Analysis of Manual and Automated Skin Tone Assignments

K. S. Krishnapriya[1], Gabriella Pangelinan[2], Michael C. King[2], and Kevin W. Bowyer[3]

[1]Valdosta State University, Valdosta, Georgia
[2]Florida Institute of Technology, Melbourne, Florida
[3]University of Notre Dame, Notre Dame, Indiana

Abstract

The Fitzpatrick scale is a standard tool in dermatology to classify skin types for melanin and sensitivity to sun exposure. After an in-person interview, the dermatologist would classify the person's skin type on a six-valued, light-to-dark scale. Various face image analysis researchers have recently categorized skin tone in face images on a six-valued, light-to-dark scale in order to look into questions of bias and accuracy related to skin tone. Categorization of skin tone on the basis of images rather than personal interview is not, on that basis alone, strictly speaking, on the Fitzpatrick scale. While the manual assignment of face images on a six-point, light-to-dark scale has been used by various researchers studying bias in face image analysis, to date there has been no study on the consistency and reliability of observers assigning skin type from an image. We analyze a set of manual skin type assignments from multiple observers viewing the same image set and find that there are inconsistencies between human raters. We then develop an algorithm for automated skin type assignments, which could be used in place of manual assignment by observers. Such an algorithm would allow for provision of skin tone annotations on large quantities of images beyond what could be accomplished by manual raters. To our knowledge, this is the first work to: (a) examine the consistency of manual skin tone ratings across observers, (b) document that there is substantial variation in the rating of the same image by different observers even when exemplar images are given for guidance and all images are color-corrected, and (c) compare manual versus automated skin tone ratings. We release the automated skin tone rating implementation so that other researchers may reproduce and extend the results in this paper.

1. Introduction

The problem of skin tone analysis from images and how skin tone affects face recognition and face analytics algorithms is currently of high interest. A significant step to understanding the impact of skin tone on face recognition and analytics accuracy is the retrospective estimation of skin tone from face images. Many existing large-scale face image datasets available to the research community generally do not have skin tone annotations included as metadata. In this paper, we report on an experiment to measure skin tone from face images in accordance with a Fitzpatrick-inspired scale and individual typology angle.

The Fitzpatrick scale is a six-valued, I (lighter) to VI (darker) skin tone rating that is widely used in dermatology [5] and dates back over 30 years [17]. Traditionally, in dermatology, Fitzpatrick skin type is assigned in-person by a trained practitioner, and a Fitzpatrick rating of this type might be considered a measure of "true" skin tone. Various researchers in face image analysis have recently used a six-valued, light-to-dark scale for rating face images [8, 31, 33, 25]. When such a rating is assigned retrospectively to images, whether manually or by an algorithm, it does not meet the strict definition of a Fitzpatrick rating. However, all six-valued light-to-dark skin tone ratings obviously have that much in common with Fitzpatrick ratings. Some recent studies in face image analysis has used a six-point, light-to-dark skin tone rating and made no reference to the Fitzpatrick scale [24, 31] while others have made explicit reference to Fitzpatrick [8, 26, 25]. Since recent research in face image analysis is coming some 30+ years into the use of the Fitzpatrick scale in dermatology, it seems unlikely that recent researchers were unaware of the Fitzpatrick scale when they chose to use a six-point, light-to-dark rating. In this paper, we analyze "apparent" skin tone (that is, the skin tone apparent in a given image), using a Fitzpatrick-inspired six-tone rating scheme which we call the Apparent Skin Tone (AST) scale. Our proposed AST

scale is in alignment with the Fitzpatrick scale, with the understood caveat that only apparent skin tone is measurable from images.

We also use individual typology angle (ITA) to measure apparent skin tone. Prior research [14, 32] assessed skin tone from images using an automated ITA measurement. ITA measurements are, again, categorized on a six-valued, light-to-dark scale - very light, light, intermediate, tan, brown, and dark.

There are three major contributions of this work. First, we present an analysis of the consistency of human rating of skin tones from images. This analysis suggests that categorical labeling of skin tone in images by human observers is subjective, and there is a level of inconsistency across observers. Second, considering the cost and time associated with manual ratings, along with the level of inconsistency, we develop an automated approach for skin tone assessment based on ITA. Automated ratings produce a level of agreement with manual ratings that is similar to the level of consistency between multiple human raters using the AST scale. The proposed algorithm presents an alternative to manual ratings and thus has obvious advantages in speed, scalability, cost, and consistency. Third, to encourage transparency and reproducibility of the experimental results, the automated skin tone assignment implementation is made available to other researchers.

2. Related Work

The Fitzpatrick scale [17] is a I (lightest) to VI (darkest) rating of skin tone, used in dermatology to classify sensitivity to sun exposure. Skin tone, of course, varies among African-American individuals and among Caucasian individuals. Additionally, face morphology varies by race, gender, and individual, independent of skin tone variation. Lester et al. [28] reviewed the research literature on Covid-19 skin manifestations in the context of the skin tone of subjects represented in research studies. A set of images from the literature were given Fitzpatrick skin tone categorization by a board-certified dermatologist [28] "with expertise in diagnosing and treating patients with skin of colour (Fitzpatrick type IV-VI)". Lester et al. also spoke to the subjective nature of skin type assessment from photographs, commenting that lighting conditions "may have led to some misclassification across one or two skin types" [28].

The Lester et al. [28] study underscores several important points. Their research on important current research issues is performed using a (single) board-certified dermatologist's subjective assessment of Fitzpatrick skin type from photos. This is evidence that, whatever shortcomings subjective Fitzpatrick ratings from images have, there is not yet anything better to replace them. Furthermore, the uncertainty due to varying illumination between images is acknowledged as possibly causing skin type misclassifications. Our experience with multiple observers rating the same set of controlled-acquisition face images is consistent with Lester et al. [28] on this point - the more varied the illumination in a set of images, the larger the potential misclassification range. In comparison to imagery in dermatology research publications, in-the-wild imagery should be expected to have even more misclassifications of skin type. In the context of our research in this paper, controlled-acquisition images such as those in the MORPH dataset will have less serious skin type misclassifications than would any of the in-the-wild datasets popular in face recognition research.

The use of skin tone ratings for face image analysis in the computer vision community appears to have started with the IARPA Janus dataset [24]. The Janus face image datasets [31] have meta-data for six-valued, light-to-dark skin tone ratings obtained via Amazon Mechanical Turk, a crowd-sourcing website for completion of discrete tasks (here, assignment of skin tone ratings to face images). However, [31] provides no basis for how the six skin tone levels were derived. Lu et al. [30] analyzed the Janus dataset and reported that recognition accuracy generally degraded with darker skin tone. Buolamwini and Gebru's [8] Gender Shades study appears to have been the first to explicitly state the name of the scale used as being Fitzpatrick. A "board-certified surgical dermatologist" provided the definitive labels for the Fitzpatrick skin type for images collected off the web in the Gender Shades study [8] reported that each of the three gender classification tools studied was more accurate for lighter skin types than for darker. There are also studies in the literature which are contradictory to these findings. For example, Muthukumar et al. [34] followed up with another study on gender classification tools and suggested that skin tone may not be the driving factor for accuracy differences. Krishnapriya et al. [25] analyzed the distribution of skin tone ratings for images sampled from the center and from the high-similarity tail HST of the impostor distribution for African-American males. They reported that same-skin-tone image pairs occur more frequently in the HST of the impostor distribution, but that darker skin tone does not appear to be a driving factor [25].

Cook et al. [13] analyzed recognition accuracy differences based on race meta-data and on a measure of skin reflectance. Exploiting the 18% gray background in a controlled enrollment image (similar to that in MORPH), they computed a measure of relative skin reflectance for each subject. They reported that darker skin tone is associated with longer image acquisition times and lower similarity scores for genuine image pairs, and that the skin reflectance measure was a better predictor than self-reported race labels. Groh et al. [20] annotated dermatology clinical images using Fitzpatrick skin type labels as well as ITA. They trained a deep neural network model to classify skin con-

ditions and reported that the skin type in the images on which a model was trained affected the accuracy scores across Fitzpatrick skin types. Bahmani et al. [6] developed a data-driven skin color measure by leveraging the dichromatic reflection model and considering different illuminations across the face. They reported that their approach over black and white subjects with uncontrolled illumination produced a meaningful progression from darker to lighter skin tone without relying on consistent background, illumination or camera sensitivity.

Howard et al. [22] collected face images from 345 subjects. These were analyzed in the $L^*a^*b^*$ color space and characterized with Face Area Lightness Measures (FALMs) to capture the intensity of reflected light on face skin. The FALMs and self-reported Fitzpatrick types were compared to ground-truth measurements from a calibrated dermatological device. The authors report that (1) intra-subject FALMs vary significantly by image and (2) appropriate estimation of skin tone from images requires controlled-acquisition images corrected for neutral grey background.

A number of studies have used six-values, light-to-dark skin tone ratings assigned by viewers from examining images (rather than in-person interviews) [28, 8, 31, 33, 34]. However, there is minimal research on the inter-rater agreement or overall consistency of the ratings. Many of the studies involve a single observer [28, 8, 33, 34] and a few had multiple raters [24, 25, 21] (with a maximum of eight). Unlike the IARPA Janus dataset skin tone annotations, the study in [25] specified how they merge multiple observer ratings to a rating for the image. Scholars used the ITA [20, 32, 14] for skin tone measurements and a six-level categorization as specified in this paper. None of these previous studies have documented the level of agreement that can be expected from two or more observers assigning six-tone ratings to the same images. In addition, they have not (1) investigated consistency on color-corrected images with exemplar images provided or (2) compared any automated approach to manually assigning skin types.

3. Dataset and Preprocessing

We used the MORPH dataset [2] containing mugshot-style images for this study. MORPH was originally assembled and distributed to support research in face aging [35]. African-American males comprise the largest cohort of MORPH, with 36,838 images of 8,850 subjects in the curated version used by Krishnapriya et al. [26]. We focused on this single cohort to analyze skin tone data independent of factors like gender and race. MORPH'S set of African-American male subjects is larger than IJB-C's set [31] of 3,531 subjects and the 562-subject set used by Cook et al. [13] (which, unlike the IJB-C and MORPH datasets, is unavailable to the research community).

In this experiment, we sample 500 image pairs from two different regions of the African-American male impostor distribution: the center and the high-similarity tail. (The motivation for sampling these two regions of the impostor distribution is that no pairs in the center are in danger of being false matches, and by definition the HST is where the false-match pairs exist. It is important to understand whether or not there is a difference in skin tone distribution in these two regions.) Each of these two sets of image pairs has nearly 1,000 unique images, as some individual images are repeated in multiple image pairs (21 in common). There are 982 unique images of 915 persons from the center and 967 unique images of 872 persons from the HST.

Of these two image sets, 13 from the center and 33 from the HST had insufficient background - that is, there were not enough visible background pixels for color-correction. The two groups were combined for pre-processing with one set of the 21 common images removed. During Dlib face detection and cropping for automated ratings, 11 images had a "failure to detect" result. Ultimately, we presented 1871 total images (959 from the center and 933 from the high similarity tail minus one set of the 21 common images) for rating by the six raters and automated system. Black rectangles are added over the eye regions of example images shown in this paper in an effort to protect individual anonymity and privacy.

3.1. Color Correction

MORPH images are acquired in a controlled environment with the subject standing in front of an 18% gray background. This experiment is designed to normalize the face images so that the 18% gray region is the same on average across all images. The motivating hypothesis is that varying color quality between images may contribute to inconsistent manual Fitzpatrick ratings.

The 18% gray is defined based on reflection, i.e. an 18% gray surface reflects 18% of the light that hits it [3]. The idea of 18% gray in photography is to achieve middle gray to human perception. In different color spaces, middle gray may be defined differently; for example, in CIELAB, middle gray is defined to be 46.6% brightness [19], while in 24-bit color space, it is given by RGB (119, 119, 119) [1]. This experiment uses the latter measure.

We must also remove gamma non-linearity from images prior to correcting color. PC monitors have intrinsic non-linearity and apply a power of 2.2, denoted gamma, for displaying images [16]. In order to make true 18% as middle gray, as $0.18^{1/2.2} * 255 \approx 119$, we calculate the color-correction factor of each image with non-linearity removed.

The color-correction steps, with examples given for the R color channel, are as follows:

- Semantic segmentation of person and background in the given face image using a pre-trained model.

- Compute the linear version of the image by raising each pixel to $\gamma = 2.2$:
  $(R, G, B) = (R^{\gamma}, G^{\gamma}, B^{\gamma})$
- Extract all (R, G, B) pixel values corresponding to the linear background.
- Find the mean linear background pixel value:
  R_{avg} = mean(all R components of linear background)
- Compute the color-correction factor based on 18% gray background:
  R_{const} = $0.18/R_{avg}$
- Apply the color-correction factor to all pixels in the linear image:
  $R_{corrected}$ = R_{const} * R_{linear}
- Finally, add back non-linearity:
  $R_{final} = R_{corrected}{}^{1/2.2}$

We used pre-trained model DeepLab V3 [11] to segment the person and the background. DeepLab has Xception [12] as its network backbone and was pre-trained on ImageNet [15]. Subjectively, color-correction improves the visual quality of the original image and makes the background more consistent across images (see Figure 1). Additional analysis is planned to assess the degree to which the color-correction step improves consistency of auto-ratings across multiple images of the same individual.

Figure 1: Original (top) and color-corrected images (bottom).

4. Manual Assignment of Skin Tones

The manual ratings task was conducted in the same laboratory environment for all raters with the color-corrected images displayed on two monitors. Six different viewers independently examined the same set of 1,871 images to assign an AST score to each face. Alongside the ratings images, a set of exemplar images were displayed to encourage the use of consistent reference points (see Figure 2). Exemplars were selected from the well-known IJB-C dataset (consisting of 3,531 subjects with 31,334 images) that has per-subject skin tone annotations in its metadata [31]. The six raters each assigned an AST score without knowing the region of the impostor distribution that an image came from and without knowing each others' ratings.

Figure 2: Exemplar images of AST ratings I-VI.

Figure 3 shows inter-rater agreement via the difference between the maximum and minimum ratings for each image. For the 1871 images rated by all six raters, 2.5% of images had no difference in rating among the six raters. The six raters agreed within one-skin-tone-difference from min to max rating for 966 images (51.6%) and within two-skin-tone-difference for 1604 images (87.7%).

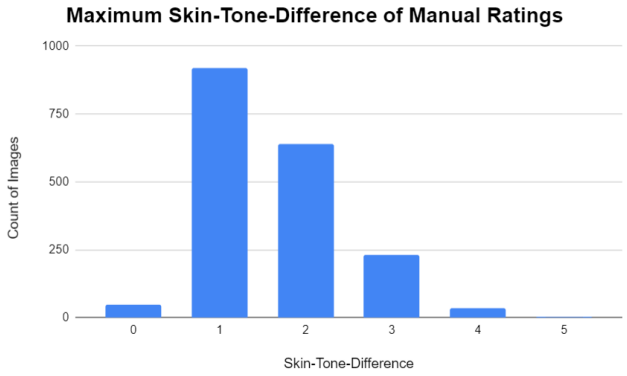


Figure 3: Count of images with given skin-tone-difference in each manual rating set.

In determining how best to combine the six individual ratings ("rating set") for each image into a consensus manual rating ("CMR"), we considered both rounded mean and mode of each set. Figure 4 shows the level of difference between automated ratings and CMR ("AvM Diff") using rounded mean versus mode of each rating set, as well as the mean, median and mode (metrics) of all ratings with the respective determination of CMR. Since the metrics of using rounded mean versus mode for the CMR are highly similar, we chose to use rounded mean to minimize the impact of misclicks (i.e. unintentional skin tone misclassifications) from individual raters on a given rating set. In the following sections of the paper, CMR refers to the rounded mean of the rating set of a given image.

AvM Diff	Mean as CMR		Mode as CMR	
	Count	%	Count	%
0	466	24.91	410	21.91
1	913	48.8	922	49.28
2	392	20.95	413	22.07
3	91	4.86	112	5.99
4	7	0.37	11	0.59
5	2	0.11	3	0.16

Metric	Mean as CMR
Mean	5.25
Mode	5
Std Dev	0.65

Metric	Mode as CMR
Mean	5.35
Mode	5
Std Dev	0.66

Figure 4: Analysis of rounded mean versus mode as CMR.

5. Automated Assignment of Skin Tones

Automatic estimation of skin tone from face images would give ratings that are 100% consistent across multiple runs on the same image. Different research groups using the same software on the same images would get the same ratings, which cannot be said of CMR. Automated ratings ("auto-ratings") would also be cheaper and faster to acquire than CMR, enabling larger-scale experimental studies.

The ITA is calculated in the CIELab color space where L represents the lightness, a represents the chromaticity coordinate from green to red, and b represents the chromaticity coordinate from blue to yellow. In this approach, we utilized ITA for representing the skin color [10], and it is calculated according to equation 1.

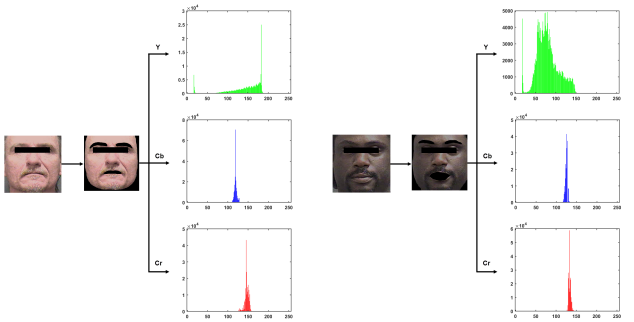
$$ITA = \frac{arctan(\frac{(L-50)}{b}) * 180}{\pi} \quad (1)$$

ITA measurements are categorized into six skin type groups - very light (skin type I), light (skin type II), intermediate (skin type III), tan (skin type IV), brown (skin type V), and dark (skin type VI).

5.1. Implementation Workflow

The selection of suitable color space for skin detection is an important factor in determining a higher probability of success. Prior research [9, 36] has shown that the $YCbCr$ color space is preferable to RGB for color segmentation analysis because in the latter space the brightness (luminance) component is not decoupled from the color information (chrominance). $YCbCr$ is also preferred over HSV for straightforward transformation and efficient separation of color and intensity information even for images with nonuniform illumination conditions. We can effectively utilize the chrominance information in $YCbCr$ color space for modeling the human skin color, and hence we propose thresholding on $YCbCr$ color space channels for skin detection. The Y channel representing the brightness cannot be constrained here because when we are evaluating different datasets, and the images can be taken in different lighting conditions. Hence, with the Y channel, it is difficult to determine if the variation in distribution is caused by different skin color or different lighting condition. In Figure 5, we can see similar Cr and Cb distributions of skin color for Caucasian (see Figure 5a) and African-American (see Figure 5b) images, and they do not seem to be affected by the variations in luminance. The evaluation of Cb and Cr channels across different sets of images showed that they are consistent across different demographic groups, and hence we can achieve better skin detection based on the thresholding on those two channels from an input image. The ranges mentioned for Cb and Cr in equation 2 were found to be the most suitable and representative of skin color for different sets of images we have tested, and slight variations of these ranges were also found to be adopted in many human skin detection studies [9].

$$\text{pixel} = \begin{cases} \text{skin}, & \text{if } 136 \leq \text{Cr} \leq 173 \& 77 \leq \text{Cb} \leq 127 \\ \text{non-skin}, & \text{otherwise} \end{cases} \quad (2)$$



(a) (b)

Figure 5: Frequencies of Y, Cb, and Cr values across face skin pixels for (a) Caucasian (b) African-American.

This automated approach utilizes color-corrected images for skin tone assignments. The BiSeNet (Bilateral Segmentation Network) [37] model used for the face skin segmentation task was pre-trained on the CelebAMask-HQ [27] dataset that has 30,000 face images from CelebA [29] and CelebA-HQ [23]. The eyes and lips regions are masked out intentionally to avoid any noise or occlusions like sunglasses while estimating the actual skin tone. The extracted face skin may contain over-exposed or under-exposed skin pixels due to illumination conditions. Thresholding on the $YCbCr$ color space is done to select the best skin pixels that are representative of a person's skin tone. The auto-rating workflow steps are:

1. Face detection and cropping on color-corrected images using Dlib face detector (Figure 6a).

2. Semantic segmentation of face skin from detected face using BiSeNet pre-trained model (Figure 6b).

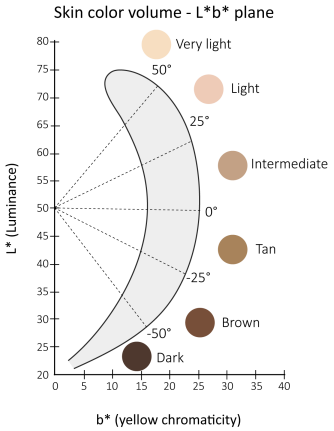
3. Conversion to $YCbCr$ color space and application of thresholding given in equation 2 on Cb and Cr channels for skin pixel selection (Figure 6c).

4. Calculation of mean pixel value and corresponding ITA using equation 1.

5. Mapping final ITA measurement to skin type group: light (type I) to dark (type VI) (Figure 6d).

(a) (b) (c) (d)

Figure 6: Auto-rating workflow: (a) face detection, (b) face skin segmentation, (c) Cb and Cr thresholding, (d) skin tone estimate.



Individual Typology Angle (ITA°)	Skin Type
ITA° >= 50°	I
25° <= ITA° < 50°	II
0° <= ITA° < 25°	III
-25° <= ITA° < 0°	IV
-50° <= ITA° < -25°	V
ITA° < -50°	VI

Figure 7: ITA classification thresholds based on minimal overlap of CMR mapped onto the standard ITA scale.

6. Manual versus Automated Assignments

We devised the Apparent Skin Tone (AST) scale to mimic the in-person Fitzpatrick scale for measurement of apparent skin tone from a face image. The AST scale is communicated to raters via exemplar images shown in Figure 2: one male and one female for each skin type (I-VI) as specified in IJB-C annotations [31]. These twelve images were manually selected by the authors with the goal of having a consistent set of exemplars. Raters are asked to reference the presented face images against the exemplars in their determinations of skin type.

We compared the consistency of CMR to automated ITA values with ranges for skin types I-VI specified in [10]. Mapping the CMR-assigned images on the ITA scale showed that there is substantial overlap across the ITA ranges; all images with the same CMR do not fall into a single corresponding ITA category. In order to relate AST-based CMR and auto-assigned ITA values, we determined custom threshold ranges for the ITA skin types, as shown in Figure 7. These thresholds were selected to minimize overlap between CMR values mapped onto the standard ITA scale.

6.1. Inter-Rater Variability

CMR and auto-ratings were assigned for 1871 images from the center and HST as previously described. Figure 8 shows the distributions of skin tone ratings for center and HST images. CMR and auto-ratings are remarkably consistent regardless of the image's location in the impostor distribution.

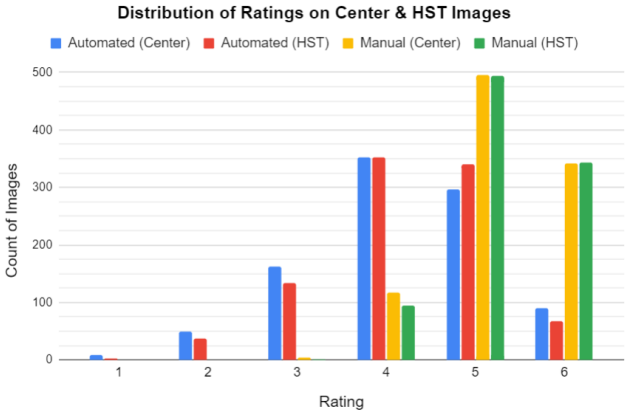


Figure 8: Distribution of CMR and auto-ratings on center and high-similarity tail images.

Figure 9 shows the count of skin-tone-difference between CMR and auto-rating values for images from the two regions of the impostor distribution. For center images, ratings agreed on 24.6% of images, within one skin-tone-difference for 71.9% of images and within two for 94.5% of images. For HST images, ratings agreed on 25.1% of images, within one skin-tone-difference for 75.5% of images and within two for 95% of images. These two results suggest that region of impostor distribution has little effect on CMR and auto-rating. That is, congruent with [25], we find no clear evidence to support the idea that images with darker skin tone ratings are more frequent in the false match pairs. However, this deserves to be examined with a more substantial image set in future studies.

For the entire set of 1871 images, CMR and auto-ratings agreed on 73.7% of images within a one-tone difference and 94.7% within a two-tone difference. We also see from Figure 8 that CMRs were higher in general than auto-ratings. In fact, the mean auto-rating is 4.2 (with standard deviation of 1 and mode of 4), while the mean CMR is 5.2 (with standard deviation of 0.65 and mode of 5).

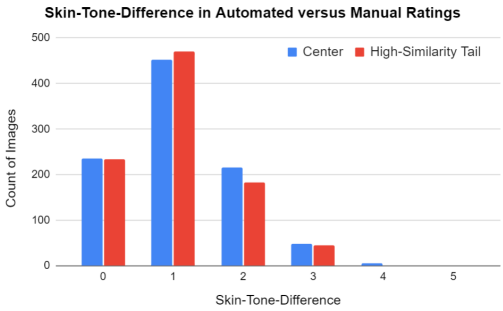


Figure 9: Count of images with given skin-tone-difference in CMR and auto-ratings.

One potential reason CMR tended toward higher values than auto-ratings is the presented image. Manual raters were shown the entire color-corrected image, while the automated system was presented only with skin pixels of the face ("face mask"). Figure 10 shows four images with a four-skin-tone difference between CMR and auto-ratings. The three face images in the top row had CMRs of 6 on the AST scale. Their corresponding face masks in the bottom row were given ITA-based auto-ratings of 2.

Figure 10: Full face images for AST-based manual ratings and corresponding face masks for ITA-based auto-ratings.

We speculate that manual raters consider skin tone in context - that is, considering factors like illumination and shadows. The four subjects in Figure 10 are illuminated on their forehead, cheek and nose regions. In the more shadowed regions of the images (e.g. jaw and neck), skin appears visibly darker. Manual raters may look at these shadow regions to better assess skin tone, understanding that highlighted regions appear lighter than the rest of the subject's skin. However, the automated rating system only received skin pixels from the face (which already tend to be lighter due to illumination) and calculated ITA based on mean pixel value alone, without consideration of other factors affecting apparent skin tone.

6.2. Intra-Rater Variability

The intra-rater variability is evaluated here by examining the consistency of CMR and auto-ratings on different images of the same individual. In theory, a subject's apparent skin tone should be the same across multiple instances since the images were taken in the same controlled 18% grey setting and subsequently color-corrected. In practice, however, variations in illumination and pose can cause the subject's skin tone to appear lighter or darker in each instance.

Of the 1,638 subjects represented in the image set, only 25 were represented in 3-4 image instances. The subset of the MORPH data set used in this study did not contain a sufficient number of multi-image subjects for a statistically significant assessment of intra-rater variability. Anecdotal examples, however, can provide insight into a limitation of the automated system.

Table 1 gives the auto-ratings (A) and CMR (M) for the five subjects with four image instances. For each subject, CMR is either the same or within one-skin-tone difference across all four of their image instances. Auto-ratings, alternately, vary by one and two-skin-tone differences across all same-subject instances.

Subject	Image	Rating (A)	Rating (M)
1	A	4	6
	B	2	5
	C	4	6
	D	4	5
2	E	6	6
	F	5	6
	G	6	6
	H	6	6
3	I	3	5
	J	4	5
	K	3	5
	L	5	5
4	M	5	6
	N	4	6
	O	5	6
	P	5	5
5	Q	5	6
	R	5	6
	S	4	5
	T	5	6

Table 1: Same-subject skin tone ratings across multiple image instances.

Highlighted in the first two rows of Table 1, CMR and auto-ratings for Subject 1 were least consistent across images and most consistent for Subject 2. The visual differences between Subject 1's four images (shown in the top row of Figure 11) are apparent. The backgrounds of images 1B and 1C vary in shade from the consistent 18% gray visible in images 1A, 1D, and 2E-H; this may have resulted from the erroneous inclusion of hair (1B) and shirt (1C) pixels in the segmented background used for the color-

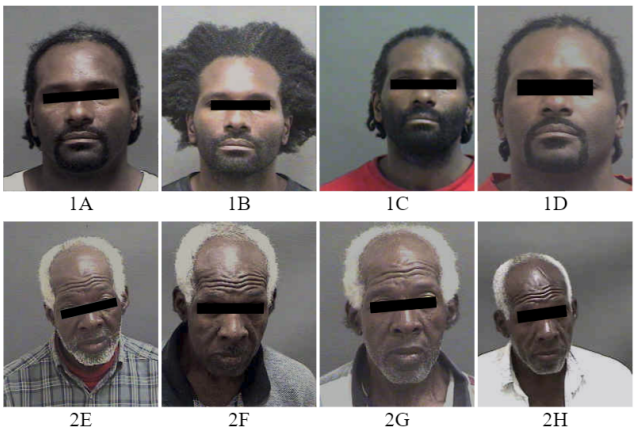


Figure 11: Subject 1 images A-D and Subject 2 images E-H from Table 1.

correction step. Additionally, while all of Subject 1's images have some highlights in the forehead, nose and cheek regions, image 1B in particular is brighter and has wider highlight regions than the other images. While the CMR is relatively consistent across the four image instances of Subject 1 (two ratings of skin type 6, two of type 5), the auto-rating is affected by the increased illumination in 1B, giving a rating of 2 versus otherwise consistent 4s (1A, 1C, 1D).

Subject 2, alternately, was consistently rated across nearly all image instances. Visually, Subject 2 has very dark skin. Logically, lighting variations for subjects with skin tone at the extremes of any tone-rating scale (very light or very dark) are less likely to cause a shift in tone assignment than subjects with middling skin tones. On the ITA scale in Figure 7, for example, types III-V have 25° ranges, while types I and VI include *any* value above 50° or below -50°, respectively.

7. Conclusion and Discussion

This paper systematically analyzes approaches to estimate a person's skin tone from an image with an 18% gray background using the Morph dataset. It describes a method for manual rating and proposes an automated approach for greater ease of use, scalability, and reproducibility.

The categorical labeling of skin tone by human observers can be subjective and inconsistent. The same images may be rated differently by different raters. Several studies mention that labeling is subjective even by trained practitioners [7, 4]. Allowing for one- and two-skin-tone-differences, manual raters agreed on 51.6% and 87.7% of images, respectively (see Figure 3).

While rating skin tone from a color image may seem simple in theory, it is a challenging task in practice. Prior research has been conducted on re-purposed images with skin tone rating assessed by humans on a six-valued, light-to-dark scale (as with the Fitzpatrick scale [18, 17]), and more recently by computer-based individual typology angle measurement [32]. Both of these techniques are flawed in that the human ratings (for example, see Figure 12) and automated ratings (Figure 13) are often on non-ideal images that have been taken in non-controlled environments.

Figure 12: Example misclassifications in the IJB-C dataset.

Figure 13: Example misclassifications in the DiF dataset. All given ITA values map to the"dark" skin type.

The effort to automate the process of skin tone labeling from images is complex. The Lester et al. study [28] states that "it is unlikely that lighting issues alone would result in skin types V or VI appearing as skin type I–III." In our experiment, this statement does seem true for consensus manual ratings but not for automated ratings, where variations in lighting appeared to be a determining factor in ITA value (see Figure 8). Furthermore, human raters may have grown aware of the fact that they were viewing images of African-American individuals and considered this factor in their assignment of skin tone ratings. In future studies, only a segment of skin from a given image will be shown to manual raters to reduce bias and remove image context.

The analysis of ITA distributions for images consistently rated as I to VI with CMR shows promise in our quest to perform retrospective labeling of images using the AST scale as a guide. The results, however, will be inherently noisy. We have not found strong evidence to support that this noise is due to the experience level of the raters but more likely can be attributed to the range of colors for skin-related pixels that are induced by varying lighting conditions and sensor characteristics. The automated-to-consensus-manual consistency is as good as the consistency between individual raters. However, the efficacy of the automated approach on other datasets (e.g. in-the-wild face images) is unknown and will be assessed in future work.

8. Acknowledgements

The authors express their thanks to the reviewers for providing numerous suggestions that resulted in improvements in the content and presentation of the paper.

References

[1] Cie color calculator. *brucelindbloom*. http://www.brucelindbloom.com/index.html?ColorCalculator.html.

[2] Morph dataset. https://www.faceaginggroup.com/morph.

[3] John Aldred. What is middle grey and why does it even matter?, 2018. https://www.diyphotography.net/what-is-middle-grey-and-why-does-it\even-matter/.

[4] Olasimbo Ayodeji Arigbabu, Sharifah Mumtazah Syed Ahmad, Wan Azizun Wan Adnan, and Salman Yussof. Recent advances in facial soft biometrics. *The Visual Computer*, 31(5):513–525, 2015.

[5] O. Arosarena. Options and challenges for facial rejuvenation in patients with higher fitzpatrick skin phototypes. *JAMA Facial Plastic Surgery*, 2015.

[6] Keivan Bahmani, Richard Plesh, Chinmay Sahu, Mahesh Banavar, and Stephanie Schuckers. Sreds: A dichromatic separation based measure of skin color. *arXiv preprint arXiv:2104.02926*, 2021.

[7] Diana Borza, Adrian Sergiu Darabant, and Radu Danescu. Automatic skin tone extraction for visagism applications. In *VISIGRAPP (4: VISAPP)*, pages 466–473, 2018.

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research 81: Conference on Fairness, Accountability, and Transparency*, 2018.

[9] Douglas Chai and King N Ngan. Face segmentation using skin-color map in videophone applications. *IEEE Transactions on circuits and systems for video technology*, 9(4):551–564, 1999.

[10] Alain Chardon, Isabelle Cretois, and Colette Hourseau. Skin colour typology and suntanning pathways. *International journal of cosmetic science*, 13(4):191–208, 1991.

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[12] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR 2017*.

[13] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 40(1), 2019.

[14] S Del Bino, J Sok, E Bessac, and F Bernerd. Relationship between skin response to ultraviolet exposure and skin color type. *Pigment cell research*, 19(6):606–614, 2006.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[16] Graham Finlayson and Gerald Schaefer. Hue that is invariant to brightness and gamma. In *Proceedings of the 12th British Machine Vision Conference*, 2001.

[17] T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of Dermatology*, 124(6):869–871, 1988.

[18] U.S. Food and Drug Administration. Your skin. https://www.fda.gov/radiation-emitting-products/tanning/your-skin, (last accessed July 2020).

[19] Scott Geffert. Adopting iso standards for museum imaging. *imagingetc.com, Inc*, 2008.

[20] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *arXiv preprint arXiv:2104.09957*, 2021.

[21] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *arXiv preprint arXiv:arXiv:2104.02821*, 2021.

[22] J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Reliability and validity of image-based and self-reported skin phenotype metrics. *arXiv preprint arXiv:2106.11240*, 2021.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[24] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR 2015n*.

[25] K.S. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K.W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1), 2020.

[26] K.S. Krishnapriya, K. Vangara, M. C. King, V. Albiero, and K.W. Bowyer. Characterizing the variability in face recognition accuracy relative to race. In *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[27] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR 2020*.

[28] JC Lester, JL Jia, L Zhang, GA Okoye, and E Linos. Absence of images of skin of colour in publications of covid-19 skin manifestations. *British Journal of Dermatology*, 183(3):593–595, 2020.

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV 2015*.

[30] B. Lu, J. Chen, C. D. Castillo, and R. Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 40(1), 2019.

[31] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark - c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018.

[32] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[33] Vidya Muthukumar. Color-theoretic experiments to understand unequal gender classification accuracy from face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[34] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R. Varshney. Understanding unequal gender classification accuracy from face images. In *https://arxiv.org/abs/1812.00099*, 2018.

[35] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition*, 2006.

[36] Khamar Basha Shaik, P Ganesan, V Kalist, BS Sathish, and J Merlin Mary Jenitha. Comparative study of skin color detection and segmentation in hsv and ycbcr color space. *Procedia Computer Science*, 57(12):41–48, 2015.

[37] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.