one for each patch on the lower two rows of our target. Each of our 90 subjects was assigned, as described above, to one of these classes, using images captured by the Nikon D1. Next, face colors (extracted as in Section 3) for all people were color-corrected (as in Section 2) and collected together according to class label, across all cameras, to form skin class descriptions. Figure 4 shows ellipses fitted to these classes. It is evident that no 2D projection of the data would provide good class separation, so we classify in a 3D color space. This is consistent with the best representations found for skin color in prior work [5, 6].

4.2. Classification experiments

"Leave-one-out" cross-validation was used to evaluate classification performance. For each person in turn, skin color classes are formed using data from all other people (and all cameras), and the remaining "test" person is classified against this set. The Mahalanobis distances between the test person's color data and each class are computed and sorted, with lower distances indicating a better match. The rank of the correct class is determined, and ideally should be ranked first. Multiple test images from different cameras are used for each person, and the results are averaged. To evaluate the overall classification performance, results are averaged across all people tested in this way, to produce an average rank of the correct class label. Also, we compute an ROC curve indicating how often the correct label is the top one, among the top 2, among the top 3, etc.

We performed this cross-validation for different choices of 1) color transform type (e.g. "skin", "all", etc.), 2) skin pixel luminance bounds $[L..U]$, and 3) classification color space (i.e. color-corrected sRGB, or chromaticity or Hue-Saturation-Value spaces derived from the corrected sRGB). Classification results for all combinations of these choices are shown in Table 3. The best result, as judged by average rank of the correct class label, was obtained in the corrected sRGB color space, using color transforms based only on skin-colored patches, and $[L..U]$ bounds of 33-98%. For this case, the average rank of the correct label was 1.58, with the top choice being correct 64% of the time. The correct label was among the top two 85% of the time, and never lower than fifth.

When comparing classification of images of the same person from two different cameras, the top selected labels agreed 66% of the time. Classification correctness did not drop significantly for cameras with lower image resolution, although more people became unclassifiable due to calibration color target detection failures. In

Table 3: Skin Color Classification Results
Rank of correct class label, averaged across all test subjects, for different combinations of color transforms, classification color spaces, and skin color estimation methods.
Color spaces: R = sRGB, C = chromaticity, H = HSV

Luminance	*Skin*			*Skin+Gray*			*All*		
Bounds L-U	R	C	H	R	C	H	R	C	H
33-98%	1.58	1.97	2.05	2.01	2.28	2.15	2.08	2.14	2.28
33-90%	1.61	1.89	2.06	2.05	2.21	2.34	2.29	2.51	2.63
50-98%	1.72	1.64	1.87	2.18	2.40	2.34	2.12	2.34	2.40

general, classification performed best when color transforms were based only on skin-colored target patches, declining by about 40% (in terms of average rank of correct label) when "skin+gray" or "all" patches were used. Classification in the sRGB color space was significantly better than in chromaticity, which in turn was better than in HSV. Choice of $[L..U]$ bounds had negligible effect.

We believe this is promising performance, especially given the small separation in sRGB space between the 16 skin patch colors that served as the basis for the class labels. For many tasks, a sparser density of skin classes may suffice, and performance would increase significantly. We hope to improve performance by experimenting with different classifiers and improved face color sampling methods.

5. CONCLUSION

By transforming automatically generated facial skin tone estimates into a corrected color space obtained through an easy-to-use calibration procedure, we can relate, cluster, and classify skin tones consistently - across multiple people, capture devices, and illuminants - to enable a wide variety of applications. We believe our color correction and automated skin color estimation algorithms are themselves of interest, and we validated each independently. All steps in our image processing and classification pipeline are of low computational weight, making them practical for a wide variety of systems.

6. REFERENCES

[1] E. Angelopoulou, R. Molana, K. Daniilidis, "Multispectral skin color modeling," *CVPR'01*.

[2] *Colour Engineering* Ed. by P. Green & L. MacDonald. Wiley, 2002.

[3] T. Darrell, G. Gordon, M. Harville, J. Woodfill, "Integrated person tracking using stereo, color, and pattern detection," *Intl. J. of Comp. Vision(37)*, No. 2, pp. 175-185, June 2000.

[4] T. Fukuda, S. Itou, F. Arai, "Recognition of human face using fuzzy inference and neural network," *Wkshp. Robot & Human Comm.*, 1992.

[5] F. Imai et.al., "Spectral reflectance of skin color and its applications to color appearance modeling," *Color Imag. Conf.*, 1996.

[6] S. Jayaram, S. Schmugge, M. Shin, L. Tsap, "Effect of colorspace transformation, the illuminance component, and color modeling on skin detection," *CVPR'04*.

[7] M. Jones, J. Rehg, "Statistical color models with application to skin detection," *CVPR'99*.

[8] M. Nischik, C. Forster, "Analysis of skin erythema using true-color images," *IEEE Trans. Medical Imaging(16)*, no. 6, 1997.

[9] Y. Raja, S. McKenna, S. Gong, "Tracking and segmenting people in varying lighting conditions using colour," *Proc. Automatic Face and Gesture Recognition*, 1998.

[10] H. Takiwaki, "Measurement of skin color: practical application and theoretical considerations," *J. Med. Invest(44)*, 1998.

[11] N. Tsumura, et. al., "Image-based color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin," *ACM Tran. Graphics(22)*, no. 3, July 2003.

[12] Y. Vander Haeghen, J. Naeyeart, I. Lemahieu, "Consistent digital color image acquisition of the skin," *Intl. Conf. Eng. in Med. and Bio.*, 1998.

[13] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features,", *CVPR'01*.

[14] M. Yang, D. Kriegman, N. Ahuja, "Detecting faces in images: a survey," *IEEE Trans. Pat. Anal. Mach. Intell(24)*, no. 1, 2002.