What the eye doesn't see: Using infrared to improve face recognition of individuals with highly pigmented skin

Alex G. Muthua [a], Rensu P. Theart[a,*], M.J. Booysen[b]

[a] *Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa*
[b] *Faculty of Engineering, Stellenbosch University, South Africa*

Abstract

Face recognition technology has become commonplace in security and access control applications. However, their performance leaves a lot to be desired when working with highly pigmented skin tones. One reason for this is the training bias introduced by under-representation in existing datasets. The other is inherent to pigmentation – darker skins absorb more light and therefore could reflect less discernible detail in the visible spectrum. We show how this can be enhanced by incorporating the infrared spectrum, which electronic sensors can perceive. We augment existing datasets with images of highly pigmented individuals, captured using the visible, infrared and full spectra We fine-tune state-of-the-art face recognition systems and compare the performance of these three spectra. We also assess the impact of narrow and wide cropping, different facial orientations, and sunlight and shaded conditions. We find a marked improvement in the accuracy and in the AUC values of the ROC curves when including the infrared spectrum, with performance increasing from 97.5% to 99% for highly pigmented faces. Including different facial orientations and narrow cropping also improves the performance, and can therefore be deemed as recommended best practices for future research.

Keywords: Face recognition, Highly pigmented skin, Infrared, Convolutional neural network, Biometrics, Receiver operating characteristic (ROC)

Highlights

- Infrared imaging improves the performance of existing face recognition algorithms.
- The improvement increases when both the full spectrum is used.
- Capturing multiple face orientations gives more robust face recognition models.
- Narrow image cropping removes some facial features but improves performance.

1. Introduction

Face recognition technology has become widespread, especially in the fields of security and access control. Although this is not a novel technology per se, the advent of deep convolutional neural networks (CNNs) has improved its performance and effectiveness and led to its adoption in a wide variety of commercial applications. However, despite substantial advances in recent years, it still faces some challenges. One problem that has been under increased scrutiny recently is a degraded performance for certain skin tones (Buolamwini and Gebru, 2018; Wang et al., 2019; Grother et al., 2019; Cook et al., 2019; Krishnapriya et al., 2020). Individuals with highly pigmented skin, are often adversely affected by this poor performance as opposed to their counterparts with lightly pigmented skin. Buolamwini and Gebru (2018) showed that commercially available face recognition software performs worse for individuals with highly pigmented skin (African-American) than for those with lightly pigmented skin (Europeans[1]) individuals: they found a 12.9%

*Corresponding author
Email address: rptheart@sun.ac.za (Rensu P. Theart)

[1] We use the term 'European' to represent groups termed 'Caucasian' in the literature reviewed.

error rate for the former but only 0.7% for the latter. Various reasons can be found for this, such as the choice of algorithm, training dataset and spectrum of light used for the image capture.

The effect of demographic factors on face recognition has been the subject of many years of research (Grother et al., 2019; Drozdowski et al., 2020; Howard et al., 2019; Du et al., 2020; Nagpal et al., 2019; Krishnapriya et al., 2020; Cook et al., 2019). Meissner and Brigham (2001) reviewed research into the well-known phenomenon of "own-race bias": people are better at recognising the faces of people of their own race than of other races. With the advent of neural network algorithms to perform face recognition, further studies were conducted to see whether this bias still existed.

1.1. Impact of data: Training data bias

A common reason given for the poor performance of face recognition for individuals with highly pigmented skin is training bias, or the lack of training data with balanced representation of these individuals.

Algorithms that predate CNN technology that were evaluated in the Face Recognition Vendor Test (Phillips et al., 2010) performed better for the algorithm developers' own demographic. showed better performance for the demographic from which the developers of the algorithm hail . They found, for example, that algorithms developed in European countries and the USA performed better on European individuals, while those developed in Asian countries performed better on Asian individuals. Wang et al. (2019) also found that equitable training reduced the effects of demographic bias, but did not eliminate them.

Similar studies were conducted to assess whether these problems persisted in the algorithms. Although CNNs brought about a marked increase in general accuracy, studies by Buolamwini and Gebru (2018) and Cook et al. (2019) showed that some level of bias still existed. Krishnapriya et al. (2020) found higher false match errors for African-Americans than for Europeans. Furthermore, they found that to achieve an operational false match rate (1 in 10,000), different similarity thresholds would be required for each demographic.

Neural networks, in general, require large amounts of data to produce accurate recognition results. This has the downside of requiring an intensive data collection process when creating or improving systems. Moreover, the nature of the data used also affects the performance of the system, resulting in substantial performance differences between tests done in a lab and those done using real-world data. Thus, the over-representation of individuals with lightly pigmented skin in popular face datasets, such as Labeled Faces in the Wild (LFW) (Huang et al., 2007) and MORPH (University of North CarolinaWilmington, 2019), has an impact on the performance of algorithms when used to recognise individuals with highly pigmented skin. For this reason, recent studies have ensured that individuals with different skin tones are represented evenly in new datasets. These studies serve to both reveal the performance impact caused by unbalanced datasets (Buolamwini and Gebru, 2018; Cavazos et al., 2021; Krishnapriya et al., 2020) and bolster the performance of subsequent algorithms (Wang et al., 2019; Terhörst et al., 2020). A summary of some of the popular, publicly available datasets that are used in research is given in Table 1. For convenience we use the terms 'HPS individual' and 'LPS individual' for the two types of image.

1.2. Impact of light: Dynamic range

The bias in training datasets alone does not explain the poor performance of existing methods when detecting faces of HPS individuals. Physics is also at play: darker surfaces reflect less light than lighter surfaces. The amount of light from the full spectrum of a light source, typically the sun or artificial light, is therefore reflected differently according to the level of skin pigmentation. Specifically, less light energy is reflected, and thus captured by sensors, from faces of HPS individuals than from those of LPS individuals under the same conditions.

One consequence is that the dynamic range (a measure of the difference between the lightest pixel and the darkest pixel in an image) of a HPS individual's face is smaller than that of a LPS individual's face under similar lighting conditions. This in turn limits the ability of algorithms to discern the edges of HPS individuals' facial features while also reducing the amount of information conveyed by the natural shadows on the face. In addition, only part of the reflected light spectrum, the part in the visible spectrum (450 to 700nm), is perceived by the human eye and by cameras that capture the visible spectrum. Outside this band, though, exists a wider electromagnetic spectrum that can be captured by various sensors.

The infrared spectrum, lying just outside the visible light band at 700nm to 2000nm, has been of particular interest in the face recognition field. Li et al. (2007) showed that effects of lighting such as direction,

Table 1: Popular face image datasets

Dataset	Illumination	Spectrum	Landmarks	%HPS	Reference
ColorFERET	Artificial	VIS	No	8	Terhörst et al. (2020)
LFW	Mixed	VIS	No	14	Wang et al. (2019)
MS-Celeb-1M	Mixed	VIS	No	14	Wang et al. (2019)
VGGFace2	Mixed	VIS	No	16	Cao et al. (2018); Parkhi et al. (2015)
IJB-A	Mixed	VIS	No	21	Buolamwini and Gebru (2018)
RFW	Mixed	VIS	No	25	Wang et al. (2019)
PPB	Artificial	VIS	No	46	Buolamwini and Gebru (2018)
MORPH	Artificial	VIS	No	80	Krishnapriya et al. (2020)
CASIA-Face-Africa	Indoor with external sources and outdoor	VIS, NIR	Yes	100	Muhammad et al. (2021)

Notes:
Illumination: type of illumination used in the images captured (natural/ artificial/ mixed).
Spectrum: light spectrum in which images are captured (VIS - visible spectrum, NIR - near-infrared spectrum).
Landmarks: facial features (position of eyes, nose, cheeks, etc.) marked in the images used.
%HPS: portion or percentage of the dataset consisting of images of individuals with highly pigmented skin.

intensity and shadows can change the appearance of a face. Unlike traditional visible light images, infrared images exhibit improved illumination invariance, reducing such effects. Zhang et al. (2008) also noted that infrared images provide better contrast and may contain rich texture details that are absent from visible light images. Fortunately, some cameras can sense beyond the visible spectrum into the infra-red (IR), thermal and ultra-violet ranges. In fact, most CCD (charge-coupled device) and CMOS (complementary metal oxide semiconductor) sensors found in digital cameras can detect rays at the NIR (near-infrared) spectrum range (700nm to 1000nm) (Li et al., 2007). Typically, an infrared cut-off filter is used to block these components in standard cameras.

Researchers have posited that exploiting this extended spectrum could improve face recognition of HPS individuals. One such approach was presented by Boutarfass and Besserer (2018). By removing the infrared cut-off filter from a digital camera, they obtained a "full spectrum" image, containing both visible and NIR light, and found that this improved face recognition, achieving 78% accuracy compared to 56% for visible light alone. They also found that the blue channel visualisation of the visible light image was less clear than that of the full spectrum image. This hints at the conveyance of less information in visible light images.

2. Related work: Recognition of highly pigmented faces

Four recent papers have considered the problem of how to improve face recognition of HPS individuals (Muhammad et al., 2021; Wang et al., 2019; Yang et al., 2021; Terhörst et al., 2020). Their novel methods have had varying degrees of success. We summarise their methods and results in Table 2 and review them below.

Wang et al. (2019) took a two-pronged approach. They created a balanced testing dataset, called RFW (Racial Faces in the Wild), based on the pre-existing LFW (Labeled Faces in the Wild) dataset. With equal proportions of African, Asian, Indian and European individuals, RFW provides a benchmark to test for variations in performance based on skin pigmentation. A second contribution is the use of a deep information maximization adaptation network (IMAN) in their face recognition algorithm. This aims to alleviate the poor performance of HPS recognition by learning facial features that are invariant between HPS and LPS individuals. In this way, representations at group level can be more similarly matched to the global or source distribution.

Yang et al. (2021) and Terhörst et al. (2020) both introduced novel losses to improve face recognition of HPS individuals. Terhörst et al. (2020) introduced a penalisation term in their classifier's loss function. This

Table 2: Research into the mitigation of poor performance of face recognition for HPS individuals

Title	Dataset	Existing method	Novelty	Improvement in HPS vs LPS recognition
Racial Faces in the Wild (Wang et al., 2019)	RFW	SOTA(ArcFace, VGGFace, SphereFace), Commercial API(Face++, Microsoft, Baidu, Amazon)	Introduces balanced test dataset. Domain adaptation used to align global and group distributions	Yes. Overall increase in performance for both HPS and LPS individuals. Difference in accuracy drops from 8% to 3.5%
Comparison-Level Mitigation of Ethnic Bias in Face Recognition (Terhörst et al., 2020)	ColorFERET, LFW	FaceNet	Introduces fairness loss to equalise group performance	Yes. Reduction in MAD of subgroups by up to 52.7%
RamFace: Race Adaptive Margin Based Face Recognition for Racial Bias Mitigation (Yang et al., 2021)	RFW	SOTA(ArcFace, CurricularFace)	Uses a racial bias loss to optimise for different races	Marginal reduction in difference between accuracy of LPS and HPS individuals from 1.57% to 1.15%
CASIA-Face-Africa (Muhammad et al., 2021)	CASIA-Face-Africa	SOTA (ArcFace, lightCNN, SphereFace)	Uses infrared images. Introduces database with large ratio of sub-Saharan HPS individuals	No change in metrics evaluated

Notes:
Dataset: training/testing dataset used.
Existing method: existing algorithm/architecture used to measure the current performance of face recognition
Novelty: new ideas/methods/aspects introduced in the paper to improve performance.
Improvement in HPS vs LPS recognition: improvements gained through the paper's novel method of recognising individuals with highly pigmented and lightly pigmented skin.

forces the performance distributions of different ethnicities (and to some extent, the distributions of skin pigmentation) to be similar and thus ensures equivalent performance for individuals from different groups. Yang et al. (2021) proposed a racial bias loss function that derives different optimal margins for different races during training.

In a paper with similarities to our approach, Muhammad et al. (2021) created a dataset consisting entirely of images of HPS individuals, called CASIA-Face-Africa. This was with the aim of providing a benchmark dataset for the performance of face recognition systems on HPS individuals. It could also act as an augmentative dataset, to increase the number of images of HPS individuals available to researchers and developers of face recognition systems. The dataset also includes infrared images. This makes it possible to analyse the effect of different light spectra on face recognition of HPS individuals. However, the paper does not include the effect of training models on their dataset; rather, it is used solely for testing based on the pre-defined weights on existing models. Further, the dataset does not include the full spectrum images considered here, and the paper does not mention the effect of the light spectrum used to capture the images.

CASIA-Face-Africa does not compare the face recognition performance for LPS and HPS individuals (because it contains only HPS individuals), but the three other papers do so and have found some improvements. Wang et al. (2019) found an improvement in the difference between recognition of LPS and HPS individuals from 8% to 3%. Terhörst et al. (2020) measured the mean absolute deviation (MAD) of the true positive rate for each subgroup evaluated from the mean true positive rate for all subgroups. They achieved a reduction in this MAD value of up to 52.7% with the LFW dataset. Yang et al. (2021) found a drop in the standard deviation between face recognition performance for different races, as well as an overall improvement in performance. They found that state-of-the-art models (ArcFace in this case) achieved a performance of $96.36\% \pm 0.78\%$ on the RFW dataset, while their RamFace model achieved $96.43\% \pm 0.68\%$.

2.1. Contribution

The above review shows that two problems remain to be solved in the field of face recognition of HPS individuals: the data bias in many of the training sets and the resulting algorithms trained with them, and the reduced dynamic range in images of HPS individuals due to light absorption. In this paper we use our own database of 542 HPS individuals, comprising more than 3,000 images, taken in the Cape Town region of South Africa. These images capture three different light spectra (visible, IR, and a combination of the two) and contain a variety of poses or orientations (front-facing, looking right, looking left, looking up and looking down). We use an existing state-of-the-art face recognition algorithm (VGGFace) to evaluate the effect of including IR, either on its own or in combination with the visible spectrum.

Using pre-trained networks and our own dataset, we investigated the effect of including infrared. We further assessed the effect of narrow cropping, various face orientations, and full sun and shaded lighting conditions. We recommend best practices to follow to help improve the performance and robustness of face recognition systems.

3. Method

3.1. Creation of database

All the images of highly pigmented faces captured in our dataset were taken in the Cape Town region of South Africa. This provided a large proportion of HPS individuals. We captured three light spectrum modes for each image: visible light, infrared and full spectrum. To capture all three spectra, we used a Raspberry Pi NoIR camera module, containing a Sony IMX219 8-megapixel sensor, which does not have a built-in IR cut-off filter. We captured visible light images and infrared images using optical filters that can block specific wavelengths of light. We used a SCHOTT BOROFLOAT 33 with a cut-off wavelength of 710nm to capture visible spectrum images and a SCHOTT RG715 filter with a cut-on wavelength of 715nm to capture near-infrared images.

The images were captured outdoors during the day with the sun as the primary light source. Since sunlight contains high intensities of light in the three spectra we needed, we did not need to use external illumination. For each individual in the database, we captured 5 to 7 images in each spectrum mode. We used five different orientations: front-facing, looking left, looking right, looking up and looking down. Additional front-facing images were captured for some individuals, to provide a test set and to evaluate the effect of varying face orientations described in Section 4.3. An example of the images for a single individual is shown in Figures 1, 2 and 3

Figure 1: Visible spectrum image of the same individual.

Figure 2: Infrared spectrum image of the same individual.

Figure 3: Full spectrum image of the same individual.

Face detection is an important preprocessing step for face recognition. CNNs by nature operate on every pixel in an image. Therefore, it is usually ideal to limit the number of irrelevant pixels not containing part of the face. Various face detection techniques have been developed as a result. These include R-CNN, Fast RCNN, Faster RCNN, MTCNN, YOLO, Mask-CNN and many others. For our dataset we used Amazon's AWS Rekognition software to detect the bounding boxes of faces in the captured images. We noted that the default bounding box obtained tended to cut out some facial features, especially the top of the head and the chin. We therefore also created a duplicate set with a wide cropping (increasing the bounding box size by 15%) to evaluate the impact of cropping, as discussed in Section 4.4. The resulting images were then downscaled to a resolution of 224 x 224 pixels, to conform to the structure of VGGFace, which is explained below. The breakdown of the database is shown in Table 3.

Table 3: Database statistics

	Visible	**Infrared**	**Full Spectrum**
Images	3,050	2,753	3,130
Individuals	546	542	548
Test images	289	289	289

3.2. CNN architecture and model

Parkhi et al. performed various face recognition analyses on both the VGGFace (Parkhi et al., 2015) and VGGFace2 (Cao et al., 2018) datasets. As part of this study, they developed a set of what they called "VGG Face Descriptors", which are publicly available. The set consists of a set of CNN models that achieved over 97% accuracy on the Labeled Faces in the Wild dataset. The models are publicly available and have the option of loading a complete set of weights, obtained by Cao et al. (2018) when training on the VGGFace2 dataset. This makes it possible to evaluate other datasets easily through transfer learning and fine-tuning – described below.

A CNN model can be viewed as consisting of two parts: a feature extractor and a classifier. The first is an initial series of convolutional layers that act as feature extractors, meaning their function is to produce a multi-dimensional vector that uniquely represents the image that is provided as input. The second consists of a few layers that convert the multi-dimensional vector into a single output corresponding to one of a pre-specified number of classes that the overall model is set to recognise. Transfer learning is the process by which a model is initialised with pre-trained weights and parameters, followed by freezing the feature-extracting layers such that their weights do not change during training. Thus, only the classification layers are updated to convert these weights into varying sets of classes in different datasets. This makes it possible to use facial feature extractors in their optimised state to identify images as the specified individuals in our database. This can save time and computing power as we can skip the process of training the model to extract features when sample weights are already available.

Fine-tuning follows the same process, but does not freeze the feature-extracting layers. Thus, the features extracting layers, which means they can be updated. Since there are still initialised weights, the extent of the updating is limited by the degree of difference between the recognition task and the database used to obtain the initial weights. In the field of face recognition, this usually means there is little updating unless the new dataset includes variations such as different poses or, in our study, images captured with different light spectra.

By performing transfer learning and freezing the initial feature-extracting layers, we can evaluate the performance of current state-of-the-art systems using different light spectra. Fine-tuning the model weights, and thus retraining the feature-extracting layers, makes it possible to evaluate any improvement gained

(a) Wide cropped image (b) Narrow square cropped image (c) Narrow tight cropped image

Figure 4: Different ways of cropping images

by tuning the model for different light spectra. Since the different light spectra carry different amounts of information, we hypothesised that retraining the feature-extracting layers could further improve performance.

Malli (2021) provides a Tensorflow implementation of the descriptors from Parkhi et al. (2015), along with the pre-trained weights obtained. This provided the basis for our implementation, which we describe below.

3.3. Protocol evaluation

We considered various evaluation protocols to investigate the effect of not only infrared light but also other factors important for the general face recognition field. We first considered a spectral comparison. We trained and tested models using our three light spectra, visible, infrared and full spectrum, and recorded their performance. We also evaluated the effects of wide and narrow cropping.

As mentioned, the face detection algorithm used to detect faces in the captured images has tight boundaries. This leads to features being omitted, especially the ears, top of the head and chin. We therefore created two duplicates of the original set of images. In the first duplicate we increased the face bounding box by 15% so as to include these features. In the second, we made the bounding box square so as to include the ears (since faces are longer than wide, this tends to be the result). To ensure that the face size was the same in all the images we padded the narrow images with zeros, rather than scaling them up, as is generally done. Figure 4 shows these two ways of cropping the image.

The orientation of the face is important for the extraction of features as it alters the appearance of the features. Ideally we would capture all the possible poses that could be expected during real world evaluation or testing. However, the front-facing pose tends to provide the most useful and identifiable features. Although our dataset contained five orientations, during training we evaluated the effect of using only front-facing images. Further, although the training set contains multiple orientations, the test set in this database contains only front-facing images. Therefore, we evaluate the effect on performance when limiting the training set to only front-facing images, as opposed to training on all available orientations.

The illumination of the faces is important for the recognition performance. The lack of an IR cut-off filter can saturate the camera sensor when capturing full spectrum images under direct sunlight. Similarly, indirect sunlight can make the infrared and visible light images darker and they will therefore have a smaller dynamic range. To take these differences into account we evaluated images taken in the shade under indirect lighting and in unshaded, directly sunlit conditions.

3.4. Training

The four hyperparameters we considered when training the models in this study were the optimiser, batch size, learning rate and number of epochs. After performing initial tests on all available optimisers in the Tensorflow framework, we selected the Adam, SGD and AdaGrad optimisers to train the neural networks. These three were the only options to produce reasonably high accuracy in an initial sample dataset of approximately 400 images.

We chose a batch size of 16. Again, during initial tests this produced the best performance when scaling by powers of 2. An exception was the face orientation evaluation. Because limiting the dataset to front-facing images reduces the number of images per class, we further reduced the batch size to 8 to avoid

overgeneralisation (there are one to three front-facing images per individual). When we used the Adam optimiser with fine-tunable weights, we encountered a saddle point that prevented the neural network from being trained effectively. Increasing the batch size to 128 overcame this problem, but there was a resulting drop in performance.

We used a learning rate of 1e-5. Since the weights were initialised with the pre-trained weights and only slight tuning was required, there was no need to alter the learning rate or use a decaying function for it.

Lastly, to determine the number of epochs, we trained each model for over 500 epochs. The points at which the validation loss began to increase while training accuracy stagnated were used to estimate an average number of epochs. In the fine-tuning case, where weights and parameters are fully tunable, we chose 10 epochs. In the non-tunable case, the Adam optimiser required 10 epochs while the SGD and AdaGrad optimisers required 100.

For each evaluation protocol we obtained several models. This was as a result of iterating the following parameters:

- Light spectra: Three light spectrum images are considered: visible (VIS), infrared (IR) and full spectrum (FS).
- Architectures: The VGG Face descriptors provide two architectures for face feature extraction: VGG16 and ResNet50.
- Tunable modes: The pre-defined weights from VGGFace are either non-tunable (NT) or fine-tunable (FT) during training.
- Optimisers: Three optimisers are used: Adam, SGD and AdaGrad.

3.4.1. Training, validation, test dataset split

Once all the images had been captured and compiled as detailed in Section 3.1, a test set was created. To avoid overlap between the training and the test set, we considered only individuals with more than one front-facing image. For every individual in this group, we selected a single random front-facing image to include in the test set. This produced 329, 305 and 336 test images in the visible, infrared and full spectrum images respectively. The common images from these three sets were then drawn out to ensure an even test set size for all scenarios, leaving us with 289 images, as shown in Table 3. The training set comprised the remaining images not selected for the test set.

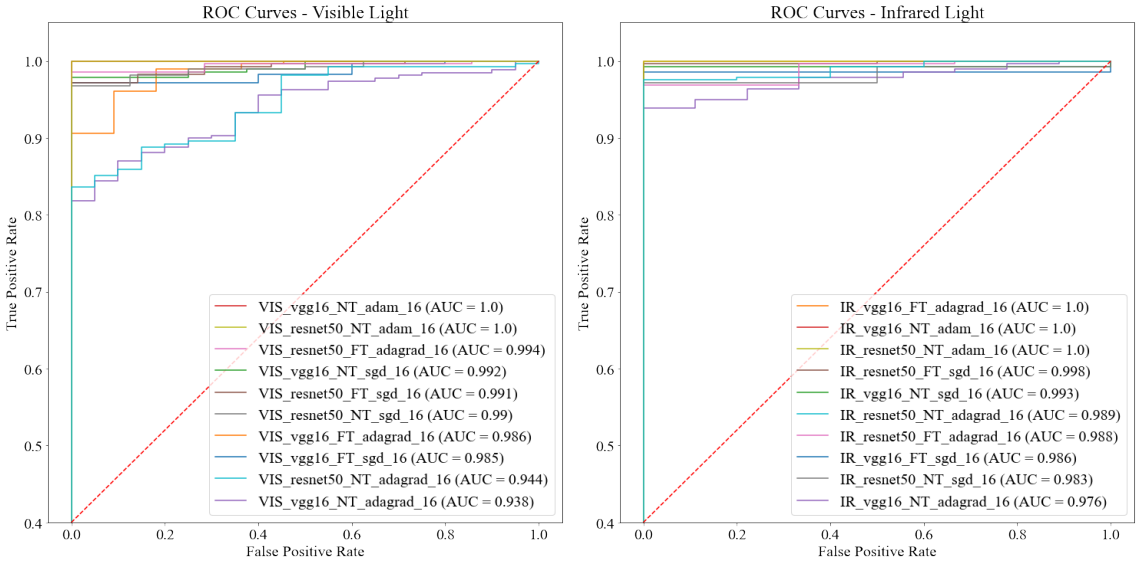
The validation split used during training of the model is 15%. This is drawn from the training set at random, and we made no special considerations. All three sets (training, validation and test) were shuffled before each training run for each model.

4. Results

4.1. Metrics

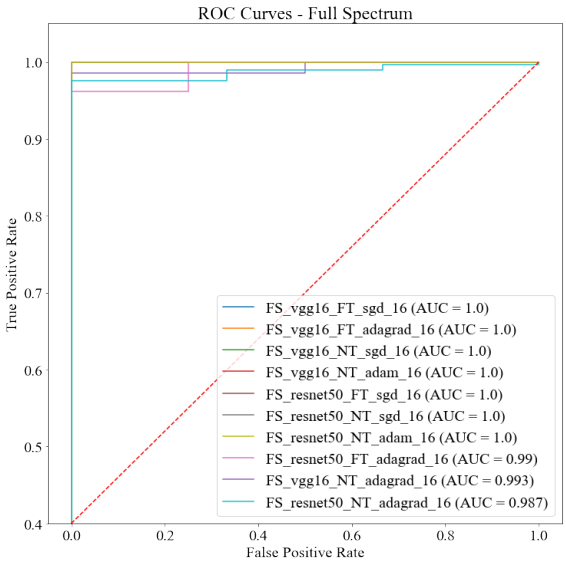
We use two metrics to report the results. The first is the positive identification accuracy, i.e. the percentage of correctly identified test faces. The nature of the training and test datasets means this result is a closed-set identification since all the faces in the test set are contained in the training set as well.

The second metric is the receiver operating characteristic curves (ROC) curves, which show the predictive ability of each model at various prediction or classification thresholds. Specifically, it maps the true positive rate and the false positive rate at certain distinct thresholds. This can be used to set a minimum cut-off on the prediction threshold to avoid models producing false matches in cases where the test image produces a low prediction score. An important metric that can be obtained from these ROC curves is the area-under-curve (AUC) value. A larger AUC value tends to indicate higher true positive rates at low false positive rates, which is ideal for face recognition systems.



(a) Visible spectrum models

(b) Infrared spectrum models



(c) Full spectrum models

Figure 5: ROC curves - spectral comparison

4.2. Spectral comparison

Tables 4, 5 and Figure 5 compare the recognition performance of the three light spectra under consideration. The tables show the positive match accuracy obtained and the figure shows the ROC curves for each spectrum. It is clear that the visible spectrum images produce the poorest performance. As an example, the VGG16, SGD accuracy for the non-tunable case is 98.5% for the visible spectrum images, compared to 99.3% and 99.7% for the infrared and full spectrum images. The difference in thefine-tunable case is even more noticeable, with an increase in accuracy from 97.6% to 99.7% and 99.1%. This trend of slightly lower accuracies and smaller AUC values for the visible spectrum images is consistent across all optimisers and tunable modes. The ROC curves for these images are also less sharp, which corresponds with the smaller AUC values. The difference in the performance of the infrared and full spectrum images is much harder to discern. Their accuracies are quite similar, though a comparison of the ROC curves in Figures 5c and 5b shows slightly larger AUC values for the full spectrum images. These results strongly suggest that infrared and full spectrum images perform better for HPS individuals than visible spectrum images.

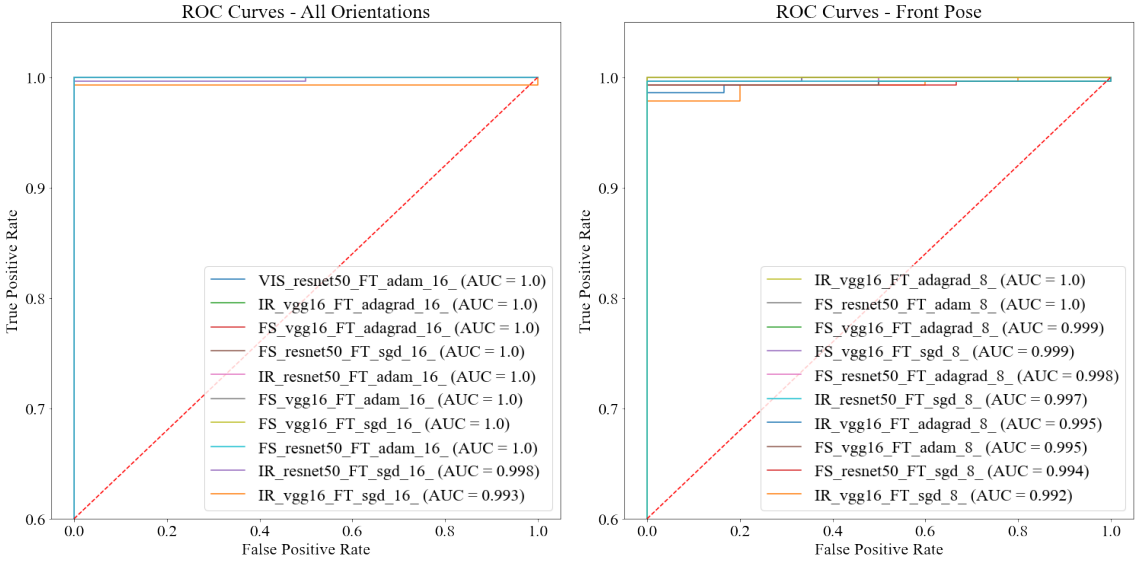
Table 4: Accuracy for a model with non-tunable weights.

Non-tunable (NT) weights						
VGG16						
Optimiser	**Visible**		**Infrared**		**Full Spectrum**	
	Accuracy	*AUC*	*Accuracy*	*AUC*	*Accuracy*	*AUC*
Adam	99.4	1.000	100.0	1.000	99.7	1.000
SGD	98.5	0.992	99.3	0.993	99.7	1.000
AdaGrad	96.0	0.938	98.7	0.976	99.4	0.993
ResNet50						
Optimiser	**Visible**		**Infrared**		**Full Spectrum**	
	Accuracy	*AUC*	*Accuracy*	*AUC*	*Accuracy*	*AUC*
Adam	98.8	1.000	100.0	1.000	99.4	1.000
SGD	98.2	0.990	99.7	0.983	99.4	1.000
AdaGrad	95.1	0.944	97.7	0.989	99.1	0.987

Table 5: Accuracy for a model with fine-tunable weights

Fine-tunable (FT) weights						
VGG16						
Optimiser	**Visible**		**Infrared**		**Full Spectrum**	
	Accuracy	*AUC*	*Accuracy*	*AUC*	*Accuracy*	*AUC*
Adam	0.3	-	0.3	-	0.3	-
SGD	97.6	0.985	99.7	0.986	99.4	1.000
AdaGrad	97.3	0.986	99.7	1.000	99.1	1.000
ResNet50						
Optimiser	**Visible**		**Infrared**		**Full Spectrum**	
	Accuracy	*AUC*	*Accuracy*	*AUC*	*Accuracy*	*AUC*
Adam	0.0	-	0.3	-	0.0	-
SGD	97.9	0.991	99.7	0.998	99.1	1.000
AdaGrad	97.9	0.994	98.4	0.988	99.1	0.990

4.3. Face orientation



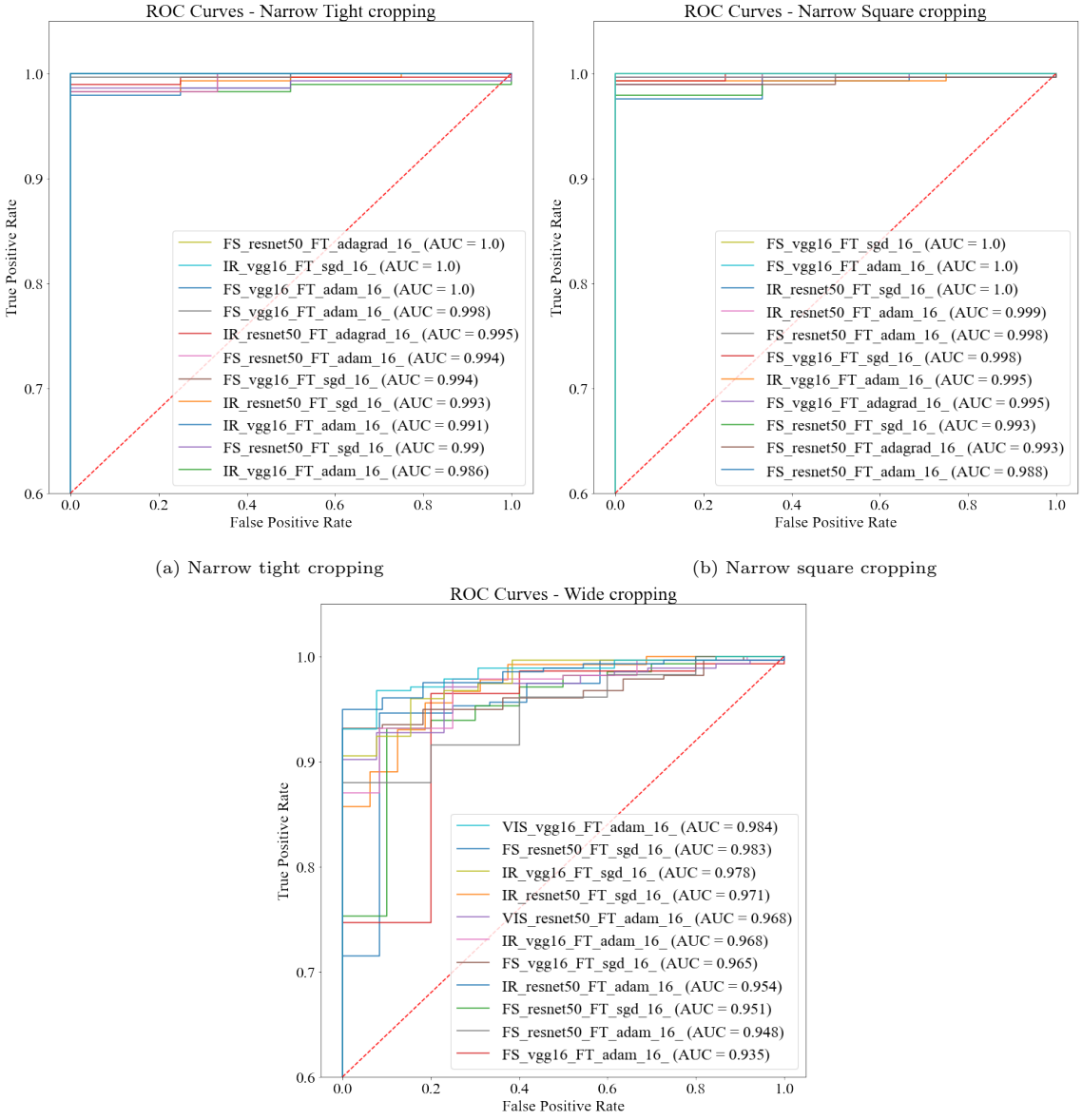
(a) All face orientations

(b) Front-facing orientation only

Figure 6: ROC curves - face orientation comparison

Figures 6a and 6b compare the performance of models trained with various poses and those trained with only the front-facing pose. These figures display the top 10 optimal models for each case. From visual inspection, we find that training using images with all five orientations seems to perform marginally better than training using only front-facing images. Models trained with the five orientations maintain AUC values approximately equal to 1.000 (with only two models missing this mark), while those trained with only the front-facing orientation have AUC values varying from 1.000 down to 0.992. This is despite only testing on front-facing images in both cases. Since all images for a single individual were taken at the same time (a period of five minutes at most), the images are likely to be very similar if only a single pose is used. We suspect that this led to a less robust model during training, which would explain the performance drop. Using several orientations enables the face recognition models to gain more generalised interpretations of faces and thus provides greater robustness. This can be seen further in the performance of the sub-optimal models (outside the top 10 ranked list) shown in Appendix A. The ROC curves for these models show much better performance when trained with several orientations.

4.4. Wide vs narrow cropping



(a) Narrow tight cropping

(b) Narrow square cropping

(c) Wide cropping

Figure 7: ROC curves - wide vs narrow cropping

Figure 7 shows the performance obtained using the wide, tight narrow and square narrow cropping pictured in Figure 4. Narrow cropping, both tight and square, produces the best results. This implies that the neural network architectures are able to draw enough information from the facial features in the narrow crop. The exclusion or inclusion of the ears has no discernible detrimental effect on the overall performance.

The wide cropping includes the missing facial features but also the background elements. Again, since most images were captured at the same location and over a short period of time, most of the background elements (trees, vehicles, benches, etc.) are the same, which may have increased the similarity between images of different individuals, to the point of possibly misclassifying them. The narrow square cropping in an attempt to include the ears also includes some background elements, though to a lesser extent. However, this type of cropping seems not to introduce enough elements to distract from the face, as in the wide cropping case, which explains the lack of a performance drop.

4.5. Shaded vs unshaded

In outdoor conditions the sun provides most of the illumination in the image. Images taken in the shade mostly use indirect sunlight and thus tend to be less saturated or darker, especially in the infrared and full spectrum cases. We therefore split the test set according to the presence or absence of shade. This gave us only 65 images of each kind, because of the 3:1 disparity in the original test set (220 images were unshaded and 70 shaded). Testing on each of the two sets, having trained on all possible images, produced similar results. This shows that even indirect sunlight is still strong enough to provide enough illumination, and specifically infrared light intensity, in the captured images, and that direct sunlight does not saturate the sensor to the point of distorting the image, even in the absence of an IR cut-off filter.

4.6. Best-performing models

Taking into account both the accuracy and the AUC values of the ROC curves that were obtained, we were able to identify a set of the best-performing models, given all possible combinations of the hyperparameters that were considered. These models were trained using narrow cropped images and using all five orientations of the face. we did not take shading into account as the performance difference was indistinguishable. Each of these models exhibited both an accuracy and an AUC value greater than 99.3%:

- Model 1 (*FS_resnet50_NT_adam*): Full spectrum, ResNet50 architecture, non-tunable weights, Adam optimiser.
- Model 2 (*FS_vgg16_FT_sgd*): Full spectrum, VGG16 architecture, fine-tunable weights, SGD optimiser.
- Model 3 (*FS_vgg16_NT_adam*): Full spectrum, VGG16 architecture, non-tunable weights, Adam optimiser.
- Model 4 (*IR_resnet50_NT_adam*): Infrared spectrum, ResNet50 architecture, non-tunable weights, Adam optimiser.
- Model 5 (*FS_resnet50_FT_sgd*): Full spectrum, ResNet50 architecture, fine-tunable weights, SGD optimiser.

On top of the accuracy and the AUC values of the ROC curves, the actual prediction scores obtained are an even better measure of the network's ability to differentiate between individuals. Table 6 shows the prediction scores produced by the five top-performing models listed above.

Table 6: Prediction scores for most accurate models.

Metric	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Accuracy	99.7	99.7	99.7	99.3	99.4
AUC Value	1.000	1.000	0.997	1.000	0.998
Average prediction score	98.3	99.5	98.7	98.3	98.8
True prediction score	98.5	99.6	98.8	98.8	99.2
False prediction score	21.2	51.6	37.6	24.4	43.7
2nd highest prediction score	0.52	0.16	0.37	0.43	0.61

Accuracy (0-100%): *Positive match accuracy obtained by selecting class with the highest prediction score*
AUC Value (0-1): *Area under the model's ROC curve. Has a maximum value of 1, and gives an indication of the true positive rate at low false positive rates*
Average prediction score (0-100%): *The average prediction score produced by the model when classifying test images (includes both true and false matches*
True prediction score (0-100%): *Average prediction score produced when the model correctly classifies a test image*
False prediction score (0-100%): *Average prediction score produced when the model incorrectly classifies a test image*
2nd highest prediction score (0-100%): *Average prediction score of the second highest class when the model classifies a test image (includes both true and false matches)*

4.7. Observations

The evaluation protocols and the list of top-performing models show that infrared and full spectrum images perform best. In addition, the Adam and SGD optimisers outperform AdaGrad. However, the architecture that is used (i.e. Resnet50 or VGG16) seems to have no discernible effect on the performance. An interesting point to note is that all the models that were trained with non-tunable weights in this list used the Adam optimiser. Thus, it can be hypothesised that using fine-tunable weights with the same optimiser could possibly outperform the top models shown here. However, this would probably require training from random initialisation to avoid or overcome the saddle point we encountered.

A second point to note is that the models with fine-tunable weights produced higher prediction scores for incorrectly classified images. This is an undesirable trait as it may suggest a lower open-set identification accuracy, as well as poorer performance for a larger dataset. However, the extent of this problem is beyond the scope of this paper.

The difference between the effects of shaded and unshaded conditions could not be discerned in our study. However, a previous study, Li et al. (2007), showed that the type of illumination used could affect the face recognition system's performance. We found that in outdoor lighting even indirect sunlight is too bright to make a significant difference. Future work should use more controlled lighting conditions to evaluate the effects.

5. Conclusion

This study evaluated the effect of using the infrared spectrum, either on its own or in combination with the visible spectrum (full spectrum), on the performance of face recognition for individuals with highly pigmented skin. The study also evaluated the effects of face orientation, cropping the image and lighting conditions. We used a fine-tuned state-of-the-art network, VGGFace, to perform face recognition.

We found that using infrared light improved performance, both in terms of identification accuracy and reduction of false positives, as exhibited in the ROC curves. Further, using a variety of face orientations produced marginally better performance than using only a single orientation (front-facing), even when the test set contained only the front-facing orientation. Finally, a narrow cropping of the image during face detection showed improved performance. The inclusion of the ears in a narrow square crop is left as an option to researchers. This might reap benefits where similarities in the background are not an issue.

Based on our investigations, we recommend the use of infrared and full spectrum images for face recognition, as these exhibited a marked improvement in performance.

In addition, as best practices, we recommend using several face orientations or poses as well as a narrow cropping of the face, while still including the ears. Although those variations produced minimal differences in the optimal cases, the performance we observed over all the parameters we considered suggests that following this recommendation will provide more robustness.

The degraded performance of face recognition for certain skin tones has been a problem for as long as the technology has existed. Reducing and eventually overcoming this defect will probably depend on the various current and ongoing research projects. Incremental gains such as that we found by using infrared light, and best practices such as those recommended here, are small steps towards making all more visible.

References

Boutarfass, S., Besserer, B., 2018. Using visible+nir information for CNN face recognition, in: 2018 7th European Workshop on Visual Information Processing (EUVIP), pp. 1–6. doi:10.1109/EUVIP.2018.8611681.

Buolamwini, J., Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on Fairness, Accountability, and Transparency. URL: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A., 2018. Vggface2: A dataset for recognising faces across pose and age. arXiv:1710.08092.

Cavazos, J.G., Phillips, P.J., Castillo, C.D., O'Toole, A.J., 2021. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Transactions on Biometrics, Behavior, and Identity Science 3, 101–111. doi:10.1109/TBIOM.2020.3027269.

Cook, C.M., Howard, J.J., Sirotin, Y.B., Tipton, J.L., Vemury, A.R., 2019. 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, 32–41. doi:10.1109/TBIOM.2019.2897801.

Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., Busch, C., 2020. Demographic bias in biometrics: A survey on an emerging challenge. URL: https://arxiv.org/pdf/2003.02488.pdf.

Du, M., Yang, F., Zou, N., Hu, X., 2020. Fairness in deep learning: A computational perspective. URL: https://arxiv.org/pdf/1908.08843.pdf.

Grother, P., Ngan, M., Hanaoka, K., 2019. Face recognition vendor test (FRVT) part 3: Demographic effects doi:10.6028/NIST.IR.8280.

Howard, J.J., Sirotin, Y.B., Vemury, A.R., 2019. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance, in: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8. doi:10.1109/BTAS46853.2019.9186002.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49. University of Massachusetts, Amherst. doi:10.1109/TTS.2020.2974996.

Krishnapriya, K.S., Albiero, V., Vangara, K., King, M.C., Bowyer, K.W., 2020. Issues related to face recognition accuracy varying based on race and skin tone. IEEE Transactions on Technology and Society 1, 8–20.

Li, S.Z., Chu, R., Liao, S., Zhang, L., 2007. Illumination invariant face recognition using near-infrared images. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 627–639. doi:10.1109/TPAMI.2007.1014.

Malli, R.C., 2021. keras-vggface. URL: https://github.com/rcmalli/keras-vggface.

Meissner, C., Brigham, J., 2001. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. Psychology, Public Policy, and Law 7, 3–35. doi:10.1037/1076-8971.7.1.3.

Muhammad, J., Wang, Y., Wang, C., Zhang, K., Sun, Z., 2021. CASIA-Face-Africa: A Large-Scale African Face Image Database. IEEE Transactions on Information Forensics and Security 16, 3634–3646. doi:10.1109/TIFS.2021.3080496.

Nagpal, S., Singh, M., Singh, R., Vatsa, M., 2019. Deep learning for face recognition: Pride or prejudiced? URL: https://arxiv.org/pdf/1904.01219.pdf.

Parkhi, O., Vedaldi, A., Zisserman, A., 2015. Deep face recognition, in: British Machine Vision Conference. URL: https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf.

Phillips, P.J., Scruggs, W.T., O'Toole, A.J., Flynn, P.J., Bowyer, K.W., Schott, C.L., Sharpe, M., 2010. Frvt 2006 and ice 2006 large-scale experimental results. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 831–846. doi:10.1109/TPAMI.2009.59.

Terhörst, P., Tran, M.L., Damer, N., Kirchbuchner, F., Kuijper, A., 2020. Comparison-level mitigation of ethnic bias in face recognition, in: 2020 8th International Workshop on Biometrics and Forensics (IWBF), pp. 1–6. doi:10.1109/IWBF49977.2020.9107956.

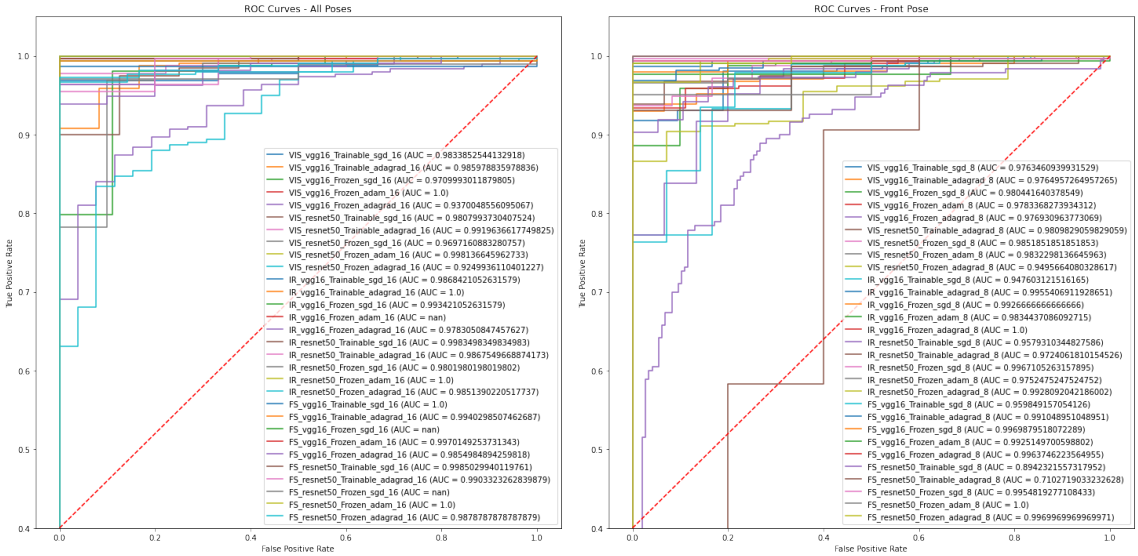
University of North CarolinaWilmington, 2019. Morph academic dataset. URL: https://uncw.edu/oic/tech/morph_academic.html.

Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y., 2019. Racial faces in the wild: Reducing racial bias by information maximization adaptation network, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 692–702. doi:10.1109/ICCV.2019.00078.

Yang, Z., Zhu, X., Jian, C., Liu, W., Shen, L., 2021. Ramface: Race adaptive margin based face recognition for racial bias mitigation, in: 2021 IEEE International Joint Conference on Biometrics (IJCB). doi:10.1109/IJCB52358.2021.9484352.

Zhang, X., Sim, T., Miao, X., 2008. Enhancing photographs with near infra-red images, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. doi:10.1109/CVPR.2008.4587825.

Appendix A. Face orientation comparison

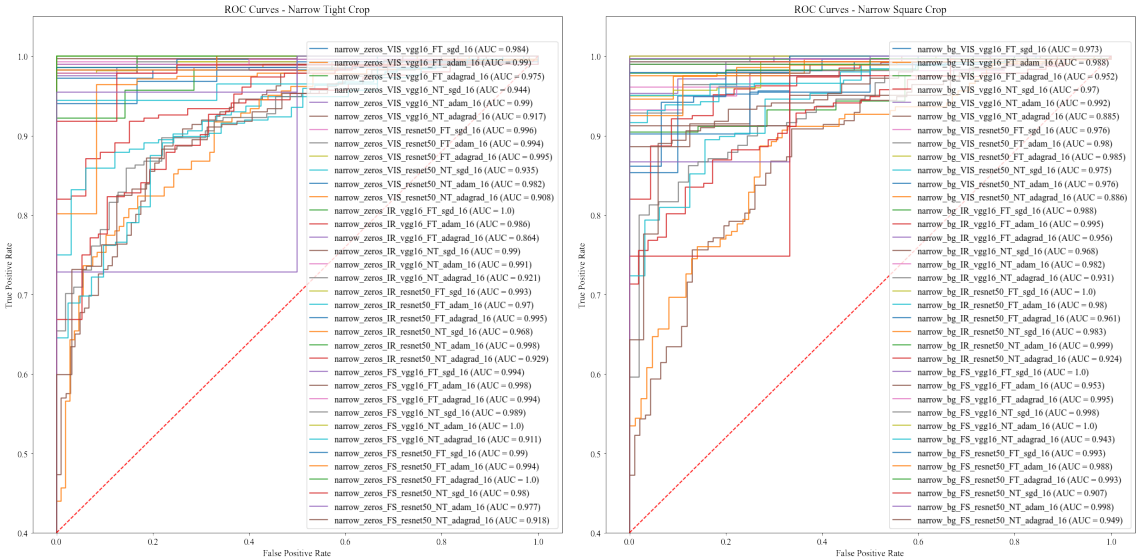


(a) All face orientations

(b) Front-facing orientation only

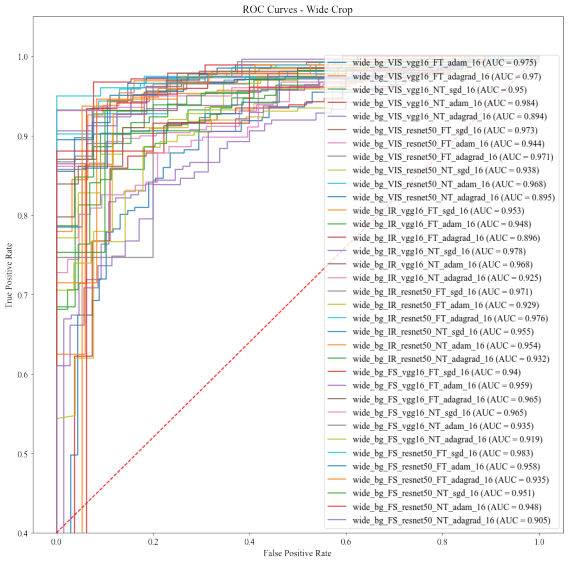
Figure A.8: ROC curves - face orientation comparison

Appendix B. Wide vs narrow cropping



(a) Narrow tight cropping

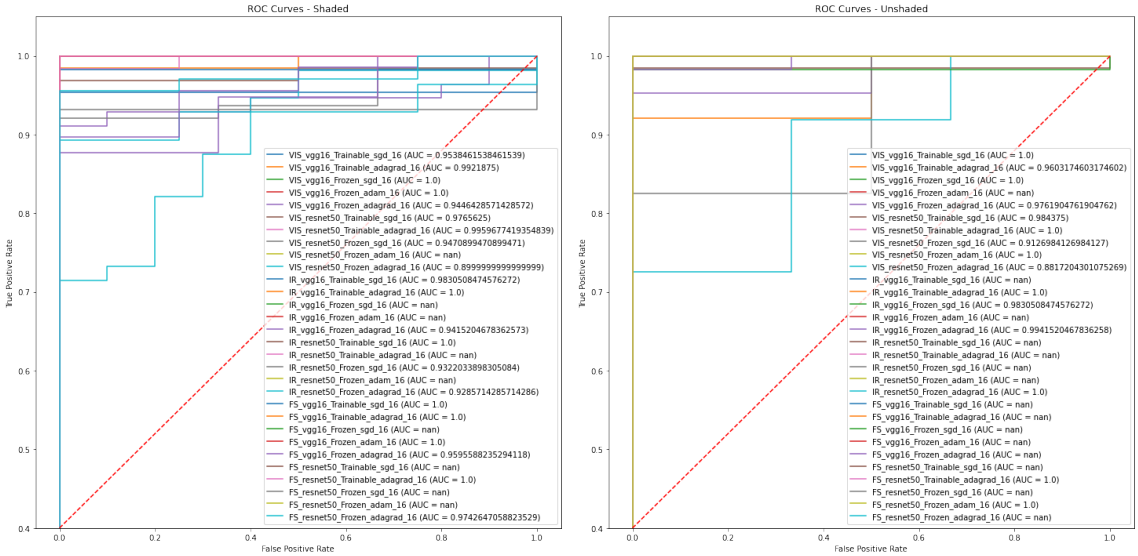
(b) Narrow square cropping



(c) Wide cropping

Figure B.9: ROC curves - wide vs narrow cropping

Appendix C. Shaded vs unshaded



(a) Shaded (b) Unshaded

Figure C.10: ROC curves - shaded vs unshaded