Toward to Reduction of Bias for Gender and Ethnicity from Face Images using Automated Skin Tone Classification

David Molina[1], Leonardo Causa[2], Juan Tapia[3]

Abstract: This paper proposes and analyzes a new approach for reducing the bias in gender caused by skin tone from faces based on transfer learning with fine-tuning. The categorization of the ethnicity was developed based on an objective method instead of a subjective Fitzpatrick scale. A K-means method was used to categorize the color faces using clusters of RGB pixel values. Also, a new database was collected from the internet and will be available upon request. Our method outperforms the state of the art and reduces the gender classification bias using the skin-type categorization. The best results were achieved with VGGNET architecture with 96.71% accuracy and 3.29% error rate.

Keywords: Gender classification, Bias, Skin-Detection.

1 Introduction

Facial recognition is the process of identifying or verifying the identity of a person using their face. It uses biometric to capture, analyze, and compare patterns based on the person's facial details. Traditionally, facial recognition has been associated with the security sector[BG18]. However, it is now expanding across many other applications. Although positive results have been achieved in this field. There are still variables that limit the practical and cross-wise application of this technology, such as the accuracy and the throughput speed.

About the accuracy, some biases have been identified, mainly in the soft biometric features: gender, race and age [BG18, TP19, WW19]. Several researchers have identified these types of biases, but there is not yet evidence about the real causes of such problems [BG18]. Considering the rapid growth in the use of these technologies, the results of facial recognition systems must not be determined by some kind of algorithmic discrimination, i.e., producing better or worse results with certain groups of people, given their gender, race or age. Different studies show that facial recognition systems have higher error rates in people with dark skin, female and young [KBK12]. These shortcomings are due, in part, to biased training processes; which is facilitated by the use of databases that do not take into account the human particularities and differences, especially in aspects such as race and gender [JH14].

Several research groups have worked on facial recognition and the associated biases. Buolamwini and Gebru [BG18] focus on evaluating the performance of three commercial clas-

[1] Universidad Andres Bello, DCI, Avenida Antonio Varas 880, Santiago Chile, d.molinagarrido@uandresbello.edu

[2] TOC Biometrics, R+D Center SR-226, leonardo.causa@toc.cl

[3] Corresponding author: Universidad de Santiago de Chile, Departamento de Informatica, juan.tapia.f@usach.cl

sifiers, IBM [Ibm20], Microsoft [Mic20], and Face++ [Fa20]. Previous research showed discrimination according to race and gender in machine learning algorithms. The analysis of facial databases such as IJB-A [KKT], revealed an over-representation of lighter-skinned, compared to darker-skinned individuals, especially female. To test the classifiers, an annotated database with 1,270 images was generated, The Pilot Parliaments Benchmark (PPB) [BG18]. The images were selected from three African countries and three European countries. They were manually grouped by gender and skin type labeled using the Fitzpatrick scale [Fit8], and the intersection of gender and skin type. Test results showed a relatively high accuracy overall. However, the error rates increase between the different groups. All classifiers provided the best results on the males than females with an error rate between 8.1%-20.6%. Similarly, classifiers showed better performance on lighter-skinned than darker-skinned individuals, with an error rate of 11.8%-19.2%. The best results are for lighter males with 100% accuracy; while the highest error rate was for darker females ranging from 20.8% to 34.7%.

Muthukumar et al. [MPR18] conducted several analyses to try to uncover the reason for the unequal performance of commercial facial recognition services in the gender classification task across intersectional groups defined by skin type and gender. In this study, a modified PPB database [BG18] was used: labels related to skin tone were classified only as light or dark and 1,204 images were used (PPB*). To perform tests on this data set, the IBM Watson classifier and a custom classifier were applied. Both systems showed better results on males than females and the highest error rate was for darker females ranging from 17.0% to 27.0%. The main finding is that the skin type is not the cause of misclassification. Besides, they have shown evidence suggesting differences in the lip, eyes and cheek structure by the ethnicity.

Wu and Wang [WW19] used deep learning method to classify facial features and to study the factors affecting face recognition, mainly the influence of the age and gender. Their results showed an average recognition rate of 83.7% using the CAS-PEAL face database [GCS04] (12,000 images). About the gender, the system performs better on males than females. Considering the age, middle-aged men presented lower performance than that of youth and the elderly; and the female had not a significant difference in the recognition rate. Dhomne et al. used Deep Convolutional Neural Network (D-CNN) algorithm based on a VGGNET architecture [SZ15] to develop a gender classification system. A gender-balanced database consisting of 200 celebrity images was used. Their results achieved 95.0% accuracy in the test dataset. Borza et al. [BDD18] compared two methods to develop an automated skin tone classification system to use in visage applications. The first method used histograms in various color spaces and Principal Component Analysis to generate a feature vector. Afterward, a Support Vector Machine (SVM) and voting schema are used to determine the skin tone. The second method uses Convolutional Neural Networks (CNN) to automatically extract chromatic features. Both methods were trained and tested on publicly available datasets with 9,951 images: Caltech, the Chicago face dataset, Minear-Park, and Brazilian face dataset. The SVM method showed an accuracy of 86.7%, and the CNN approach obtained an accuracy of 91.3%.

The relates work shows that the main problems in gender classification is a non-representative database and manual skin type classification by human experts are critical problems in this field of research. A balanced database in terms of race and gender and automated detection of skin tone could be a powerful tool to reduce biased, subjectivity, standardize criteria, and to improve the gender classification in facial recognition systems.

The goal of this paper is to develop a method for gender classification with racial analysis using automated detection of skin tone based on machine learning and deep learning algorithms. Additionally, we build an annotated gender and skin-type balanced-database to train and test this work. The database will be available to other researchers upon request.

2 Methods

2.1 Images Database

In order to study the bias of gender and ethnicity because of the subjective method used to label the skin-color, a new database was created collected images from the internet. This database is distributed equally in gender, which provides phenotype and geographical differentiation (see Fig. 1).

The database consists of 12,000 facial images of different phenotype groups. Divided into a Set 1 of dark-skinned people (black race) and a Set 2 of Asian and Caucasian people (white race).

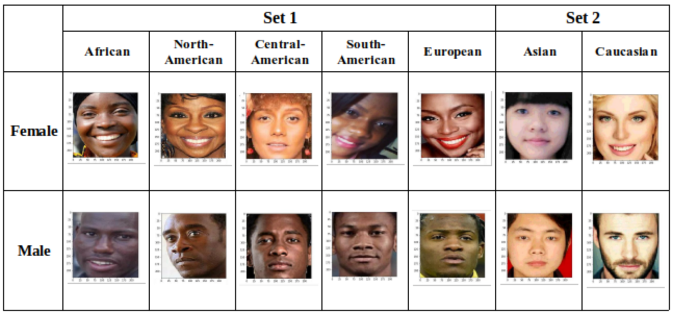

Fig. 1: Example of our annotated gender and skin-type balanced-database. Source: Self-production.

Set 1 is formed by 2,000 images of Africans, 2,000 images of African-Americans (1,000 North-Americans, 500 Central-Americans and 500 South-Americans), and 2,000 images of Europeans. The images were obtained from different existing facial databases and supplemented by Google images.

Set 2 is formed by 3,000 images of Asians and 3,000 images of Caucasians. The images were obtained from UTKFace and SCUT-FBP5500 databases. The gender in both sets is represented by 50% men and 50% female. Images dimensions are at least 250 × 250 pixels, a maximum of five images per subject in different positions, and the wild pose (without restrictions) were used.

2.2 Gender Classification System with Racial Analysis

The proposed classification method consists of two modules. Module 1 applies advanced image processing tools and machine learning to automatically classify the skin tone. It can be described as an analysis cascade of three stages: in Stage 1, the images are processed using CNN algorithms for face detection. Stage 2 uses skin segmentation based in HSV color space thresholds on face images to obtain the face skin [BDD18][ZSQ9]. Stage 3 applies K-means [Mac7] to determine the predominant color in face skin. Module 2 uses features extraction and two classifiers D-CNN, based on VGGNET [SZ15] and MobileNet [HZ17] architectures, to generate the gender classification system.

2.3 Automated Skin Tone Classification

2.3.1 Face Detection

Stage 1 uses the CNN pre-trained algorithm based on TinyFaces detector with a ResNet-101 architecture to identify faces in the images of the database. This algorithm improves the detection of small objects [HR17] and performs well on facial images with different poses and faces of different sizes.

2.3.2 Skin Segmentation

Stage 2 allows to segment the face images generated in the previous step, to obtain only the face skin (Fig. 2). HSV color space thresholds are used for this segmentation, which has been proven to give better results in skin color extraction tasks [BDD18][ZSQ9].

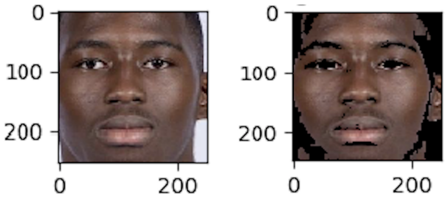

Fig. 2: Segmented image with color thresholds in the HSV color space. Source: Self-production.

2.3.3 Skin Tone Classification

The purpose of Stage 3 is to determine which is the predominant color in the skin-segmented image using the K-means algorithm. A grid search from k=2 up to 10 was used to looking at the best parameters. The best result was achieved with k=4. Each pixel on the segmented image is selected, in the RGB color space, and it is associated in one of the four clusters, depending on the chromatic differences. For each image, the most voting cluster is the predominant color cluster, i.e., it contains the type of color that is most repeated on the face, and therefore, the color with which image is classified. After all, images were associated with some clusters, the mean RGB value in Set 1, Set 2 and both are calculated to determine the thresholds of four color categories. These color categories define the predominant skin tone in the face and represent the racial analysis.

2.4 Gender Classification System

Module 2 uses deep learning to develop gender classifier. Two classification method based on D-CNN were implemented and trained: VGGNET and MobileNet. To train and test, the image database was divided into three different sets: training set, validation set and testing dataset. The output of Module 2 and the skin tone are used to analyze and evaluate the system in terms of gender, skin-type and intersectional groups.

3 Results

3.1 Automated Skin Tone Classification

Tables 1 and 2 shows the distribution of the RGB component (mean and standard deviation) for Set 1 and Set 2 by geographical zones and gender.

Zone	**Set 1 - Male**		**Set 1- Female**	
	Mean RGB	**SD RGB**	**Mean RGB**	**SD RGB**
African	84.25	27.36	90.90	31.68
South-American	96.43	35.03	111.32	40.73
Central-American	98.29	34.47	110.45	41.23
North-American	95.28	27.51	107.81	30.66
European	90.95	29.30	113.14	40.55

Tab. 1: RGB Component for Set 1.

Zone	**Set 2 - Male**		**Set 2- Female**	
	Mean RGB	**SD RGB**	**Mean RGB**	**SD RGB**
Asian	160.28	34.13	172.53	33.44
Caucasian	150.17	37.03	161.61	39.40

Tab. 2: RGB Component for Set 2.

To define the category thresholds, the mean RGB value of each set and the total database was used. Categories 1 and 2 represent dark skin tones. Categories 3 and 4 represent light skin tones. The categories 1, 2, 3 and 4 reached the following skin tone values respectively: <= 97.48, [97.48 - 129.32], [129.32 - 161.15] and >= 161.15.

3.2 Gender Classification System

The system was trained using a person-disjoint database and the parameters were adjusted using three partition sets: Train, Validation and Test. A training set of 60% (7,200) and validation set of 20% (2,400) was used. The performance of the all system was measured using the testing set of 20% (2,400) dataset. For both models, different D-CNN configurations and parameters tuning were applied. The best results were obtained for the models with data augmentation, 150 epochs, inputs image size 224×224 pixels and learning rate of $1e-4$. Some results are shown below.

3.2.1 VGG16 Net Architecture

The overall results for gender classification show a 96.71% accuracy and 3.29% error rate (4.17% Set 1 and 2.42% Set2 -lighter-skinned group). Table 3 shows the error by skin tone category and gender. In Tables 4 and 5 the error classification rate is showed by gender and geographical zone for each data set. In all tests, the error increased for the darker group (Set 1 and category 1) compared with the lighter group (Set 2 and category 4). The highest error rate was for the darker female group.

Category	**Female**		**Male**	
	Amount	**Error [%]**	**Amount**	**Error [%]**
1	22	5.76	15	3.92
2	9	4.52	8	4.02
3	7	3.48	3	1.50
4	10	2.40	5	1.20
Total	48	4.00	31	2.58

Tab. 3: Classification Error rate by Skin Tone Categories and Gender.

Zone	**Set 1 - Female**		**Set 1 - Male**	
	Amount	**Error [%]**	**Amount**	**Error [%]**
African	13	6.50	12	6.00
South-American	2	4.00	1	2.00
Central-American	2	4.00	2	4.00
North-American	6	6.00	2	2.00
European	7	3.50	3	1.50
Total	30	5.00	20	3.33

Tab. 4: Classification Error Rate by Gender and Geographical Zone on Data Set 1. ”Amount” represents the number of images miss-classified.

Zone	**Set 2 - Female**		**Set 2 - Male**	
	Amount	**Error [%]**	**Amount**	**Error [%]**
Asian	10	3.33	6	2.00
Caucasian	8	2.67	5	1.67
Total	18	3.00	11	1.83

Tab. 5: Classification Error rate by Gender and Geographical Zone on Data Set 2.

3.2.2 MobileNet Architecture

The overall results for gender classification show a 96.33% accuracy and a 3.67% error rate (4.17% Set 1 and 3.17% Set 2-lighter-skinned group) but with a small increase in error rate in this set.

In Tables 6 and 7 the error classification rate is showed by gender and geographical zone for each data set. Table 8 shows the error by skin tone category and gender. The error increased for the darker group compared with the lighter group. Unlike the previous case, there is an improvement in female classification, being the highest error rate for males, especially for the darker male group.

Zone	Set 1 - Female		Set 1 - Male	
	Amount	Error [%]	Amount	Error [%]
African	11	5.50	13	6.50
South-American	1	2.00	2	4.00
Central-American	1	2.00	3	6.00
North-American	3	3.00	1	1.00
European	7	3.50	8	4.00
Total	23	3.83	27	4.50

Tab. 6: Classification Error rate by Gender and Geographical Zone on Data Set 1.

Zone	Set 2 - Female		Set 2 - Male	
	Amount	Error [%]	Amount	Error [%]
Asian	4	1.33	15	5.00
Caucasian	8	2.67	11	3.67
Total	12	2.0	26	4.33

Tab. 7: Classification Error rate by Gender and Geographical Zone on Data Set 2.

Category	Female		Male	
	Amount	Error [%]	Amount	Error [%]
1	15	3.93	21	5.48
2	12	6.00	7	3.50
3	5	2.49	8	4.00
4	3	0.72	14	3.36
Total	35	2.92	50	4.17

Tab. 8: Classification Error Rate by Skin Tone Categories and Gender.

3.3 Comparison with Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Gender shades [BG18] is one of most relevant work about gender and ethnicity bias. A manually Fitzpatrick skin-type scale [Fit8] was used to labeling face images (PPB database) in six categories (Types I to VI). Faces labeled were grouped in two skin tone groups, lighter skin (Types I, II and III) and darker skin (Types IV, V, and VI). This classification reached 46.4% for darker skin and 53,6% for lighter skin. Our proposal clustering automatically the same four categories using K-means as reported in [HR17], a lighter skin that includes skin-tone categories 3 and 4, and a darker skin that includes skin-tone categories 1 and 2.

Table 9a presents the distribution of both databases, showing a similar proportion of skin types. The overall accuracy is presented in Table 9b. Our model shows the best results with an improvement ranged from 3.0% up to 8.8%.

In terms of gender (Table 10a) and skin type distribution (Table 10b), our results are better than test reported for the commercial classifiers [BG18]. Gender classification shows an error rate difference of 1.42% compared with the 8.1% for Microsoft, which was obtained

Skin Type	Our Work [%]	Gender Shades [%]
Darker Skin	48.5	46.4
Lighter Skin	51.1	53.6

(a)

Classifier	Accuracy [%]
Our Work	96.7
Microsoft	93.7
Face++	90.0
IBM	87.9

(b)

Tab. 9: a) Database Distribution by Skin Type. b) Overall Accuracy.

the lowest error rate difference between commercial classifiers. Considering skin-type, the error rate was of 2.5%, while the best results for commercial classifiers were obtained by Face++ with an 11.8% error rate. Table 11 shows the results by intersectional groups perform worst on darker females, but our method presents a great improvement by reducing the gap to 3.6%.

Classifier	Female[%]	Male[%]	Error[%]
Our work	96.0	97.4	1.4
Microsoft	89.3	97.4	8.1
Face ++	78.7	99.3	20.6
IBM	79.7	94.4	14.7

(a)

Classifier	Darker [%]	Lighter[%]	Error[%]
Our Work	95.4	97.9	2.5
Microsoft	87.1	99.3	12.2
Face++	83.5	95.3	11.8
IBM	77.6	96.8	19.2

(b)

Tab. 10: a) Accuracy by gender. b) Accuracy by skin type.

Classifier	**DM [%]**	**DF [%]**	**LM [%]**	**LF [%]**	**Gap [%]**
Our Work	95.6	95.2	96.7	98.8	3.6
Microsoft	94.0	79.2	100.0	98.3	20.8
Face++	99.3	65.5	99.2	94.0	33.8
IBM	88.0	65.3	99.7	92.9	34.4

Tab. 11: Overall Accuracy by Gender and Skin Type: Darker Male (DM), Darker Female (DF), Lighter Male (LM) and Lighter Female (LF).

4 Conclusion

In this ongoing research, we show that is feasible to develop an objective method to assign the skin-tone in order to improve the gender classification. This approach improves the results and reduces gender bias by ethnicity produced by the manual assignation of each categorization. This assignation is influenced by the experience of each people. Another achievement of this work is the construction of an annotated gender and skin-type balanced-database, which can be used to train and test this and other methods upon request.

References

[BG18] J. Buolamwini and T. Gebru, “Gender shades: intersectional accuracy disparities in commercial gender classification”,Proceeding of Machine Learning Research, vol. 81, pp. 1—15, 2018.

[TP19] Juan Tapia and Claudio Perez, “Clusters of Features using Complementary Information Applied to Gender Classification From Face Images,” in IEEE Access, vol. 7, pp. 79374-79387, 2019. doi: 10.1109/ACCESS.2019.2923626.

[WW19] S. Wu and D. Wang, "Efect of subjects age and gender on face recognition results," Journal of Visual Communication and Image Representation, vol. 60, pp.116–122, 2019.

[KBK12] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information, " IEEE Transactions on Information Forensics and Security, vol. 7(6), pp. 1789–1801, 2012.

[JH14] A. K. Jain Hu Han, "Age, gender and race estimation from unconstrained face images," MSU Technical Report: MSU-CSE-14-5, 2014.

[Ibm20] IBM, "Watson Visual Recognition - Overview," 2020. [Online]. Available: https://www.ibm.com/cloud/watson-visual-recognition. [Acceded: Feb. 10, 2020].

[Mic20] Microsoft Inc., "Cognitive Services - APIs for AI Developers," 2020. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/. [Acceded: Feb. 10, 2020].

[Fa20] Face++, "Face Detection - Cognitive Services," 2012-2020. [Online]. Available: https://www.faceplusplus.com/face-detection/. [Acceded: Feb. 10, 2020].

[KKT] B. F. Klare, B. Klein, E. Taborsky, A Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained f.ace detection and recognition: Iarpa janus benchmark," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1931–1939, 2015.

[Fit8] T. B. Fitzpatrick, "The Validity and practicality of sun-reactive skin types I through VI," Archives of Dermatology, vol. 124(6), pp. 869, 1988.

[MPR18] V. Muthukumar, T. Pedapati, N. Ratha, P. Sattigeri, C-W. Wu, B. Kingsbury, A. Kumar, S. Thomas, A. Mojsilovic, and K. R. Varshney, "Understanding unequal gender classification accuracy from face images," 2018.

[GCS04] W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," ICT-ISVISION Joint Research & Development Laboratory for Face Recognition, Chinese, 2004.

[SZ15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," ICLR, 2015.

[BDD18] D. Borza, A. Darabant, and R. Danescu, "Automatic skin tone extraction for visagism applications," In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS - Science and Technology Publications, 2018.

[ZSQ9] B. D. Zarit, B. J. Super, and F. Quek, "Comparison of five color models in skin pixel classification," Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV. IEEE Comput. Soc, 1999.

[Mac7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proc. of rhe Fifh Berkeley Symposium on Math., Srar. and Prob., Vol. 1, pp. 281-296, 1967.

[HZ17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017.

[HR17] P. Hu and D. Ramanan, "Finding tiny faces," In 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.