World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:11, 2024

Use of Segmentation and Color Adjustment for Skin Tone Classification in Dermatological Images

F. Duarte

Abstract—The work aims to evaluate the use of classical image processing methodologies towards skin tone classification in dermatological images. The skin tone is an important attribute when considering several factor for skin cancer diagnosis. Currently, there is a lack of clear methodologies to classify the skin tone based only on the dermatological image. In this work, a recent released dataset with the label for skin tone was used as reference for the evaluation of classical methodologies for segmentation and adjustment of color space for classification of skin tone in dermatological images. It was noticed that even though the classical methodologies can work fine for segmentation and color adjustment, classifying the skin tone without proper control of the acquisition of the sample images ended being very unreliable.

Keywords—Segmentation, classification, color space, skin tone, Fitzpatrick.

I. Introduction

BIAS in diagnostic systems in the medical field is a current issue that still does not have a standardized mitigation (whether race, gender or socioeconomic status). Systems are trained with data sets that do not include certain demographics [1]. In the dermatological field, there is a direct correlation between skin tone and the risk of skin cancer (melanoma) and diagnosis in darker skin tones is often late and in more advanced stages [2]. To assess bias, it is necessary to have labeled data, and often skin tone is a protected attribute [3]. Due to this difficulty, it is necessary to classify skin tones using the images themselves, to have a basis for a possible assessment of bias in diagnostic algorithms [4]. The work then seeks to make this classification based on the calculation of the ITA (Individual Typology Angle).

II. Methodology

A. Generalized Histogram Tresholding

Generalized Histogram Thresholding (GHT) builds on decades of research in image thresholding techniques. Traditional methods such as Otsu's method and Minimum Error Thresholding (MET) are the foundations of this field, each offering a unique approach to segmenting images and selecting the optimal threshold. Otsu's method minimizes the variance between classes, while MET seeks to minimize the classification error. Weighted percentile thresholding adds another dimension by considering specific percentile values in the histogram. GHT integrates these classical methods by using an approximate maximum a posteriori estimate from a Gaussian mixture model, allowing for smooth transitions between classes and increasing accuracy. GHT also provides the ability to adjust the widths of the histogram bins during thresholding. This method has been shown to be more efficient or on par with advanced segmentation techniques, including deep neural networks. Despite its sophisticated approach, the method can be implemented with minimal modifications to previous methods, making it a practical choice for modern image processing tasks [5].

Fernando Duarte is with Universidade Federal de São Paulo - Instituto de Ciência e Tecnologia, Avenida Cesare Mansueto Giulio Lattes, 1201. São José dos Campos - SP - CEP: 12247-014, Brazil (e-mail:f.duarte@unifesp.br).

B. Individual Topology Angle

The Individual Topology Angle (ITA) is a concept used in the field of color science to quantify skin color. It is defined based on the CIELAB color space, which is a color space specified by the International Commission on Illumination (CIE) that describes all colors to the human eye and was designed to be perceptually uniform.

ITA is calculated using the lightness (L*) and yellow/blue (b*) coordinates of the CIELAB color space [6]. The ITA formula is as follows:

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

C. Fitzpatrick Scale

The Fitzpatrick scale, developed by Thomas B. Fitzpatrick, is a classification system for human skin color. The scale categorizes skin into six different classes based on its response to exposure to ultraviolet radiation, specifically the tendency to burn or tan [7].

D. Datasets

Two datasets were used in this work: HAM10000 ("Human Against Machine with 10000 training images"), which consists of 10015 dermoscopic images with human skin lesions [8]; and the dataset of dermatological images collected in Argentina for the evaluation of Artificial Intelligence tools in its population [9].

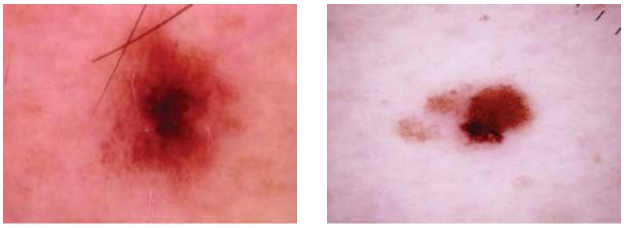
III. Development

The implementation of the ITA calculation and image segmentation using Python. Some libraries were used, such as: OS, DermITA, PIL, Numpy, Math, Matplotlib, and Skimage. For the ITA calculation, both dermITA and another implementation were used. For segmentation, GHT was used. The color space transformations were done using the Skimage.color library, and for the white color correction filter the GrayWorld algorithm library was used. The skin type

classification based on the calculated ITA number and the Fitzpatrick Scale was implemented manually in the Code based on [6]. The complete code then starts scanning the specified directory and loops through all files found in the directory, except for other directories. For each file found, it is opened as an image and checked to see if it is a valid image. Once the image has been validated, it is transformed into a Numpy image, which can be treated with the GrayWorld filter or not, depending on the experiment, and is then transformed into grayscale for segmentation using the Generalized Otsu Threshold. Once the segmentation is complete, the mask is applied to the original image (or the filtered image), and then the ITA calculation is performed. Once the ITA has been calculated, the skin tone classification is performed based on the Fitzpatrick scale. After all the images have been classified, the result is then written to a text file, where the file name, the ITA value, and the skin tone classification are separated by commas for later analysis.

A. Experiment 1

The first experiment performed was to use the DermITA library in the HAM10000 dataset to evaluate whether the results were satisfactory.



(a) ISIC_0024306.jpg – ITA = 38.48 (b) ISIC_0024313.jpg – ITA = 96.90

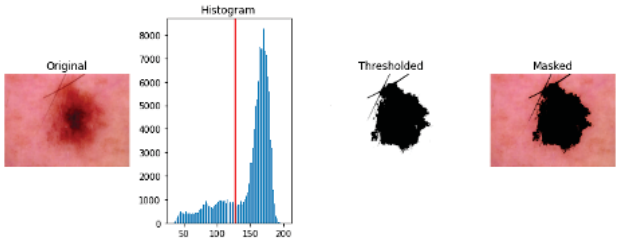
(c) ISIC_0024929.jpg – ITA = -40.85 (d) ISIC_0026735.jpg – ITA = -34.08

Fig. 1 Images of Experiment 1

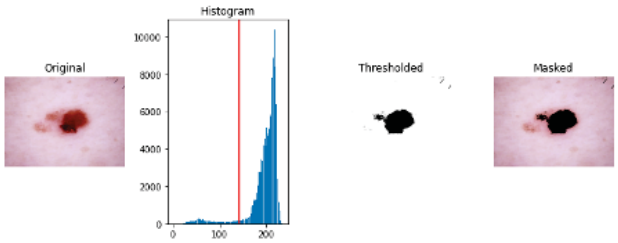
It was then possible to notice that some images with light skin tones had a very low ITA value. In the DermITA calculation, the value is obtained by the median of the three types of fragmentation made in the image, thus showing that the lesion in dark tones would be influencing the ITA calculation.

B. Experiment 2

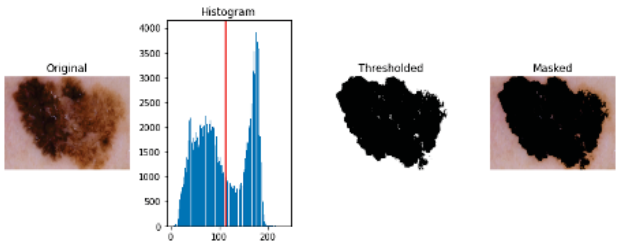
For the second experiment, segmentation was then performed on the images (HAM10000 dataset) without a color filter and then, with the segmentation mask applied, the ITA value was calculated, disregarding the pixels with a null value (where the mask was applied).



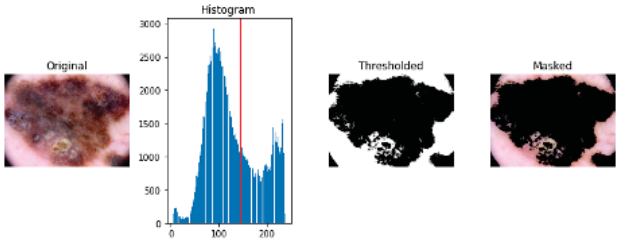
(a) ISIC_0024306.jpg – ITA = 49.36



(b) ISIC_0024313.jpg – ITA = 97.06



(c) ISIC_0024929.jpg – ITA = 83.65



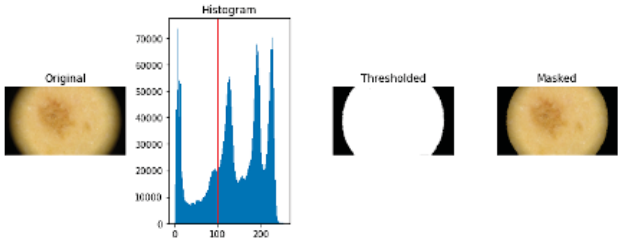
(d) ISIC_0026735.jpg – ITA = 89.67

Fig. 2 Images of Experiment 2

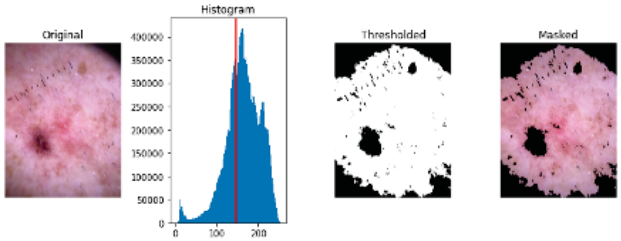
The ITA calculation obtained in the second experiment was on average higher than in experiment 1, showing that with the segmented image, the ITA calculation was performed only on uninjured skin.

C. Experiment 3

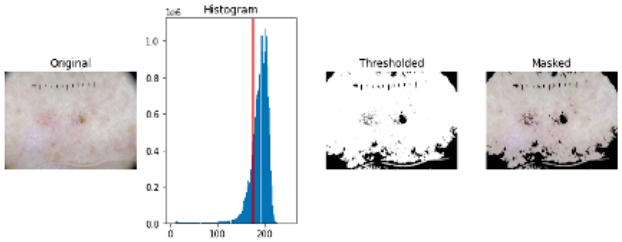
Since the second experiment showed satisfactory ITA values, the third experiment performed the ITA calculation in the same way as in the previous experiment, but using the ISIC Argentina dataset, where the skin tone classification label existed for most of the images. The skin tone classification was performed based on the ITA calculation and the Fitzpatrick scale. Some of the images are shown in Fig. 3.



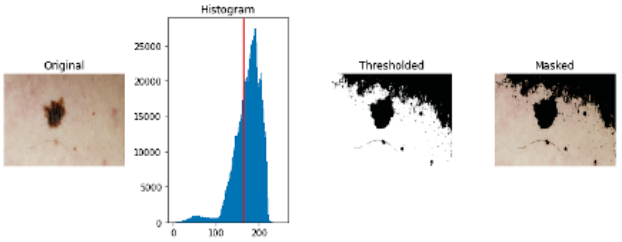
(a) ISIC_0079358.jpg – ITA = 35 – Tipo IV



(b) ISIC_0086914.jpg – ITA = 94 – Tipo I



(c) ISIC_0088904.jpg – ITA = 87 – Tipo I



(d) ISIC_0094098.jpg – ITA = 68 – Tipo I

Fig. 3 Images of Experiment 3

The values obtained in experiment number three were satisfactory in a visual analysis, however, in a quantitative analysis in comparison with the dataset labels, the classification had a very low accuracy rate, around 20%.

As an example, the classification of the dataset is shown compared to the values presented previously:

TABLE I
COMPARISON OF CALCULATED VALUES WITH DATASET LABELS

fitzpatrick_skin_type	isic_id	ITA	calculated fitzpatrick
II	0079358	35	IV
II	0086914	94	I
II	0088904	87	I
II	0094098	68	I

The first observation that can be made is that for all the images shown here as an example, the skin tone classification of the dataset is "II". This proves challenging, since the images are clearly very different in terms of color. The evaluation of experiment 3 then motivated the search for works where some type of filter or color adjustment was made to these types of images in order to achieve some standardization.

D. Experiment 4

In experiment four, an attempt was made to adjust the color of the image before the ITA was calculated. An example that was found in work carried out in this area was the adjustment of the white color, or GrayWorld. So, in this experiment, similarly to experiment 3, the ITA was calculated on the segmented image, but with the color adjusted by this algorithm. In Figs. 4-7 are the values obtained for the same images shown previously:

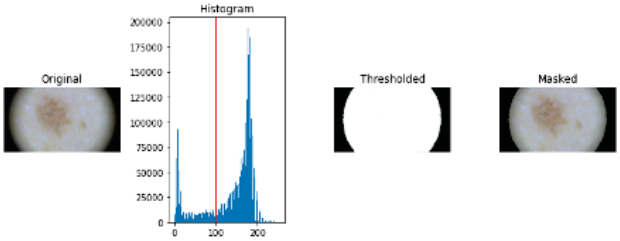


Fig. 4 ISIC_0079358.jpg – ITA = 91 – Tipo I

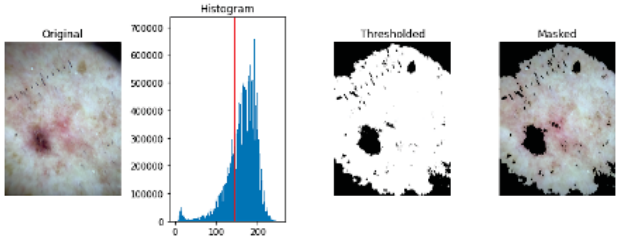


Fig. 5 ISIC_0086914.jpg – ITA = 90 – Tipo I

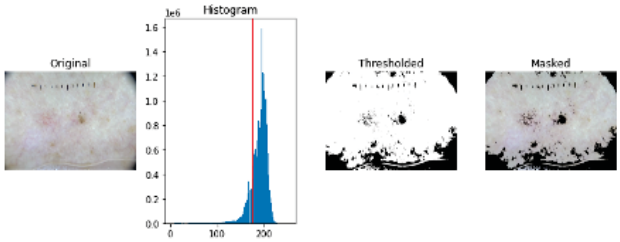


Fig. 6 ISIC_0088904.jpg – ITA = 90 – Tipo I

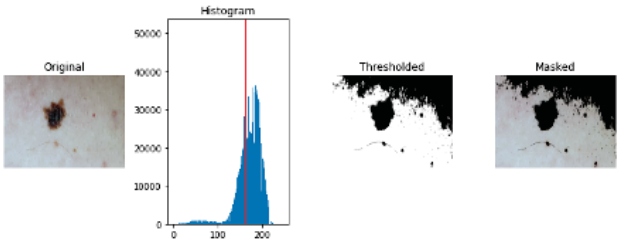


Fig. 7 ISIC_0094098.jpg – ITA = 91 – Tipo I

Experiment 4 shows homogenization both visually and in numerical values. However, the values still appear outside of what was expected when compared to the dataset labels.

E. Other Experiments

Due to the fact that it was not possible to obtain satisfactory results with the methodology used in comparison with the labels of the images from the ISIC Argentina Dataset, some other tests were performed, editing the images or applying GrayWorld + DermITA. In these experiments, it was clear that the segmentation was effective, but the color scheme of the images varied greatly, and the adjustment with GrayWorld did not bring the results to values close to the labels of the dataset.

IV. Results

Below is a summary of the main results of the experiments performed with the two datasets.

Table II shows the results in percentage of the classification obtained in the images of the HAM10000 dataset for different experiment configurations and with the results of [4]. In [4], the value that was used to classify light from dark skin was an ITA of 45. ITA less than 45 for dark skin and ITA greater than 45 for light skin. Therefore, for comparison purposes, the same criterion was used.

TABLE II
Results Classification HAM10000

	GrayWorld	Original	DermITA	Ref. [4]
Light	99.95	92.34	79.78	63.8
Dark	0.05	7.66	20.22	43.2

For the results of the skin type classification experiments using the ISIC Argentina dataset, Table III shows the percentage of hits and misses comparing the values calculated using segmentation without color correction. Here, some divisions in the data are shown, such as gender and image type.

TABLE III
Results Classification ISIC Argentina

	Total	Female	Male	Derm.	Overview	CloseUp
Acerto	19.3	18.3	20.6	19.0	20.6	0
Erro	80.7	81.7	79.4	81.0	79.4	100

V. Conclusions

Segmentation proved effective in removing lesions from images, but differences in image colors make standardization of classification difficult. Another difficult point was finding studies to make a more accurate comparison. Studies published using questionable techniques should be evaluated with caution. The work can be extended to other types of classification and segmentation techniques for possible comparison, in addition to possible different filters to try to homogenize the images in the datasets. But overall it is possible to conclude that it is not feasible to accurately classify skin tones based solely on the image, there is a wide range of variables that affects the image, such as light, distance, lens, sensor, etc. In order to make a fair evaluation of image processing techniques, it would require a more controlled data aquisition set up.

References

[1] Kumar, A., Aelgani, V., Vohra, R. et al. Artificial intelligence bias in medical system designs: a systematic review. Multimed Tools Appl 83, 18005–18057 (2024).
[2] Gupta AK, Bharadwaj M, Mehrotra R. Skin Cancer Concerns in People of Color: Risk Factors and Prevention. Asian Pac J Cancer Prev. 2016 Dec 1;17(12):5257-5264.
[3] Kinyanjui, Newton M., et al. Estimating skin tone and effects on classification performance in dermatology datasets. arXiv preprint arXiv:1910.13268, 2019.
[4] Li X, Cui Z, Wu Y, Gu L, Harada T. Estimating and improving fairness with adversarial learning. arXiv:210304243 [cs]. Published online May 11, 2021.
[5] Barron, J. T. (2020). A generalization of Otsu's method and minimum error thresholding. arXiv preprint arXiv:2007.07350.
[6] YKalb, T., Kushibar, K., Cintas, C., Lekadir, K., Diaz, O., Osuala, R. (2023, August 18). Revisiting Skin Tone Fairness in Dermatological Lesion Classification. arXiv.org. http://arxiv.org/abs/2308.09640
[7] Fitzpatrick, T.B.: The Validity and Practicality of Sun-Reactive Skin Types I Through VI. Archives of Dermatology 124(6), 869–871 (06 1988). https://doi.org/10.1001/archderm.1988.01670060015008
[8] Tschandl, P. (2018, January 1). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Harvard Dataverse. https://doi.org/10.7910/dvn/dbw86t
[9] Lara, M. a. R., Kowalczuk, M. V. R., Eliceche, M. L., Ferraresso, M. G., Luna, D. R., Benitez, S. E., Mazzuoccolo, L. D. (2023). A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population. Scientific Data, 10(1). https://doi.org/10.1038/s41597-023-02630-0