The CHROMA-FIT Dataset: Characterizing Human Ranges of Melanin For Increased Tone-awareness

Gabriella Pangelinan, Xavier Merino, Samuel Langborgh, Kushal Vangara,
Joyce Annan, Audison Beaubrun, Troy Weekes, Michael C. King
Florida Institute of Technology
Melbourne, FL, USA
{gpangelinan, slangborgh2021, kvangara2015, jannan2021, abeaubrun2013}@my.fit.edu
{xmerino2012, tweekes, michaelking}@fit.edu

Abstract

The disparate performance of face analytics technology across demographic groups is a well-documented phenomenon. In particular, these systems tend toward lower accuracy for darker-skinned individuals. Prior research exploring this asymmetry has largely relied on discrete race categories, but such labels are increasingly deemed insufficient to describe the wide range of human phenotypical features. Skin tone is a more objective measure, but there is a dearth of reliable skin tone-related image data. Existing tone annotations are derived from the images alone, either by human reviewers or automated processes. However, without ground-truth skin tone measurements from the subjects of the images themselves, there is no way to assess the consistency or accuracy of post-hoc methods. In this work, we present CHROMA-FIT, the first publicly available dataset of face images and corresponding ground-truth skin tone measurements. Our goal is to provide a baseline for tone-labeling methods in assessing and improving their accuracy. The dataset comprises approximately 2,300 still images of 209 participants in indoor and outdoor collection environments.

1. Introduction

Face analytics techniques like recognition and gender classification have well-documented accuracy disparities with regard to race [3, 11, 26, 28]. Despite myriad research efforts, little substantial progress has been made in characterizing the reasons for such challenges. As this body of work grows, it has been noted that race labels (provided as metadata for many datasets) may be insufficient to characterize people in an increasingly interracial and intercultural world. Discrete race categories like African American / Black and Caucasian / White have been criticized for obscuring "the immense phenotypical heterogeneity that exist within them." [24] Additionally, such labels rely on subjective identification, which "can change over time, place, and context." [1]

Skin tone is a more objective point of reference, but obtaining reliable measurements has proven challenging. Historically, manual assignments have been given by trained professionals (e.g., dermatologists) using the Fitzpatrick Skin Type (FST) scale in face-to-face evaluations. The datasets that include manually annotated FST values are typically limited to close-up skin photographs intended for clinical diagnosis. It is important to note that in-person FST-annotated datasets are exceedingly scarce, with less than 2% of available datasets [7, 22, 27] reporting the rating [34], even among those curated for cancer diagnostics.

Many of the face image datasets with FST values instead used manual reviewers [3, 10, 13, 21, 33], often without specialized training in skin tone classification. This raises concerns about the validity of such assignments since FST was originally conceived for application in a clinical setting. Critics argue that using FST values outside of this context is inappropriate, particularly because FST is not designed for photometric assessments which can be heavily influenced by factors like lighting and subject pose [14].

Automated methods can also be used to estimate skin tone directly from images, though they often require specialized collection conditions and pre-processing. Individual Typology Angle (ITA) [4] is perhaps the most well-known method [8, 9, 23, 31, 33]. ITA applies a mathematical formula to image information in the $L^*a^*b^*$ colorspace. As an inexpensive and highly-reproducible process, ITA is suitable for analyzing large-scale datasets. Additionally, unlike ratings from human reviewers, ITA ratings are by nature

consistent on the same image over multiple runs. However, labeling errors can occur [17], especially for face images taken in uncontrolled environments with poor exposure. More recent methods incorporate factors like skin reflectance [2, 5] and hue angle [31] in assessing tone, though they are still susceptible to variations in illumination. Without ground-truth—the measurement of a person's actual skin tone—only manual inspection can catch such errors.

Regardless of method, tone annotations cannot truly be verified without an in-person measurement, as noted in [14]. To this end, we present a dataset with the metadata necessary for such a task. In Section 2, we review previous tone-aware datasets. In Section 3, we describe the collection of the CHROMA-FIT dataset and its included metadata. In Section 4, we detail the mapping of collected RGB / $L^*a^*b^*$ measurements to relevant tone-classification scales and introduce additional skin absorbance characteristics.

2. Related Datasets

Many works examining demographic disparities in face recognition—and the datasets they analyze—rely on discrete race labels. Law enforcement mugshot databases may use self-reported race taken from driver's licenses, as in the well-studied MORPH dataset [29] and the self-collected dataset used in [16]. Race labels can also be generated by attribute classifiers [18] or crowd-sourced [12, 19].

Skin tone annotations are rarer: Table 1 lists publicly available face image datasets with skin tone labels. Four of the datasets provide only manually annotated skin tone assignments, one provides only automated ITA values for skin tone, and two provide both measures. However, all assignments were provided post-collection, by third-party observers viewing only the images or videos, without in-person evaluations.

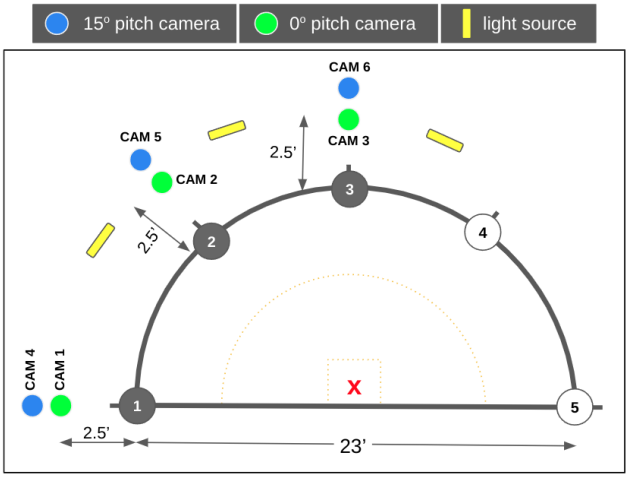
BUPT-Globalface and BUPT-Balancedface, introduced in [33], contain 2 million images of 38k persons and 1.3 million images of 28k persons, respectively. These datasets were specifically compiled to provide a balanced racial and cultural representation, and their participants are divided into two "skin bin" groups of four and eight tones. However, a limitation of these datasets is the inclusion of celebrity images, which tend to contain facial features not representative of the general population [15].

All datasets providing FST values as metadata [3, 13, 21, 33] share an issue: strictly speaking, FST assignments are given in person by a trained practitioner. This concern has been raised by several works which opted to use their own six-tone, light-to-dark scales [3, 17, 21, 25]. The Monk Skin Tone scale extends to ten tones and was specifically designed "to enable practitioners to teach human annotators and test for consistent skin tone annotations across various environment capture conditions." [30] Its creators note that the accompanying dataset "should be used only as educational content ... [not] as a training dataset for a model." [30]

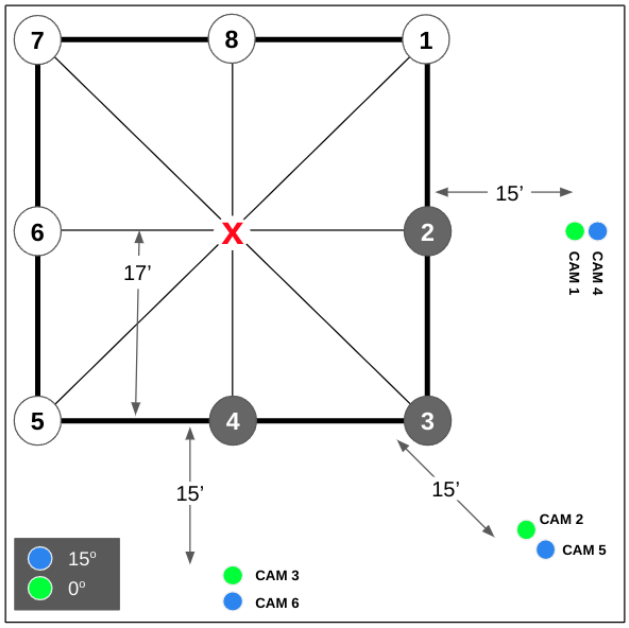
In [14], Howard et al. analyze a self-collected dataset, providing Face Area Lightness Measures (FALMs) derived from in-person ground-truth measurements and image-based estimation. However, their dataset has not been made available to the research community.

3. Data Collection

The CHROMA-FIT dataset comprises indoor and outdoor images taken with a Nikon D7500 DSLR camera equipped with an 18-140mm lens. Fig. 1 shows the setup for both collection environments.



(a) Indoor Setup



(b) Outdoor Setup

Figure 1. Data Collection Setup

Year	Dataset	# People	Content	Labeling Method	Label Type
2015	BUPT-Globalface [33]	38k	2 million images	Manual & Automated	ITA, FST
2015	BUPT-Balancedface [33]	28k	1.3 million images	Manual & Automated	ITA, FST
2017	Pilot Parliaments Benchmark (PPB) [3]	1270	1 image / person	Manual	FST
2018	IARPA Janus Benchmark-C (IJB-C) [21]	3,531	$\sim$6 images / person	Manual	FST
2019	IBM Diversity in Faces (DiF) [23]	n/a	1 million images	Automated	ITA
2022	Casual Conversations [13]	3,011	$\sim$15 videos / person	Manual	FST
2023	Monk Skin Tone Examples (MST-E) [30]	19	1515 images, 31 videos	Manual	MST

Table 1. Summary of Previous Tone-aware Datasets

For each participant, we collected 11 total images: six in the controlled indoor setting, and five in the uncontrolled outdoor setting. In both environments, five images were captured of the participant facing different angles relative to the camera (as seen in Fig. 2). Indoors, an additional close-up enrollment image was captured when the participant directly faced the camera. 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. Indoor Collection

Participants began in the indoor collection environment, shown in Fig. 1a. During the intake stage, various physical measurements were taken: height (in.), weight (lbs.), and skin tone. Skin tone measurements were captured from the forehead and forearm regions using the DSM III Skin Colormeter, which provides values in the $L^*a^*b^*$ color space, and the Pantone CAPSURE, which provides sRGB values.

The indoor images were captured with standard office lighting (i.e., 600 lux). Participants stood at the red X indicated in Fig. 1a, approximately 14 feet from each camera. Each pair of cameras was stacked: on the bottom, a camera with 0° pitch, and on top, a camera with 15° pitch. The indoor images were taken from Camera 3, with participants alternately facing positions 1, 2, 3, 4, and 5.

3.2. Outdoor Collection

After the indoor collection portion, participants were taken to an outdoor setting, visualized in Fig. 1b, and the same five-angle photo set was captured. The outdoor setup is inspired by that of [6]. In this setting, lighting was not controlled, and varied across the two-week collection period. As such, the outdoor images vary in exposure level depending on the sunlight, cloud cover, and other environmental factors of the individual day or time of day.

Participants stood at the red X indicated in Fig. 1b, approximately 32 feet from each camera. As with the indoor environment, each pair of cameras was stacked. The outdoor images were taken from Camera 3, with participants alternately facing positions 2, 3, 4, 5, and 6.

3.3. Metadata

The metadata is composed of self-reported demographic information and ground-truth skin tone information. Skin tone measurements from the forehead (FH) and forearm (FA) regions are provided in the sRGB and $L^*a^*b^*$ colorspaces.

Self-Reported Data:
(1) Age, (2) gender, (3) ethnicity

Measured Data:
(1) FH / FA sRGB, (2) FH / FA $L^*a^*b^*$, (3) FH / FA erythema and melanin, (4) height, (5) weight

4. Skin Tone Measurements

We used the Pantone CAPSURE Colormeter and the DSM III Skin Colormeter, shown in Fig. 3, to measure participant skin tone in distinct color spaces.

When held against a surface, the Pantone CAPSURE Colormeter captures a $9mm^2$ area of skin under an LED light source with a ring-like configuration, emitting various colors onto the skin. The device records an average of four measurements and discerns the nearest shade from a selection of 3000 CMYK colors, outputting a corresponding tone in the sRGB color space. It should be noted that the CAPSURE was not specifically designed for use in measuring skin tone, and occasionally yielded shade matches that made less sense in a skin tone context: e.g., a tone corresponding to a paint shade with a glossy sheen.

Alternatively, the DSM III Skin Colormeter was designed specifically for skin applications. It analyzes a $7mm^2$ section of skin under the illumination of two built-in LEDs, and outputs values in the $L^*a^*b^*$ color space, which is wider than sRGB and thus preferable for expressing the complexity of skin color. Additionally, the device provides measurements for erythema (redness) and melanin (pigmentation) based on skin absorbance characteristics.

Our collection procedure included (1) calibrating each device against its standard white background then (2) taking

Figure 2. Photo Capture Angles

readings from both devices on the forehead and forearm of each participant. We did not require any skin cleansing. Consequently, the measurements may have been influenced by presence of makeup or sunblock.

To augment the original collection, we performed conversions from each device's native color scale to both $L^*a^*b^*$ and sRGB representations (e.g., $L^*a^*b^*$ to sRGB and vice versa). For each participant, we provide eight total measurements: (1) sRGB, (2) sRGB $\rightarrow$ $L^*a^*b^*$, (3) $L^*a^*b^*$, and (4) $L^*a^*b^*$ $\rightarrow$ sRGB from both forehead and forearm regions.

These ground-truth values were next classified according to two skin tone scales. The Apparent Skin Tone (AST) scale, as used in [17], specifies six tone bins, while Google's Monk Skin Tone (MST) scale, presented in [30], specifies ten tone bins.

(a) Pantone CAPSURE (b) DSM III Skin Colormeter

Figure 3. Devices for Measuring Skin Tone

4.1. From Ground-Truth to Apparent Skin Tone

The Apparent Skin Tone (AST) scale has six values, ranging from I (lighter) to VI (darker), in alignment with the Fitzpatrick categories. The scale's name reflects its ability to be retroactively assigned, using ITA calculated from an image or from a measured $L^*a^*b^*$ value. By using the ground-truth $L^*a^*b^*$ measurement, we avoid the challenge of varying illumination in the typical ITA-from-image case.

Our collection procedure yielded four distinct ground-truth $L^*a^*b^*$ values, which were converted to ITA values then mapped to an AST rating, as described in [17]. Eq. (1) shows how to obtain an ITA value from an $L^*a^*b^*$ measurement. Fig. 4 and Tab. 2 visualize the relationship between ITA values and mapped AST tones.

$$\text{ITA} = \left[arctan\left(\frac{(L^* - 50)}{b^*} \right) \right] * \frac{180}{\pi} \qquad (1)$$

Table 2. ITA to AST Rating

Individual Typology Angle (ITA°)	AST Rating	Description
ITA° >55°	I	Very light
41° <ITA° <55°	II	Light
28° <ITA° <41°	III	Intermediate
10° <ITA° <28°	IV	Tan
-30° <ITA° <10°	V	Brown
ITA° <-30°	VI	Dark

This approach treats each AST rating as though it had been determined by an evaluator exclusively focused on a specific skin area. To establish a single ground-truth AST rating, we computed the mode among all the ratings. In

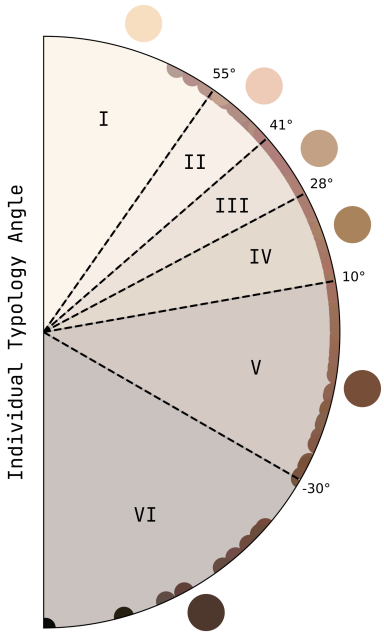


Figure 4. ITA Scale

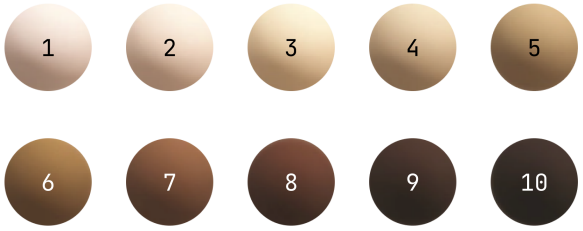
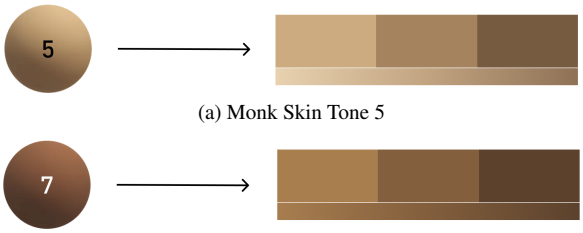


Figure 5. MST Orbs



(a) Monk Skin Tone 5

(b) Monk Skin Tone 7

Figure 6. Salient Color Extraction from MST Orbs

cases where two modes were identified for a subject, we used the median of the modes as the ground-truth rating.

4.2. From Ground-Truth to Monk Skin Tone

The Monk Skin Tone Scale (MST) is a ten-point scale, 1 (lighter) to 10 (darker), designed to express a broader spectrum of skin tones than the Fitzpatrick scale. The MST scale was derived from a synthesis of scientific research in social psychology and categorization, as well as insights from individuals of diverse ethnic backgrounds. Its purpose is to enhance computer vision systems' comprehension of diverse skin tones and to promote fairness in machine learning evaluations. The ten tones are depicted in Fig. 5.

For each measurement collected from both the forehead and forearm, we generated MST ratings using both sRGB and $L^*a^*b^*$ color spaces, resulting in a total of 8 ratings per subject. In the case of ratings derived from the sRGB color space, we first converted the measurement into a linear RGB value. Then, we compared this value with three salient colors represented in RGB, which were determined from the MST orbs. See an example of salient color extraction in Fig. 6. These salient values were obtained through a K-means algorithm applied to each pixel within the MST orb. Finally, we employed the Root Mean Square Error (RMSE) to calculate the difference between the linear RGB value and the cluster centers. The assigned MST rating corresponded to the orb with the lowest RMSE.

For ratings obtained from the $L^*a^*b^*$ color space, we computed the perceptual difference between the measurement and the salient colors represented in $L^*a^*b^*$ format. This perceptual difference was computed using the CIEDE2000 formula. Similar to the sRGB approach, we assigned the MST rating that matched the orb with the lowest cumulative CIEDE2000 value.

This approach treats each MST rating as if it were assessed by an evaluator solely focusing on a specific skin area. To establish the ground-truth MST rating, we determined the mode across all these ratings. In instances where two modes were identified for a particular subject, we calculated the median of these modes and designated it as the ground-truth MST rating. It is important to note that we excluded the Colormeter's readings for the purpose of MST determination for one subject due to apparent calibration issues specific to that individual.

4.3. Relating AST and MST Scales

Fig. 7 displays exemplar facial images categorized into tone-groups for both the AST and MST scales. It is evident that the MST scale encompasses a significantly broader range of perceptual tones, with our study participants predominantly falling within MST tones 3-8. In a general sense, these six MST ratings appear to encompass the entirety of the AST scale. In Fig. 7, where possible, we have attempted to depict the same individuals classified under both scales, offering insights into their interrelation. The distribution of participants by AST and MST rating can be seen in Fig. 8a-8b.

To further visualize the relationship between tone categorization in both scales, we have presented a heatmap

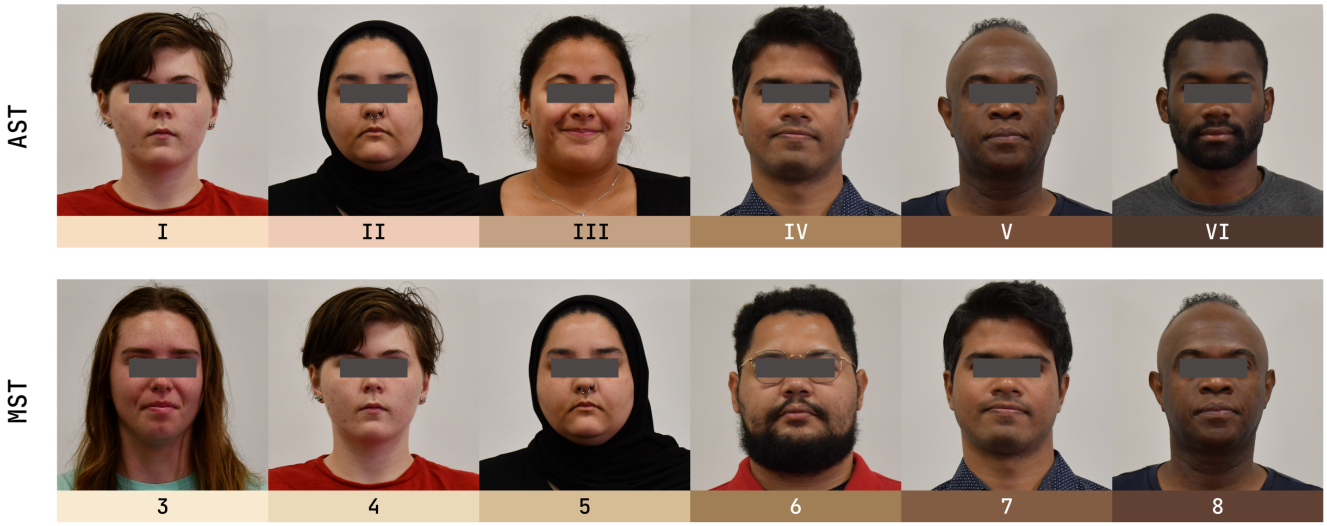
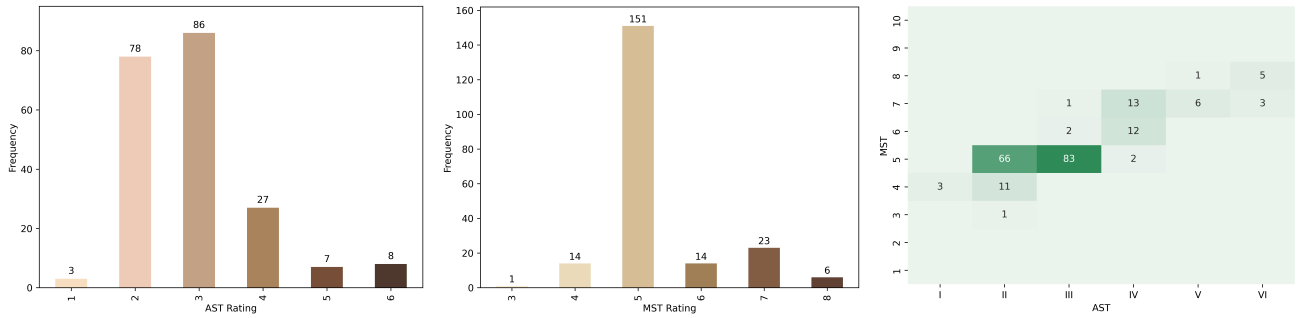


Figure 7. Participants Mapped to AST and MST Ratings



(a) AST Distribution

(b) MST Distribution

(c) AST - MST Heatmap

Figure 8. Skin Tone Ratings

in Fig. 8c. This visualization reveals that the AST scale seem to be encompassed by just six MST tones. This alignment becomes more apparent when referring to the color swatches featured in Fig. 7, as these specific tones exhibit relative similarities.

Interestingly, some people with different AST tones were grouped into a single MST tone: e.g., the exemplar individuals for AST tones II and III were both classified as MST tone 5. While this may seem counterintuitive, an examination of the tonal spectrum displayed in Fig. 6a for MST tone 5 sheds light on how each MST rating captures a relatively wide range of tones.

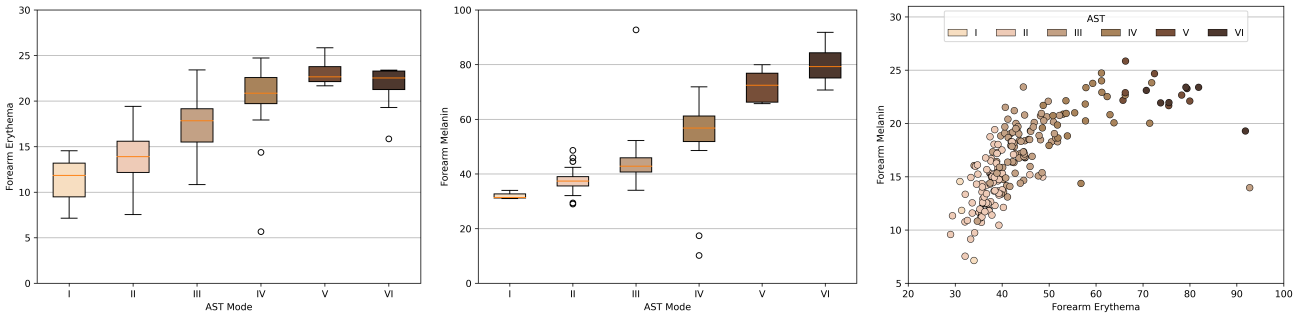
A Pearson's product-moment correlation analysis was conducted to investigate the association between AST and MST ratings within the dataset. There was a robust and statistically significant positive correlation between AST and MST ratings, with a correlation coefficient of $r(209) = 0.8422, p < 0.0005$. This strong positive correlation suggests that as AST ratings increase, MST ratings tend to increase as well, with an AST rating explaining about 70% of the variation in MST ratings.

4.4. Erythema and Melanin

Finally, we consider the erythema (E) and melanin (M) indices given by the DSM III Skin Colormeter. Both values relate to the absorbance characteristics of the skin. Erythema denotes skin redness, while melanin represents skin pigmentation.

The boxplots in Fig. 9a-9b give the distribution of forearm E and M values by AST value respectively. We have chosen to only show forearm values since the distributions for forehead are similar, and we remove the potentially confounding factor of facial cosmetics or sunscreen, which could skew the measurements. The MST data exhibited similar distributions, and as such, they were omitted for clarity.

The scatterplot in Fig. 9c plots melanin vs. erythema values, with each data point colored to indicate the corre-



(a) Forearm Erythema by AST (b) Forearm Melanin by AST (c) Forearm Erythema vs. Melanin by AST

Figure 9. Erythema and Melanin

sponding AST value.

Among the participants in our dataset, erythema and melanin values generally seem to increase in tandem with the AST ratings, indicating that individuals with darker skin tones exhibit greater levels of both erythema and melanin. This association aligns with findings from histological examinations that measure eumelanin content *in vivo* [20], providing a biological basis for the correlation observed in our data.

The bulk of existing research surrounding these characteristics is predominantly focused on dermatological disorders, such as rashes and inflammation, rather than on quantifying the "redness value" of the skin for facial recognition purposes. It is noteworthy that, in the absence of direct melanin and erythema measurements, the L^* component is a valid measure of constitutive pigmentation, while the a^* value has a strong correlation with erythema levels [32]. Thus, these colorspace components can be a proxy for skin tone analysis where the data is not available.

5. Conclusions and Future Work

The CHROMA-FIT dataset aims to provide a valuable resource for the research community in refining the accuracy of both automated and manual skin tone classification techniques. By incorporating ground-truth skin tone data, our dataset facilitates a more nuanced evaluation of prevailing methodologies, enhancing the precision of their assessment.

Looking ahead, we are committed to diversifying our dataset with an increased participant pool that encapsulates a more comprehensive spectrum of skin tones. The present iteration predominantly catalogues mid-range skin tones, specifically those classified as II-III on the Apparent Skin Tone (AST) scale and as type 5 on the Monk Skin Tone (MST) scale. Our goal is to broaden this range, expressly including underrepresented skin tones at both ends of the spectrum—namely, the very light (types 1-2) and the very dark (types 9-10) on the Monk scale—to offer a complete representation of skin tone diversity.

Moreover, our study extends beyond data compilation to include initial experiments that translate physical skin tone measurements into equivalent AST and MST values. We established a correspondence between the two scales, providing a reference for their interrelation. In our future works, we aim to delve deeper into the nuances of the Monk scale, which has not been as extensively explored in skin tone studies. Despite its relative obscurity, we support the original developers' view that the Monk scale holds significant promise for representing a broader spectrum of skin colors more equitably.

In conclusion, our dedication to curating and sharing the CHROMA-FIT dataset is driven by the aspiration to foster equality in technological representation. We hope that our contributions will support and amplify inclusivity, ensuring that individuals of all skin tones are fairly represented in datasets and, by extension, in the technologies that permeate our everyday lives.

Acknowledgment

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21110800003]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

The CHROMA-FIT dataset's compilation owes greatly to Drs. Ioannis Kakadiaris and Shishir Shah (University of Houston), Dr. Olga Korotkova (University of Miami), Dr. David Voelz (New Mexico State University), and Dr. Ram Narayanswamy. Their expertise and support were crucial in the dataset's planning and execution, for which we are immensely grateful.

References

[1] Talking glossary of genomic and genetic terms: Race, Sep 2023. 1

[2] Keivan Bahmani, Richard Plesh, Chinmay Sahu, Mahesh Banavar, and Stephanie Schuckers. Sreds: A dichromatic separation based measure of skin color. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2021. 2

[3] 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. 1, 2, 3

[4] Alain Chardon, Isabelle Cretois, and Colette Hourseau. Skin colour typology and suntanning pathways. *International journal of cosmetic science*, 13(4):191–208, 1991. 1

[5] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun 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*, 1(1):32–41, 2019. 2

[6] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nick Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, Gavin Jager, Matt Larson, Bart Murphy, Christi Johnson, Ian Shelley, Nisha Srinivas, Brandon Stockwell, Leanne Thompson, Matt Yohe, Robert Zhang, Scott Dolvin, Hector J. Santos-Villalobos, and David S. Bolme. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 593–602, 2023. 3

[7] Sergio M. M. de Faria, Jose N. Filipe, Pedro M. M. Pereira, Luis M. N. Tavora, Pedro A. A. Assuncao, Miguel O. Santos, Rui Fonseca-Pinto, Felicidade Santiago, Victoria Dominguez, and Martinha Henrique. Light Field Image Dataset of Skin Lesions. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3905–3908, July 2019. ISSN: 1558-4615. 1

[8] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J Black, and Victoria Abrevaya. Towards racially unbiased skin tone estimation via scene disambiguation. In *European Conference on Computer Vision*, pages 72–90. Springer, 2022. 1

[9] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–26, 2022. 1

[10] 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. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828, Nashville, TN, USA, June 2021. IEEE. 1

[11] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test part 3: Demographic effects, 2019-12-19 2019. 1

[12] H. Han, A. K. Jain, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017. 2

[13] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):324–332, 2021. 1, 2, 3

[14] John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. Reliability and validity of image-based and self-reported skin phenotype metrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4):550–560, 2021. 1, 2

[15] Ashraf Khalil, Soha Glal Ahmed, Asad Masood Khattak, and Nabeel Al-Qirim. Investigating Bias in Facial Analysis Systems: A Systematic Review. *IEEE Access*, 8:130751–130761, 2020. Conference Name: IEEE Access. 2

[16] Brendan Klare, Mark Burge, Joshua Klontz, Richard Vorder Bruegge, and Anil Jain. Face recognition performance: Role of demographic information. *Information Forensics and Security, IEEE Transactions on*, 7:1789–1801, 12 2012. 2

[17] KS Krishnapriya, Gabriella Pangelinan, Michael C King, and Kevin W Bowyer. Analysis of manual and automated skin tone assignments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 429–438, 2022. 2, 4

[18] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pages 365–372. IEEE, 2009. 2

[19] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1547–1557, 2021. 2

[20] P.J. Matts, P.J. Dykes, and R. Marks. The distribution of melanin in skin determined in vivo. *British Journal of Dermatology*, 156(4):620–628, Apr. 2007. 7

[21] 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. 1, 2, 3

[22] Teresa Mendonca, Pedro M. Ferreira, Jorge S. Marques, Andre R. S. Marcal, and Jorge Rozeira. PH² - a dermoscopic image database for research and benchmarking. *Annu Int Conf IEEE Eng Med Biol Soc*, 2013:5437–5440, 2013. 1

[23] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019. 1, 3

[24] Jr. Monk, Ellis P. The Unceasing Significance of Colorism: Skin Tone Stratification in the United States. *Daedalus*, 150(2):76–90, 01 2021. 1

[25] 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. 2

[26] 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. 1

[27] Andre G. C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C. R. Alves Jr, José G. M. Esgario, Alana C. Simora, Pedro B. C. Castro, Felipe B. Rodrigues, Patricia H. L. Frasson, Renato A. Krohling, Helder Knidel, Maria C. S. Santos, Rachel B. do Espírito Santo, Telma L. S. G. Macedo, Tania R. P. Canuto, and Luíz F. S. de Barros. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, Oct. 2020. 1

[28] Gabriella Pangelinan, K. S. Krishnapriya, Vitor Albiero, Grace Bezold, Kai Zhang, Kushal Vangara, Michael C. King, and Kevin W. Bowyer. Exploring causes of demographic variations in face recognition accuracy, 2023. 1

[29] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, 2006. 2

[30] Candice Schumann, Gbolahan O Olanubi, Auriel Wright, Ellis Monk Jr, Courtney Heldreth, and Susanna Ricco. Consensus and subjectivity of skin tone annotation for ml fairness. *arXiv preprint arXiv:2305.09073*, 2023. 2, 3, 4

[31] William Thong, Przemyslaw Joniak, and Alice Xiang. Beyond skin tone: A multidimensional measure of apparent skin color. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4903–4913, 2023. 1, 2

[32] Jennifer K. Wagner, Celina Jovel, Heather L. Norton, Esteban J. Parra, and Mark D. Shriver. Comparing Quantitative Measures of Erythema, Pigmentation and Skin Response using Reflectometry. *Pigment Cell Research*, 15(5):379–384, 2002. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1034/j.1600-0749.2002.02042.x. 7

[33] Mei Wang, Yaobin Zhang, and Weihong Deng. Meta balanced network for fair face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1, 2, 3

[34] David Wen, Saad M Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos De Blas Perez, Alastair K Denniston, Xiaoxuan Liu, and Rubeta N Matin. Characteristics of publicly available skin cancer image datasets: a systematic review. *The Lancet Digital Health*, 4(1):e64–e74, Jan. 2022. 1