https://doi.org/10.2352/EI.2023.35.5.IRIACV-327

Overcoming Deep Learning Subclass Imbalances: Comparing the Transfer of Identity Across a Racial Transformation

Andrew Sumsion, Shad Torrie, Zheng Sun, and Dah-Jye Lee

Abstract

As facial authentication systems become increasingly advantageous technology, the subtle inaccuracy under specific subclasses grows in importance. As researchers perform data augmentation to increase subclass accuracies, it is critical that the outcomes of data augmentation approaches are understood. We specifically research the impact the data augmentation method of racial transformation has upon the identity of the individual according to a facial authentication network. This demonstrates whether the racial transformation maintains critical aspects of individual identity or whether the data augmentation method creates the equivalence of an entirely new individual for networks to train upon. We utilize a top-performing racial transformation from previous top research articles methods and display the embedding distance distribution of augmented faces compared with the embedding distance of non-augmented faces. We demonstrate that identity is transferred through certain racial transformations while other racial transformations fail to maintain a unique facial outcome. Our results suggest only utilizing certain racial transformations (from current racial transformation results) for data augmentation in order to obtain the highest overall accuracy and accuracies across the subclass of race.

Introduction

As neural networks are used more frequently to solve a wide variety of problems, the problem of overfitting to a dataset is the topic of much research [1]. The broad concept behind overfitting is that the neural network achieves high accuracy on the dataset that is trained on but does not generalize well across data that is not seen in the dataset or subsets of data that are not accurately represented in the dataset. One example of this is seen in facial recognition systems that achieve high overall accuracy, but do not maintain their accuracy across subsets of the supplied dataset: the human race. One of the causes of this low subset accuracy is the dataset being weighted more heavily on one particular race than another.

In solution to this overfitting to certain races caused by an unequal dataset, many have proposed using racial transformation to transform the race of dataset images to obtain a higher accuracy across race [2, 3, 4]. This approach typically uses a Cycle-GAN (Generative Adversarial Network) to transform an image of an individual from one race to another race and back. They use a race classifier as well as a face locator network as discriminators. They propose using their method to augment the training dataset of a facial recognition network to result in having a dataset of equal race to train upon.

Our paper more fully researches this idea and contributes to the field on the ability and inability of using certain racial transformations as positive and negative cases. We do this by researching specifically how much of the human identity itself is transferred during race transformation. This information will provide future researchers with information on how to best augment their datasets for obtaining the highest overall accuracy as well as the highest accuracy across data subsets of race.

Our approach is built around analyzing the results that others have already built and combining them in a method to demonstrate new information. We do this by first performing racial translation based on the work of a top racial translation system that provides pre-trained weights. Then, we take the images and run them through a high-accuracy face recognition network. Finally, we analyze the embeddings through multiple two-sample z-tests that come from the face recognition network to determine to what extent race transformation can be used to improve face recognition models.

Background

The overall accuracy of the face recognition task is continuously improving. One benchmark that is often used is the Labeled Faces in the Wild (LFW) dataset. Currently, the top-performing network reports a 99.833% overall accuracy [5]. This is a remarkable accomplishment, but there still remains improvements that can make the system more robust.

As the overall accuracy of face recognition systems increases, the latest improvements are being made across subsets of data that have lower accuracy. One of these subsets of facial data is the human race. In 2019, the Racial Faces in the Wild (RFW) dataset was published. In their paper, they discuss the inaccuracies across races. Their dataset is comprised of images from the MS-Celeb-1M [6] dataset. They defined four subsets of race: Caucasian, Asian, Indian, and African. They then selected an equal number of images from each of the races and these images comprised their dataset. They then proceed to demonstrate the need for racially balanced datasets in order to obtain equal accuracies across subsets of data. They then used their racially balanced dataset to evaluate four commercial APIs and 4 SOTA (State of the Art) algorithms for facial recognition. Despite all eight of these top algorithms reporting remarkably high accuracies, the accuracy across each race was lower than their reported overall accuracy. The lowest report accuracy for a subset of race was 75.83%. This demonstrates that although top networks report high accuracies, they do not maintain this accuracy across all subsets of data [7].

We also note that the current state-of-the-art algorithm for the LFW dataset that we discussed above cited their results on the RFW dataset. They were able to achieve the lowest racial accuracy of 98.950% accuracy on the Asian subset of the RFW dataset [5]. This is a remarkable improvement, however, in comparison to the overall 99.833%, this means that their percentage error for

this particular subset of data is still over six times as large. This demonstrates the remaining opportunity for improvement that can be supplied by providing a racially balanced dataset.

Race Transformation

We chose to analyze the racial transformation model, VGG1200-Races, based on the race translation accuracy and availability of pre-trained weights [2]. The VGG1200-Races model was developed by a team of researchers from Durham, UK. They note that to overcome the bias in a training dataset you must either perform pre-processing methods, in-processing methods, or post-processing methods. They developed their model to be used in an augmentation method to augment the training dataset to be racially balanced. They did this by performing adversarial image-to-image transfer by using the CycleGAN approach [8]. In the CycleGAN approach, one takes an input, transfers it to another domain, and then transfers it back to the original domain. As the transfers happen, they use discriminators to distinguish how well the transfer took place and that it actually is part of the desired domain.

In VGG1200-Races, they used the initial domain as the input image, a person of a specific race. The domain that they transferred to is an image of the same person but as a different race. After the transfer, they will then transfer the image back into the original domain, the initial race of the individual. They provided an example of two domains: African face images and Caucasian face images. The transfer is switching from one of these races to the other [2]. In order to train the network to transfer race, they used a face locator and a race-identifying network as their discriminators. This allowed the network to be trained to transfer race from one race to another. They trained their network using the VGGFace2 training dataset [9], so in our paper, we use the VGGFace2 evaluation dataset.

In order to utilize the VGGFace2 evaluation dataset, we need race labels for each of the images. We found the ethnicity-recognition-dataset that provides labels for the ethnicity of the VGG-Face2 evaluation dataset. In order to ensure correct labeling of the ethnicity of the dataset, they had three individuals of different ethnicities perform the labeling to avoid the bias from "the well-known other-race effect" [10]. They followed the standard four ethnicities: African American, East Asian, Caucasian Latin, and Asian Indian. These are the same labels provided by RFW [7] and other racially focused datasets only with slightly different labels for the same category. For this paper, we use the terms given in the RFW dataset namely African American as Black, East Asian as Asian, Caucasian Latin as Caucasian, and Asian Indian as Indian.

Using the VGG1200-Races race transformation network, we were able to take the VGGFace2 evaluation dataset and transfer the race of all the images. One of the first steps is the requirement to identify a face. VGG1200-Races used DLIB in their code, so to maintain consistency we also used DLIB [11]. In the ideal case, after racial translation, the dataset will be entirely racially equal as each original image will correspond to an image of every race. However, due to the occasional face that is not identified as a face by DLIB, a few images are lost. We have included tables that indicate the number of images that were created through translation in Table 1 as well as the combined total of original and generated images in Table 2. We display some example images of translated images in Figure 2.

Initial Race	Transferred Race	Number Images Obtained Through Translation
Asian	Black	16776
Asian	Caucasian	16776
Asian	Indian	16776
Black	Asian	9061
Black	Caucasian	9061
Black	Indian	9061
Caucasian	Asian	118353
Caucasian	Black	118353
Caucasian	Indian	118353
Indian	Asian	8428
Indian	Black	8428
Indian	Caucasian	8428

Table 1: Race Translation Table: Results

Race	Original Dataset Count	Transformed Images	Number Images Total
Asian	18751	135842	154593
Black	10699	143557	154256
Caucasian	130942	34265	165207
Indian	9004	144190	153194

Table 2: Race Translation Table: Total Images

Face Recognition

After racial translation, we need to perform face recognition on each of the faces in the dataset, along with the output images from the race translation model. In order to perform this face recognition we used the model, Facenet [12]. This model supplied two different pre-trained models on GitHub [13], one for CASIA-Webface [14] and one for VGGFace2 [9]. As we are using VGGFace2, we used the second supplied face recognition pre-trained model. They reported this model to have an accuracy of 99.65% [15]. In order to determine to what extent identity is transferred through race translation, we passed each of the images, original and generated, through this face recognition model. We then saved the embeddings that were returned in order to analyze the results more fully.

Analysis Cases

After we perform race transformation and facial recognition on all the original images as well as the transformed images, we need to analyze the facial recognition embeddings. This will supply information on how to best use racially transformed images as a data augmentation method in facial recognition training. We initially consider whether the transformed images are considered the same individual by facenet. Next, we consider two approaches to use the racially transformed images: as additional positive cases for the individuals already pictured or to use them as entirely different individuals. We analyze the pros and cons of both of these scenarios.

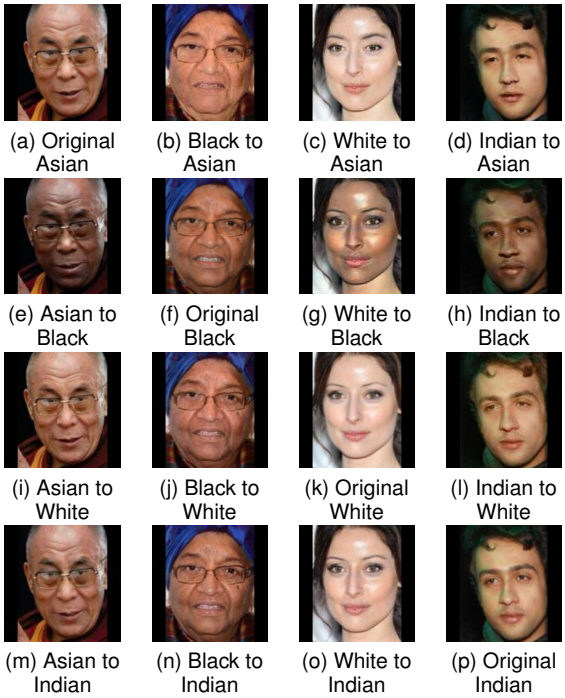


Figure 1: Example Race Translation Images

In order to numerically compare our results, we performed various two sample z tests with the equation for the z-score defined:

$$Z_{score} = \frac{(\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

with μ_1 and μ_2 being the mean of the two populations and σ_1 and σ_2 being the standard deviation of the population. We note that we use σ_1^2 and σ_2^2 (population standard deviation) instead of $s_1^2/\sqrt{n}$ and $s_2^2/\sqrt{n}$ (sample standard deviation) as instead of sampling the validation dataset of VGGFace2 we are performing a census. This means that our desired population of interest is either the entire VGGFace2 validation dataset and also the images that have their races transferred.

Case 1: Identity Similarity of Transferred Images.

The first two sample z test that we run is to determine after racial transformation if all the images from an individual correspond to one person in the racially transformed dimension. We do this by having the first population being the difference of all possible combinations of the racially transformed images from each individual. We have the second population as the difference between the transformed images and all possible negative cases within the original race to the person from which the image was transformed. Upon these definitions we define our null and alternate hypotheses:

$$H_o : \mu_{Transformed} = \mu_{OriginalNegativeCases}$$

$$H_a : \mu_{Transformed} < \mu_{OriginalNegativeCases}$$

with $\mu_{Transformed}$ defined as the difference of all possible combinations of the racially transformed images that came from the same individual. $\mu_{OriginalNegativeCases}$ is defined as the difference between the transformed images and all possible negative combinations of the individual from which the transformed image came from within the same race. The implications of rejecting the null hypothesis would imply that the transformed images do not maintain the same person across the racial transformation. The implications of failing to reject the null hypothesis would provide support to the transformed images maintaining the same person through the transformation.

Our results are seen in Table 3. From these results, we fail to reject the null hypothesis for certain transformations and reject the null hypothesis for other transformations. We fail to reject the null hypothesis for Asian transferred to Black, Asian transferred to Caucasian, Asian transferred to Indian, Black transferred to Asian, and Indian transferred to Caucassian. We reject the null hypothesis for Black transferred to Caucasian, Black transferred to Indian, Caucasian transferred to Asian, Caucasian transferred to Black, Caucasian transferred to Indian, Indian transferred to Asian, and Indian transferred to Caucasian.

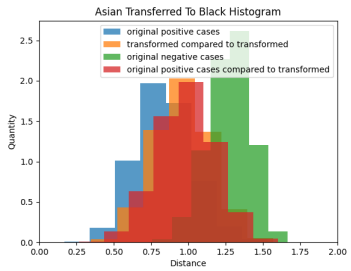
These results imply that for races in which we failed to reject the null hypothesis that the transformed images do not represent a specific individual as there is too much variation in between images. This implies that for the images from the transform that we failed to reject the null hypothesis that one should not use them as data augmentation for facial recognition. This is because the augmented images do not correspond to one another. So, upon performing facial recognition network training these augmentation methods would result in additional noise to the network instead of additional data to learn off of. An example of a failing to reject the null hypothesis is seen as the red and green histograms in Figure 2a. This can be compared to Figure 2b which gives an example of rejecting the null hypothesis in the red and green histograms. This provides a visual understanding of the difference between failing to reject and rejecting the null hypothesis.

Original Race	Transformed Race	Z score	p-value	H_o
Asian	Black	1.202	0.115	FTR
Asian	Caucasian	1.341	0.090	FTR
Asian	Indian	1.278	0.101	FTR
Black	Asian	1.430	0.076	FTR
Black	Caucasian	1.727	0.042	R
Black	Indian	1.815	0.035	R
Caucasian	Asian	2.129	0.017	R
Caucasian	Black	1.754	0.040	R
Caucasian	Indian	1.907	0.028	R
Indian	Asian	1.711	0.044	R
Indian	Black	1.822	0.034	R
Indian	Caucasian	1.595	0.055	FTR

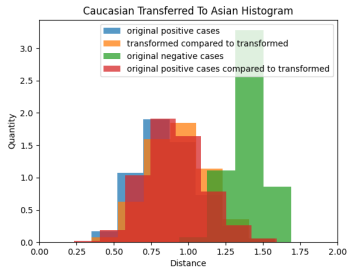
Table 3: Results for Case 1. FTR: Fail to Reject null hypothesis. R: Reject the null hypothesis.

Case 2: Race Transfer as Additional Positive Cases

The second two sample z test that we run compares the result of the difference of the embeddings for various images of a single individual with the difference of the various images of a single



(a) Example of Failing to Reject the Null Hypothesis



(b) Example of Rejecting the Null Hypothesis

Figure 2: Plots demonstrating the population distributions for two of the transformations. For case 1 look at the comparison between the red and the green. For case 2 look at the comparison between the blue and the orange.

individual after we have transferred their race. To test this, our null and alternate hypotheses are defined as:

$$H_o : \mu_{original} = \mu_{transformed}$$

$$H_a : \mu_{original} < \mu_{transformed}$$

with $\mu_{original}$ as the population mean for the difference in embeddings for combinations of original images of the same individual and $\mu_{transformed}$ as the population mean for the difference in embeddings for combinations of transformed images of the same individual.

In this case, the implications of rejecting the null hypothesis would imply supporting evidence to use the race-transformed images as a completely separate individual during training. They could be used as additional negative cases for the original individual from which the augmentation was performed. On the other hand, the implications of failing to reject the null hypothesis imply supporting evidence that race-transformed images should be used as additional positive cases during training. This would supply additional positive cases for the individual on which the data augmentation was performed.

Our results from performing the two sample z-test are seen below in Table 4. With an alpha value of 0.05, we fail to reject the null hypothesis in all possible cases. This supports the use of using racially transformed images as additional positive causes by rejecting the notion that there is a statistically significant difference between $\mu_{original}$ and $\mu_{transformed}$.

Original Race	Transformed Race	Z score	p-value	H_o
Asian	Black	-0.446	0.328	FTR
Asian	Caucasian	-0.479	0.315	FTR
Asian	Indian	-0.380	0.352	FTR
Black	Asian	-0.450	0.326	FTR
Black	Caucasian	-0.314	0.377	FTR
Black	Indian	-0.217	0.414	FTR
Caucasian	Asian	-0.243	0.404	FTR
Caucasian	Black	-0.232	0.408	FTR
Caucasian	Indian	-0.113	0.455	FTR
Indian	Asian	-0.340	0.367	FTR
Indian	Black	-0.317	0.376	FTR
Indian	Caucasian	-0.258	0.398	FTR

Table 4: Results for Case 2. FTR: Fail to Reject the null hypothesis. R: Reject the null hypothesis.

Results

Our results provide evidence for using specific transformations from race transformation as additional positive cases and not using other racially transformed images for training face recognition systems. This can be seen as in order to be used as a negative case they must reject the null hypothesis for case 1 and fail to reject the null hypothesis for case 2. In order to be used as an entirely separate individual for training then the null for case 1 and for case 2 must be rejected (this option never happened in our study). In order to be used as only a negative case, then we must fail to reject case 1 and then reject case 2 (this option also never happened in our study). Under the circumstances where the null hypothesis from cases 1 and 2 were both rejected, we suggest not using these images for training. Our results of what races to use and what cases to not use are given in Table 5.

Original Race	Transformed Race	Case 1 H_o	Case 2 H_o	Outcome
Asian	Black	FTR	FTR	Don't Use
Asian	Caucasian	FTR	FTR	Don't Use
Asian	Indian	FTR	FTR	Don't Use
Black	Asian	FTR	FTR	Don't Use
Black	Caucasian	R	FTR	Positive Cases
Black	Indian	R	FTR	Positive Cases
Caucasian	Asian	R	FTR	Positive Cases
Caucasian	Black	R	FTR	Positive Cases
Caucasian	Indian	R	FTR	Positive Cases
Indian	Asian	R	FTR	Positive Cases
Indian	Black	R	FTR	Positive Cases
Indian	Caucasian	FTR	FTR	Don't Use

Table 5: Results for Case 2. FTR: Fail to Reject the null hypothesis. R: Reject the null hypothesis.

Future Work

This paper contributes to the discussion on how to best augment a dataset using racially transformed images. However, there is much work that remains to be done on this subject. While the work done for racial transformation typically has racial transformation used as additional positive pairs for training [2], we demonstrate that images from certain racial transformations should not be used for training. We demonstrate that certain racial transformations do not transfer all images from one individual into the new racial domain and maintain the requirement that they are the same person after the racial transformation. Further work can be done to train a facial recognition network to demonstrate the improved accuracy that comes from not using certain noisy data augmentation race transformations.

This study was performed on one specific racial transformation network. There are dozens more racial transformation networks that have been developed. Performing this same study, or similar studies will provide crucial information about what racial transformation approaches can be used for what racial transformations. This will provide crucial feedback on these approaches of data augmentation and how to best utilize them. As this research is being performed, the field would also benefit from increased accuracy of racial transformation networks in general.

One of the main drawbacks of not using specific racial transformations is that it will limit the balancing of the dataset, which is often the reason behind utilizing racial transformation networks. Future work will benefit from using a facial recognition network as an additional discriminator. This additional discriminator will require the network to maintain identity across the racial transformation. This way, all the images will be able to be used as a data augmentation method for facial recognition training.

As a novel idea, additional research could benefit from performing race transformation and facial recognition in the same network. The first half of the network could transfer the race of an image while the second half of the network could benefit from having multiple races as it could utilize the novel parts of each race while performing facial recognition. This would provide a novel approach on the approach to both race transformation and facial recognition technologies.

Conclusion

As facial recognition technologies are increasing in use, it is crucial to continue their research and development. In order to maintain a constant improvement of facial recognition technologies, subclasses that have lower accuracies must be considered. One such subclass is the human race. Recent studies have utilized racial transformation networks to augment the dataset to result in a more balanced training dataset. Our work demonstrates that while certain racial transformations should be used to create additional positive cases, other racial transformations should not be used to augment facial recognition training. Our results on what racial transformations to utilize and which to not are displayed in table 5. Our decisions were made by demonstrating the racial transformation of each image in the VGG-Face2 evaluation dataset, performing facial recognition on the original and the transferred race images, and performing two separate two sample z-tests. We also presented possible future work to build upon our results.

References

[1] H. Li, J. Li, X. Guan, B. Liang, Y. Lai, and X. Luo, "Research on overfitting of deep learning," in *2019 15th International Conference on Computational Intelligence and Security (CIS)*. IEEE, 2019, pp. 78–81.

[2] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon, "Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 18–19.

[3] J. Ge, W. Deng, M. Wang, and J. Hu, "Fgan: Fan-shaped gan for racial transformation," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020, pp. 1–7.

[4] Y. H. Kim, S. H. Nam, S. B. Hong, and K. R. Park, "Gragan: Generative adversarial network for image style transfer of gender, race, and age," *Expert Systems with Applications*, vol. 198, p. 116792, 2022.

[5] G. G. Chrysos, S. Moschoglou, G. Bouritsas, J. Deng, Y. Panagakis, and S. Zafeiriou, "Deep polynomial neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4021–4034, 2021.

[6] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[7] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the ieee/cvf international conference on computer vision*, 2019, pp. 692–702.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[9] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[10] A. Greco, G. Percannella, M. Vento, and V. Vigilante, "Benchmarking deep network architectures for ethnicity recognition using a new large face dataset," *Machine Vision and Applications*, 2020.

[11] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[13] T. Esler, "facenet-pytorch," 11 2022. [Online]. Available: https://github.com/timesler/facenet-pytorch

[14] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[15] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

Author Biography

Andrew Sumsion received his B.E. degree in Electrical Engineering with minors in Computer Science and Math from Brigham Young University in 2022. He is currently a Ph.D. student in the Electrical and Computer Engineering Department at Brigham Young University. His research interests include Computer Vision, Deep Learning, and Robotic Vision.

Shad Torrie received his B.E. degree in Computer Engineering with a minor in Computer Science from Brigham Young University in 2022. He is currently pursuing a Ph.D. degree in the Electrical and Computer Engineering Department at Brigham Young University. His research interests include Computer Vision, Deep Learning, Human Computer Interface and Robotics.

Zheng Sun received his Bachelor of Engineering degree from Sun Yat-sen University in 2017. He is pursuing a Ph.D. degree in the Electrical and Computer Engineering Department at Brigham Young University. His work focuses on computer vision and machine learning.

Dr. D. J. Lee received his Ph.D. degree in electrical engineering from Texas Tech University in 1990 and MBA degree from Shenandoah University in 1999. He served in the machine vision industry for eleven years before joining Brigham Young University faculty in 2001. He is currently a professor and the director of the Robotic Vision Laboratory in the Electrical and Computer Engineering Department at BYU. He cofounded Smart Vision Works, Inc. in 2012. His research includes vision systems and devices with artificial intelligence, high-performance visual computing, real-time robotic vision, and visual inspection automation applications.