Transformation Tolerance and Demographic Robustness of Machine-based Face Recognition Systems

by

Ashika Verma

B.S. Computer Science and Engineering, Massachusetts Institute of Technology (2022)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author . . . .
Department of Electrical Engineering and Computer Science
August 5, 2022

Certified by . . . .
Pawan Sinha
Professor of Brain & Cognitive Sciences
Thesis Supervisor

Certified by . . . .
Kyle Keane
MIT Quest for Intelligence Research Scientist
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Transformation Tolerance and Demographic Robustness of Machine-based Face Recognition Systems

by

Ashika Verma

Submitted to the Department of Electrical Engineering and Computer Science
on August 5, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Face recognition is widely acknowledged to be a very complex visual task for both humans and computers. Previous studies which analyze robustness of facial recognition systems have revealed that the ability to recognize faces becomes worse as the blur levels of face images increases, and that naturalistic color is important for facial recognition at high blur levels. Additionally, previous studies of current state of the art face recognition technologies have found bias in face recognition amongst different races, resulting in a worse recognition performance for people of color. In this study, we evaluate the performance and robustness of a current state-of-the-art facial recognition neural network architecture (ResNet-101) trained on an augmented facial identity dataset (Augmented Casia Webface) and perform a thorough comparison between White, Black and East Asian identities. We created a full-color, a grayscale and many hue-shifted datasets and then Gaussian blurred each dataset at different intensities and compared how AI systems perform relative to humans and amongst the different races.

Thesis Supervisor: Pawan Sinha
Title: Professor of Brain & Cognitive Sciences

Thesis Supervisor: Kyle Keane
Title: MIT Quest for Intelligence Research Scientist

Acknowledgments

There are many people to thank here at MIT that made these past four years achievable.

I would first like to thank my parents, Ajay and Rashmi Verma, for always being supportive throughout the M.Eng.. Throughout this year, there were times when I wanted to give up but they encouraged me and gave me the space and resources that I need to complete my education.

Second, I am extremely grateful to my supervisor, Dr. Kyle Keane, for all of the support in my research project. He readily accepted me when I needed guidance this year and set me up in a way to succeed. He was also always ready to call and work out any problems I ran into and I appreciate all of the time and energy he invested in me and my success.

Third, I would like to thank another incredible advisor, Anna Musser, for preparing me and teaching me out the research world works. She taught me and gave me the confidence to write my own work and present it confidently, which was crucial for writing up this thesis.

Fourth, I would also like to thank my thesis advisor, Professor Pawan Sinha. Professor Sinha oversaw my research and his support gave me to confidence to work on this project without doubting myself.

I would also like to thank everyone who made this research and education possible these past four years. I'd like to thank my friends for always being there for me and also pushing me through my education. Additionally, there are countless MIT professors, staff, and administrators who made learning a fun experience and made a space for me here at MIT.

MIT is an incredible adventure, and it is the people you meet along the way that truly make it incredible. I am excited to start my next phase of life, and I am proud to everything I learned at MIT with me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

From unlocking personal devices to criminal justice applications, face recognition is a commonly used biometric authentication method in the world today. As such, vulnerabilities within face recognition systems have real world implications. Face recognition systems have been used to establish probable cause for arrests, such as in cases involving identity theft [14], passport fraud [29], and the U.S. Capitol mob where the actions of assailants were filmed and uploaded to YouTube [35]. The implementation of face recognition systems have been particularly successful at identifying suspects involved in driver's license fraud. For example, over 10,000 people within the state of New York were found to illegally possess more than one driver's license using face recognition systems [14]. As such, it can be seen that face recognition, when accurate, can be a useful tool for increasing the reach of our judicial system. Given the ubiquity and power of face recognition systems, it is paramount that such systems be able to perform robustly under real world conditions and also free from certain societal biases before being fully deployed and trusted. Characterizing potential vulnerabilities to these systems, like discrepancies between test sets and training set material, should thus be explored.

This project dived into real world scenarios that may tamper and introduce degradations to data as well as how those degradations affect gender and racial biases that already exist within face recognition. This project is divided into a face recognition problem with the goal of presenting how different scenarios and degradations affect face recognition of those of different races and genders.

1.1 Availability of Face Recognition

Since the introduction of AlexNet in 2012 [21], the number of deep learning methods that recognize faces has exploded, with networks such as ResNet [15], VGGNet [38], and FaceNet [37] serving as the baseline for high accuracy recognition networks [43]. Several companies with substantial investment in AI technologies, such as Google, Meta and Baidu, have declared success on the profoundly important task of face recognition, with Meta developing a face classifier (DeepFace [41]) with an accuracy of 97.35% on the famous benchmark dataset ”Labeled Faces in the Wild (LFW)” [16]. The accuracy of DeepFace has since risen above 99.80% in the span of only three years. These companies have even released simple APIs for this technology to be used by the public, such as Amazon Rekognition[1], Google's Cloud Vision API [3], and Microsoft Face Service [4].

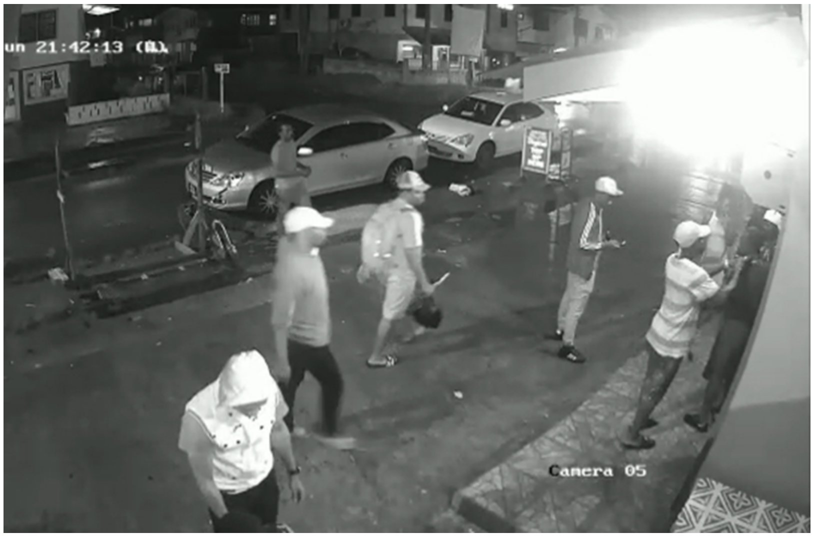

Figure 1-1: Example of low quality surveillance footage.

1.2 Robustness of Face Recognition

Although face recognition systems have now reached incredibly high accuracy rates and are widely available, these systems are not as robust as we might think and really only works in ideal situations. The real world images which these models are used on are not drawn from the same distribution as the training set, which can lead to vulnerabilities in these systems. For example, in surveillance footage, much of the data is of low resolution, taken in unusual lighting conditions, and occasionally in grayscale as shown in Figure 1-1 [13]. This is inconsistent with typical training data fed through networks which is high-quality and restricts pose and lighting parameters. Research indicates that when testing photos originating from an uncontrolled environment, the accuracy of face recognition systems decreases [25], leading to models that are not necessarily robust enough to deploy in the real world.

On the other hand, humans have an incredibly robust ability to recognize faces under strange conditions. We are able to recognize faces under strong degradations, such as identifying celebrities without eyebrows or identifying friends from long distances [36]. By identifying similarities and discrepancies between humans and machines, we stand to gain deeper insights into understanding how humans achieve their remarkable robustness in recognition performance while also potentially improving general knowledge regarding how computational vision systems can be enhanced to exhibit human-like robustness. Studying these neuroscience concepts on neural networks may give us more insight into the incredibly complex task that humans execute naturally: recognizing faces.

1.3 Gender and Racial Bias

In addition to these scenario-introduced degradations such as grainy security footage, there are many steps to machine learning where racial and gender biases can be introduced and ultimately sways the final results of a machine learning model. The first place where these biases can be introduced is within gathering the data for the training dataset. For face recognition, lots of the public datasets are composed of faces scraped from photo datasets like Wikipedia, Facebook, or Flickr. While this seems harmless for the most part, one must also take into account who is using the platforms that the data is being collected from, and whether or not the audience and the photos being uploaded to these platforms are diverse enough to not introduce bias. For example,around 30% of the users of Flickr are from the United States, 60% of the users are male, and 80% of the users are below the age of 55 [2]. With these statistics, one would expect any machine learning model trained on this dataset to be much better at detecting and recognizing a white man in his 30s over a black woman in her 60s. Although these datasets might have demographic imbalances, researchers can easily review the demographic composition of the datasets because they are public and raise awareness by auditing them to call attention to possible biases that may be introduced.

Research has also found that models trained on biased data result in algorithmic discrimination [8]. Within the word embedding space, researchers have found that Word2Vec encodes societal gender biases. In the experiment, researchers trained an analogy generator using Word2Vec which fills in missing words. For example, “man is to computer programmer as woman is to X”. The completed result of this example was “homemaker”, which conforms to the stereotype that men are associated with programming and that women are associated with homemaking [8]. On top of the direct bias visible in the model, this Word2Vec model is widely available and thus commonly used, allowing gender biases to trickle down into many other systems that rely on the model. In addition to word embeddings, research has found that commercial gender classifiers which use images as input significantly favor white men over black women [9]. Some of the key takeaways from Buolamwini et al. were that all of the commercial gender classifiers performed better on male faces over female faces, all the classifiers performed better on light faces over darker faces, and that all the classifiers performed worse on darker female faces [9].

1.4 Project Overview

The ultimate goal for this thesis is to evaluate how a current state of the art face recognition systems perform under certain degradations, to evaluate how different races and genders' recognition performances are affected by those degradations, and to compare the performance of these recognition systems to the human ability to recognize faces. We will accomplish this by creating a dataset with different racial subgroups and an equal number of male and female subjects, creating an algorithm which calculates how well a neural network can recognize and organize faces, evaluating the algorithm over different degradations alongside human results, and comparing and contrasting the results from the different racial and gender groups.

Chapter 2

Background and Related Works

This chapter goes into more depth on how face recognition systems work and how bias can be introduced to face recognition systems. This chapter also discusses how human face recognition can be affected by certain image degradations and how we could link face recognition systems with human perception.

2.1 Face Recognition Systems

Face recognition within the space of artificial intelligence is the ability to confirm or recognize the identity of an individual using a photo of an individual's face. These face recognition systems are used to recognize individuals in photos, videos, or in real time.

Early approaches to face recognition began with taking a face and manually defining landmarks, then using distance metrics across the face to identify an individual. From then until 2012, landmark based recognition dominated the face recognition space with techniques like Eigenface [42], Gabor [26], and LBP [7], reaching accuracies of up to 95% on a popular face recognition benchmark, Labeled Faces in the Wild (LFW) [16]. However, much changed when AlexNet won the ImageNet competition in 2012 by a large margin using a method called deep learning [21]. Deep learning works by using a cascade of multiple layers of processing units to extract different sets of features from a large volume of data. Deep convolutional neural network (CNN) layers automatically learn features from images, many of which were designed for years in previous research especially regarding face recognition, and additionally have more layers which learn higher levels of abstraction. In face recognition, the combination of all of these layers of abstraction finally represent a facial identity with an unprecedented level of stability.

2.1.1 Where is Face Recognition Used?

Face recognition is used worldwide, from biometric authentication on phones to biometric border checks in Europe. As companies got a hold of more computational power and deep neural networks became popular, face recognition became a standard feature in modern technology. In 2014, Facebook publicly released DeepFace, which was a photo-tagging software embedded in their website [41]. The same year, Chicago, for the first time, arrested a man based on face recognition technology which was acquired via a 5.4 million dollar federal grant [11]. Face recognition started to trickle in as a security feature for personal devices in 2015, with its introduction in Windows Hello and Android's Trusted Face, then later in 2017 with Apple's introduction of Face ID [5]. Since then, face recognition use has exploded, with China rapidly increasing its usage on its citizens, retailers experimenting with the technology to track shoplifters, and even Taylor Swift's security team using the technology to identify stalkers [5]. Given the ubiquity of these systems today, it is incredibly important that these face recognition systems are robust and able to be used in a growing and unique set of scenarios.

2.1.2 How does Face Recognition Work?

To perform face recognition, there are three key components: face detection and adjustment, face extraction, and face classification as shown in Figure 2-1. When presented with an image, the first step is to find the face within the image with face detection software, which has been created previously using deep learning methods. Next, with a face landmark detector, the face in the image is aligned using the landmarks and cropped. Afterwards, a face recognition module, or a deep convolutional network, is trained and tested with these aligned face images. Using the trained face recognition module, one performs feature extraction where test images are passed through the network and given a deep feature representation. The last step is face classification, where one can calculate the distance between face representations to either determine if two images have the same identity, or to identify an individual in the image.

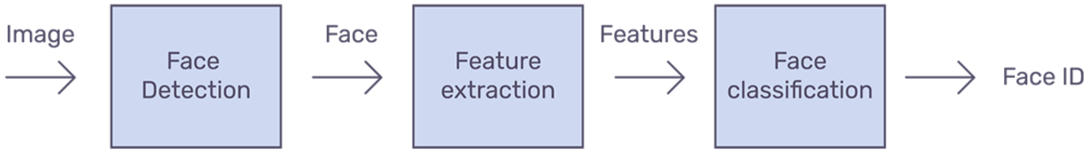

Figure 2-1: Simple face recognition pipeline.

2.1.3 Face Recognition Datasets

A prerequisite to training a successful deep neural network is a sufficiently large dataset. In the early stages of deep face recognition, models were usually trained on private datasets. For example, Facebook trained their model DeepFace [41] on 4M images of 4K people and Google trained their model FaceNet [37] on 200M images of 3M people. Both of these models achieved groundbreaking performance, but their work was not reproducible as the training data was proprietary. To address this issue, CASIA-Webface [27] provided the first ever widely-used public dataset which consisted of 0.5M images of 10K celebrities from around the world. Since then, there have been more large scale databases used for training face recognition models such as VGGFace2 [10], MillionCelebs [47], and MS-Celeb-1M [12]. For testing, the gold standard for face recognition is the Labeled Faces in the Wild (LFW) [16] which contains around 13,000 images of faces collected on the web.

Table 2.1: Commonly used Face Recognition Datasets for Training

Datasets	Publish Time	# photos	# subjects
Facebook [41]	2014	4.4M	4K
CelebFaces+ [40]	2014	202,599	10,177
CASIA WebFace [45]	2014	494,414	10,575
Google [37]	2015	>500M	>10M
MS-Celeb-1M (Challenge 1) [12]	2016	10M	100,000
VGGFace2 [10]	2017	3.31M	9,131
MillionCelebs [47]	2020	18.8M	636K

2.1.4 Data Bias

Deep learning networks rely on finding patterns in training data to generalize representations to new data. For example, to train a network to recognize an apple, one would present many example images of apples to the network and ensure that those images are labeled as apples. That way, when the network sees another apple, it would classify the image as an apple. However, if all of those apples were red and the network is told to classify a green apple, the network could falter, as a green apple was not present in the data that the network was trained on. The network would have a positive bias towards red apples.

These large scale face recognition datasets such as CASIA-WebFace [45], VGGFace2 [10], and MS-Celeb-1M [12] are usually created by scraping websites such as Google Images. Usually these datasets also consist of famous celebrities in some formal setting. For example, a common photo would be of a celebrity on the runway smiling as shown in Figure 2-2. These types of photos are incredibly different from daily life photos as external factors such as high camera quality or professional lighting are much more common in these celebrity photos. Given this type of training data, researchers have found that the performance of a model trained on one dataset can drop when tested on a completely different dataset. For example, a model trained on VGGFace [30] achieved a 98.95% accuracy on LFW [16] but only obtained 26%, 52%, and 85% on the Ugly, Bad, and Good partitions of the GBU [33] dataset, which is a dataset made up of difficult to recognize, average difficulty to recognize, and easy to recognize photos respectively [32].

Figure 2-2: Example of celebrity image.

2.1.5 Degradations on Face Recognition

Images in the real world are not so clear-cut and face recognition systems will run into images that have certain degradations. However, it has been shown that degraded image examples using increased Gaussian blur, Gaussian noise, and JPEG encoding have a destructive impact on the accuracy of a pre-trained ImageNet classifier, with destructive rates reaching 80% to 90% [22]. The impact of adversarial attacks and physical facial disguises (such as wigs or makeup) on pre-trained networks has been shown to decrease accuracy of face recognition, with VGG-Face achieving 33.76% Genuine Acceptance Rate (GAR) at 1% False Acceptance Rate (FAR) and 17.73% GAR at 0.1% FAR [31]. Other work has focused on the role of illumination and lighting on the accuracy of the face recognition models, indicating that accuracy can increase by roughly 40% when a model is trained with images that are illuminated from various angles. This could be due to a wide variety of factors, such as textural values in the image changing as a result of the illumination or the minimization between classes leading to increased false classifications [25] [18]. Some work has been done regarding the impact of color cues on recognition of faces by humans, revealing a statistically significant impact of color cues on recognition accuracy when the image quality is degraded ([46]). Similarly, the impact of related color cues on the recognition of objects by pre-trained networks has been explored, showing that different methodologies for transforming colored images to grayscale produce different accuracies by the network, with a coefficient of variation of up to approximately 10% [20]. Little work has been performed regarding the impact of color cues on the recognition of faces by pre-trained networks that are currently used as the 'gold standard' for face recognition and the comparison of those results to human-subject research.

2.1.6 Racial and Gender Bias within Face Recognition Systems

Within the data bias field, demographic bias is an urgent issue that has yet to be solved. Within the most common datasets used for training, white, middle-aged men tend to appear more frequently than other demographic groups as shown in Table 2.2 and Figure 2-3. It has recently been shown that around 80% of the existing large face datasets are biased towards "lighter skin" faces compared to "darker skin" faces [28]. This skew towards a certain group of people nonetheless becomes amplified, and causes deep learning models to have significantly different accuracies when the models are applied to different demographic groups. A prime example of this skew was demonstrated in Wang et al. [17], where researchers created the dataset Racial Faces in-the-Wild (RFW) [44] and demonstrated that commercial APIs work unequally in verifying faces for different races, with the maximum difference between mean error rates between Caucasians and African faces being 8.38% [17] as shown in Table 2.3.

Some research has been done on creating fair datasets that represent a wider range of races and also have an equal balance of each race. One example of a representative

Table 2.2: Demographic Information of Commonly Used Face Datasets [17][6].

Datasets	Race (%)				Gender (%)	
	Caucasian	Asian	Indian	Black	Female	Male
CASIA WebFace	84.5	2.6	1.6	11.3	41.1	58.9
MS-Celeb-1M	76.3	6.6	2.6	14.5	-	-
VGGFace2	74.2	6.0	4.0	15.8	40.7	59.3
LFW	66.0	9.8	7.2	17.0	-	-

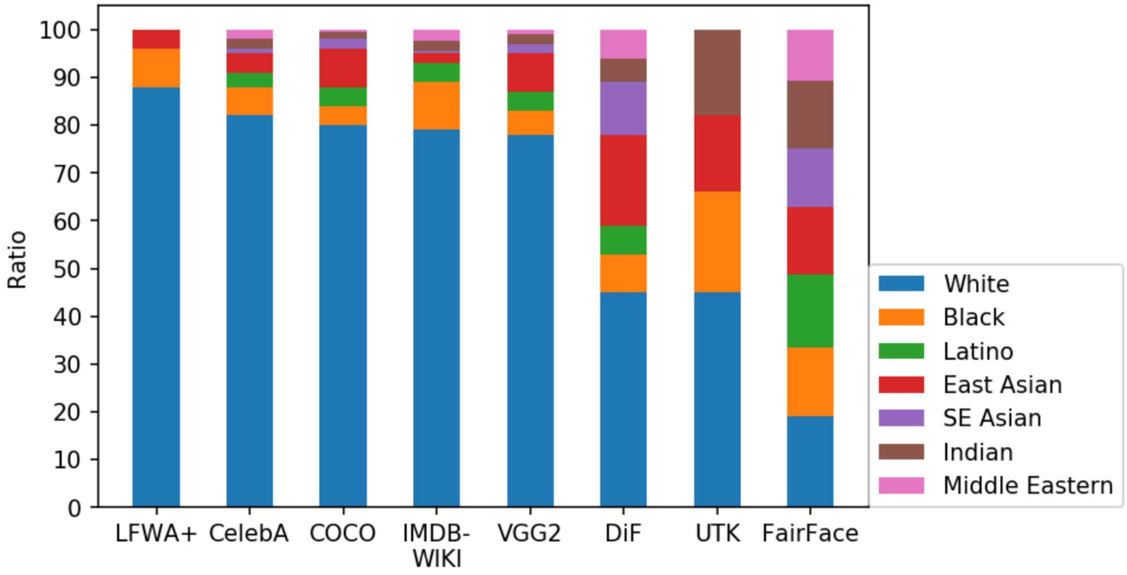

Figure 2-3: Racial composition in face datasets.

dataset is the Pilot Parliaments Benchmark, which is composed of parliament representatives from around the world and has around an equal number of lighter males, lighter females, darker males and darker females. For skin type labels, they used the Fitzpatrick six-point labeling system, which dermatologists have been using as a gold standard for skin classification. Using this labeling system which essentially works as a sliding scale for skin pigmentation, they created an intersectional dataset of 1270 individuals that range in both gender and skin color. The FairFace dataset [23] is another dataset which aims to mitigate bias by having a balanced number of images of these groups: Western White, Middle Eastern, East Asian, South East Asian, Black, Indian, and Latino. Research has found that more balanced racial datasets lead to a less biased model in the end. A model trained on the FairFace dataset achieved

Table 2.3: Commercial API Verification Accuracies for RFW [17]

Commercial APIS	LFW Accuracy (%)	Race Verification Accuracy (%)			
		Caucasian	Asian	Indian	Black
Face++	97.03	93.90	92.47	88.55	87.50
Baidu	98.67	89.13	90.27	86.53	77.97
Amazon	98.50	90.45	84.87	87.20	86.27
Microsoft	98.22	87.60	79.67	82.83	75.83
Mean	**98.11**	**90.27**	**86.82**	**86.28**	**81.89**

a gender accuracy mean of 94.89% across races while models trained on UTKFace, LFWA+, and CelebA achieved means of 89.54%, 82.46% and 86.03% respectively [23]
.

2.1.7 Evaluation of Face Recognition Models

Since many models for face recognition are designed to solve different problems, there are testing datasets that are designed to evaluate the models for different tasks. The two main classes of evaluation are face verification and face identification and are broken down further in Figure 2-4.

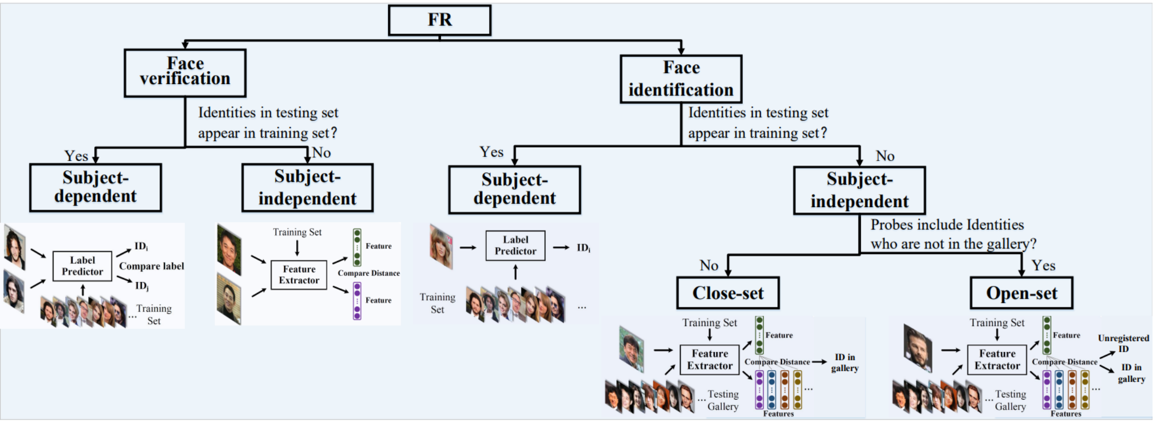

Figure 2-4: Comparison of different face recognition evaluation protocols [43].

Face Verification

The most common evaluation metrics are face verification, close-set face identification and open-set face identification. Face verification takes two images and determines whether or not the two faces are the same. The LFW testing dataset is specifically designed for face verification, and most of the accuracies presented throughout face recognition research relies on this metric.

Close-set face identification

Face identification works by taking a probe image and determining the identity of the subject. For close-set identification, there exists a gallery of identities, and models are to determine the identity of a probe image from the gallery. This type of identification is especially relevant for user driven searches such as forensic identification.

Open-set face identification

Open-set face identification works similar to close-set face identification, except some probe images may not exist in the gallery. This type of identification is most useful for face search systems, such as watch-list identification, where systems may need to reject images that do not exist in the gallery. As of now, there are few databases which cover the task of open-set face identification.

2.2 Human Face Recognition

One goal for artificial intelligence is to create systems which rival, and even surpass, that of human ability. The best way to understand how these deep neural networks could perform better is to take a look at how humans understand and conceptualize human faces. More specifically, we take a look at some examples of how humans are able to robustly recognize faces.

2.2.1 Image Color Degradation over Blur

The role of naturalistic image degradation on face recognition performed by humans has been highlighted in previous research [46]. In the study, humans were given degraded faces and asked to recognize the face. Specifically, the degradations chosen were normal full-color images, grayscale images, and hue shifted images across many levels of blur. In humans, experimental evidence showed that recognition performance on grayscale images is not significantly different than on full-color images, at least at high resolutions [46]. However, as the blur level progressively increases, humans perform significantly better in recognizing the blurred full-color images than the blurred grayscale images, showing that color cues are in fact important for recognition. When the hue of the face images is shifted 21.6° in addition to the applied blur, human recognition is at the same level as for full-color images, which is in turn significantly better than for the grayscale images [46]. The results are summarized in Figure 2-5. This suggests that color is important for low level tasks such as segmenting out different parts of the face, and not higher level diagnostic information, like identifying eye color.

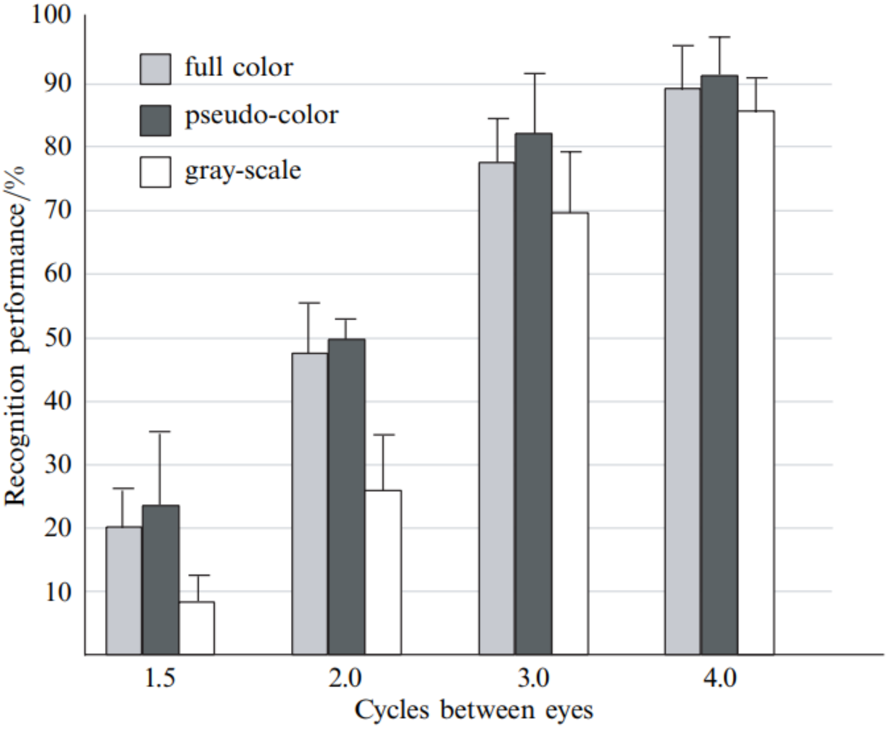

Figure 2-5: Human face recognition performance with full color, pseudo-color and grayscale images at decreasing levels of blur [46]. One "cycle" is equivalent to 2 pixels, so more cycles between the eyes corresponds to a higher resolution image.

2.2.2 Importance of Eyebrows

In another human study performed by Sadr et al. [36], the eyebrows were found as an important facial feature for face recognition by humans. Researchers erased the eyes or the eyebrows off of 50 celebrity face images, as shown in Figure 2-6. The subjects were then shown these images and asked to name the celebrities. The performance was measured by the percentage of images the subject could correctly identify. The study found that performance with images lacking eyebrows was significantly worse than the normal, unmodified images and the images without eyes. One reason for why eyebrows are important for human face recognition could be that the eyebrows convey emotion and other nonverbal signals, thus humans are biased to attend to the eyebrows to interpret these signals. Another explanation could be that eyebrows serve as a "stable" facial feature. Since eyebrows are large and generally are high-contrast facial features, eyebrows can survive a substantial amount of image degradation, like viewing from a far distance, and thus have become important for human recognition.

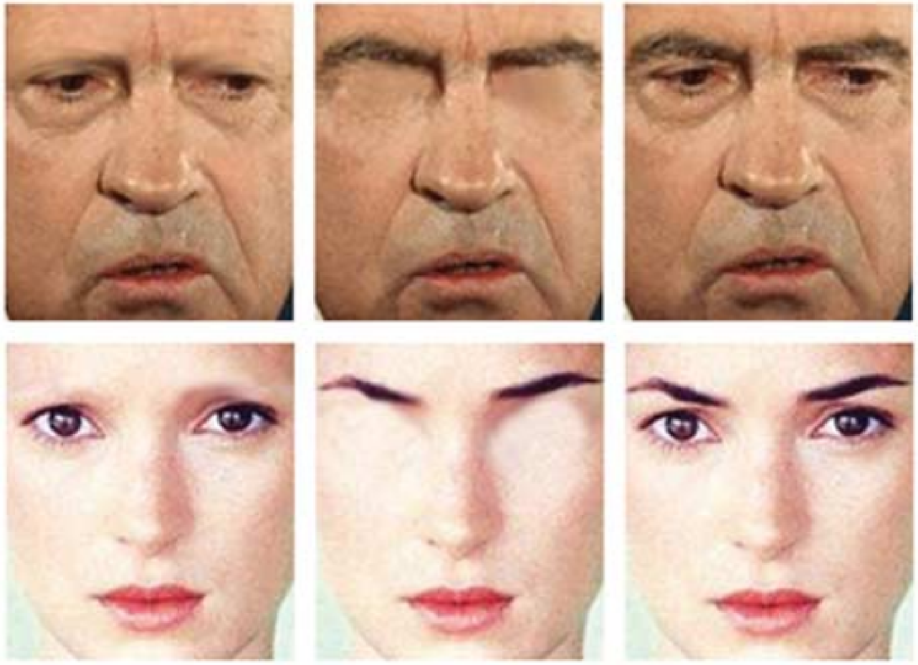

Figure 2-6: Sample images of President Nixon and Winona Ryder with no eyebrows, no eyes, and no alteration.

2.2.3 Other Human Results

Other noteworthy degradations summarized in Sinha et al. [39] are the following:

- Vertical or horizontal compression of face images do not affect recognition performance (Figure 2-7).

- Contrast polarity inversion dramatically decreases recognition performance (Figure 2-8).

- High-frequency information on its own does not result in good face recognition performance (Figure 2-9).

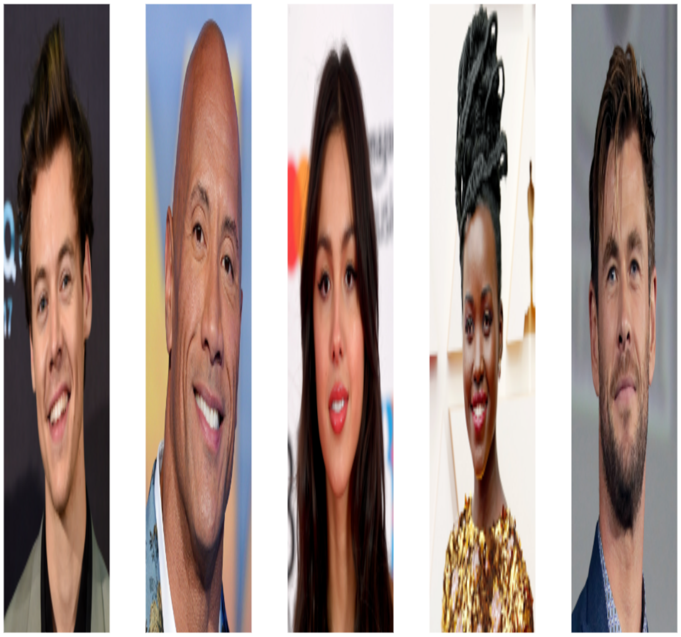

Figure 2-7: Images of celebrity faces that have been compressed to 25% of their original width. The celebrities from left to right are as follows: Harry Styles, Dwayne "The Rock" Johnson, Olivia Rodrigo, Lupita Nyong'o, and Chris Hemsworth.

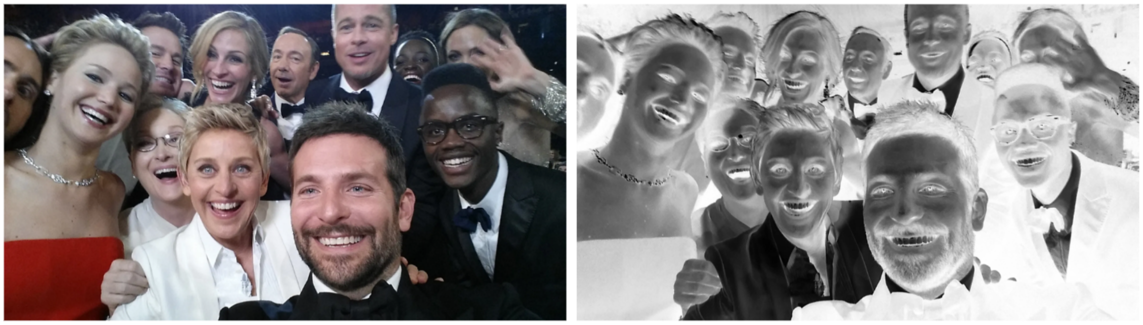

Figure 2-8: Normal image and negative contrast image containing several well-known celebrities. (Photographed during the 2020 Oscars.)

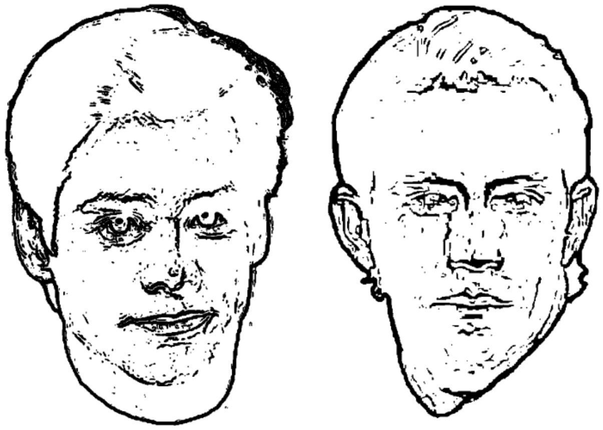

Figure 2-9: Sample high-spatial frequency information images of Jim Carrey and Kevin Costner.

Chapter 3

Exploration of Methods

This chapter steps through design decisions and experiment elements given the problem we are trying to solve. This chapter discusses the degradations that we apply in the experiments, the model we chose to evaluate, and the evaluation metric.

3.1 Degradations

The first step was to choose the degradation we would like to apply to images in order to compare to humans. Within the human results, the choices were vertical and horizontal compression, color cues over different levels of blur, creating line drawings from images, removing certain facial features, and contrast polarity. For this experiment, we wanted to choose a degradation that is encountered often in the application of face recognition, so we experimented with color cues over varying levels of blur. In other words, how does face recognition technology perform at full color, grayscale, and at different shifts of hue at increasing levels of blur.

3.1.1 Grayscale

There are number of ways to convert RGB (red-green-blue) images to grayscale images, such as the average method and the weighted method. The average method is as follows:

$$\text{Grayscale} = (R + G + B)/3$$

This method is simple but does not work as well as expected as humans react differently to red, green and blue. Generally, humans are more sensitive to green light, less sensitive to red light and even less sensitive to blue light. Due to this sensitivity, the more common conversion is using the weighted method:

$$\text{Grayscale} = 0.299R + 0.587G + 0.114B$$

This method is also called the luminosity method and weighs red, green and blue according to their wavelengths. Throughout the paper, when we refer to grayscale images, we use this method to convert to grayscale.

3.1.2 Hue

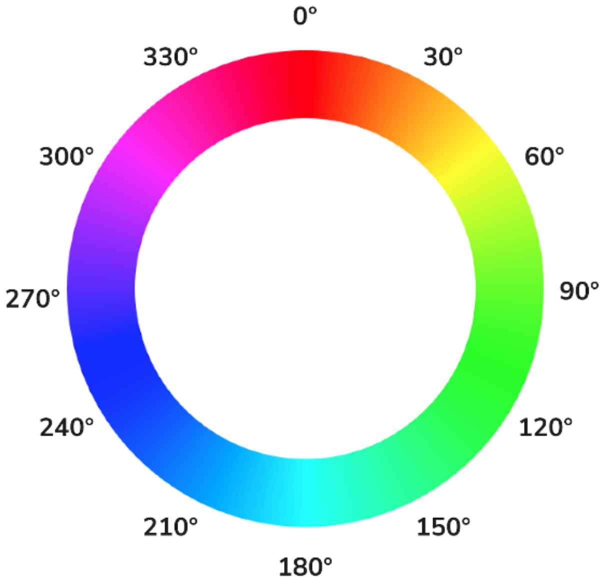

Figure 3-1: The color wheel. All of the colors in this wheel have the same saturation and the same lightness, but differ in hue.

Hue is defined as the degree which stimulus can be described as red, orange, yellow, green, blue, or violet. Hue is represented by a single number which corresponds to a position on the color wheel, and is thus represented by degrees ranging from 0 degrees to 360 degrees as shown in Figure 3-1.

Apply a hue shift

The easiest way to apply a hue shift is to convert an RGB image to an HSB (hue-saturation-brightness) image and then shift the hue values of all the pixels in the image by adding a certain amount to the Hue value H. In other words, one could shift the hue by the desired angle a to get the new hue H':

$$H' = H + a \mod 360^\circ$$

In Figure 3-2, the image is hue shifted by 180° which translates each pixel's color to the color opposite of the color wheel from Figure 3-1.

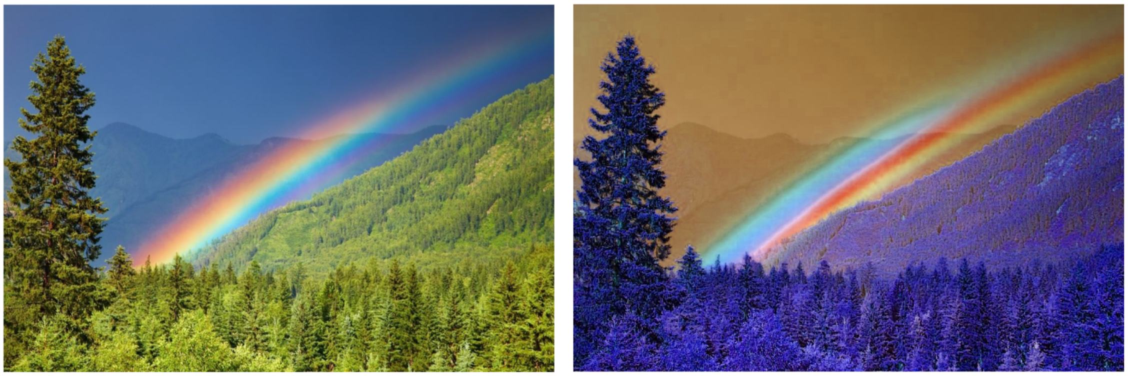

Figure 3-2: Example image of a rainbow hue shifted by 0° and 180°.

3.1.3 Gaussian Blur

In order to blur an image, there are multiple different methods. The version that is generally known to be the smoothest way to blur an image is through Gaussian blur, which is a type of image-blurring filter that uses a Gaussian function (which also expresses the formula for a normal distribution) for calculating the transformation to apply to each pixel.

The formula for a two dimensional Gaussian function is as follows:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Using this formula, one can create a Gaussian matrix of radius r, or of size $(2r + 1) \times (2r + 1)$ and standard deviation σ to convolve the image. When we refer to an image of blur radius r, we create a $(2r + 1) \times (2r + 1)$ Gaussian matrix where $\sigma = \frac{r}{2}$ and convolve the image with the matrix. As the blur radius increases, the strength of blur increases, and as the blur radius decreases, the strength of the blur decreases. An example of an image with blur radius 5px and blur radius 30px is shown in Figure 3-3.

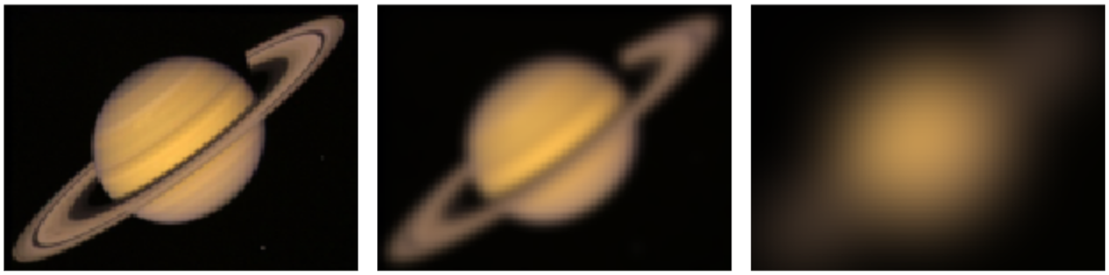

Figure 3-3: Example image of Saturn at blur radius 0px, 5px, and 30px.

3.2 Evaluated Neural Network

In these experiments, we chose to use the ResNet-101 architecture for our model.

3.2.1 Residual Network

The main goal of a residual network is to create a deeper network with more layers. It was created to solve the degradation problem within neural networks, where much of the accuracy of a model is concentrated in a handful of layers and then degrades rapidly. In other words, if one were to take out a layer in a traditional deep CNN, the layer could possibly be important and rapidly decrease the accuracy of the network by nearly 40%.

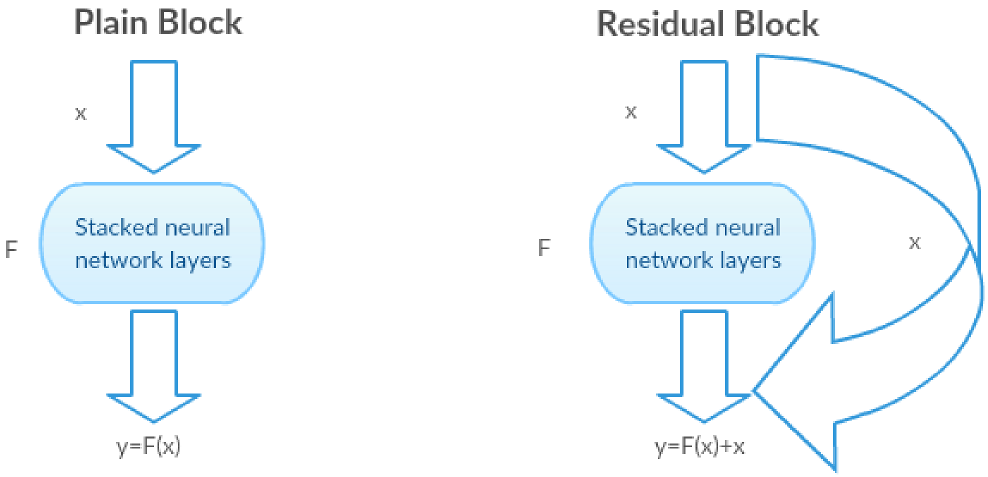

Figure 3-4: Comparison between plain neural network block and a residual neural network block.

Instead of learning a direct mapping from $x \rightarrow y$ (or from input to actual output) using a function $H(x)$ (a few stacked non-linear layers), we will define a residual function using $F(x) = H(x) - x$ where $F(x)$ represents the stacked non-linear layers and x represents the identity function where the input is equal to the output. We can reframe this equation to get $H(x) = F(x) + x$.

If the identity mapping x is optimal, we can easily push $F(x)$ to 0, or push the residuals to 0, rather than creating a function that tries to map x to x, or an identity function. In other words, it is much easier to come up with a solution like $F(x) = 0$ rather than $F(x) = x$ using a stack of non-linear convolutional neural network layers. Figure 3-4 shows a comparison between a plain block and a residual block. This function $F(x)$ in a residual network is called the residual function.

There are two types of residual connections:

- The identity x can be used directly if the input and output are the same dimension:

$$y = F(x, \{W_i\}) + x$$

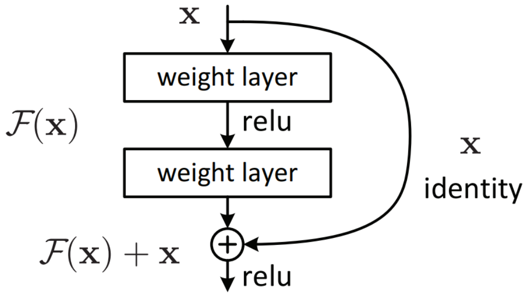
x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 3-5: One block of the ResNet architecture.

- The dimensions can different between the input and output, so the network can either A) perform the same identity mapping but with extra zeros as padding, or B) the projection shortcut can be used to match dimension using the formula below:

$$y = F(x, \{W_i\}) + W_s x$$

Within ResNet-101, each block is 3 layers deep, as shown in Figure 3-6.

3.2.2 Evaluated Model: ResNet-101

The specific model that we used contains 104 convolution layers, 104 batch normalization layers, 100 element-wise layers, 1 padding layer, 2 pooling layers, 33 total layers and 1 flatten layer. The network was originally trained as a classifier, but for generalization on new faces, the final classification layer was removed to turn the network into an encoder. This encoding gives us a unique vector for each image that is passed through the network. The network is available for download on the Wolfram Neural Net Repository [19].

3.3 Training Dataset

The ResNet-101 model was trained on the Augmented CASIA-WebFace dataset [27]. The original dataset, CASIA-WebFace, is a collection of 494,414 facial photographs

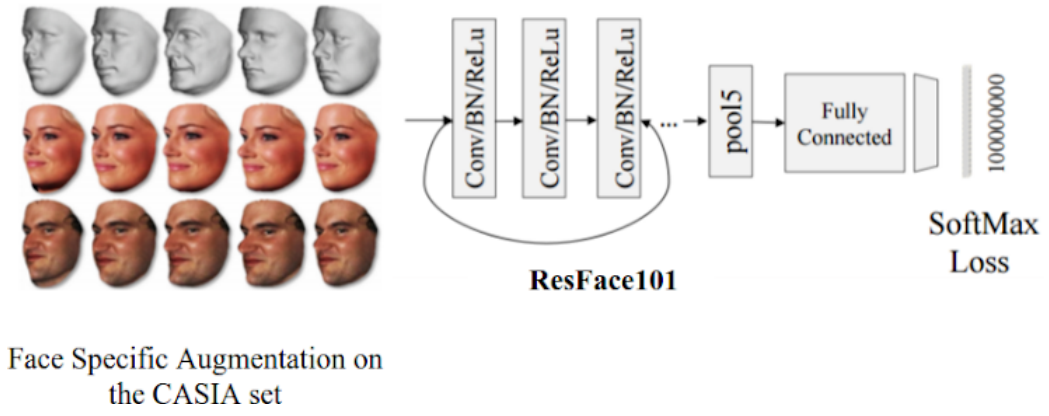

Figure 3-6: Summary graphic of ResNet-101 Trained on Augmented CASIA-WebFace Data.

of 10,575 subjects. Additionally, a far greater per-subject appearance was achieved by synthesizing pose, shape and expression variations from each single image.

3.3.1 CASIA-WebFace Dataset

The original CASIA-WebFace dataset has a racial distribution of 84.5% Caucasian, 2.6% Indian, 1.6% East Asian, 11.3% Black as shown in Figure 3-7.

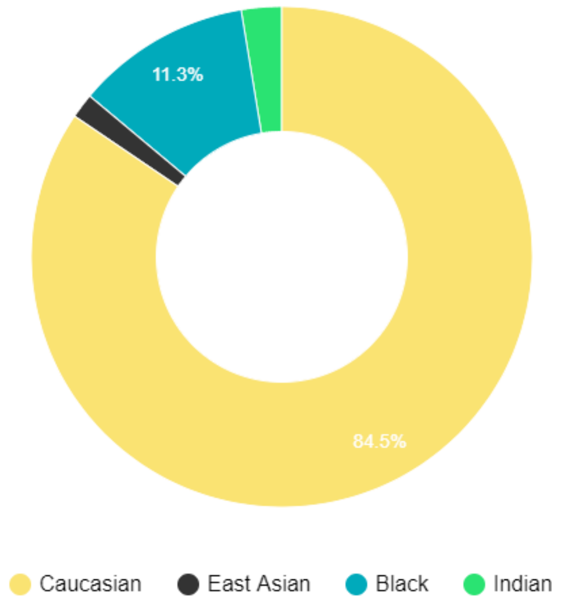

Figure 3-7: Racial breakdown of CASIA-WebFace dataset.

Table 3.1: Verification Accuracy of ResNet-34 on RFW [44]

Training Database	LFW Accuracy (%)	Race Verification Accuracy (%)			
		Caucasian	Asian	Indian	Black
Casia WebFace [45]	99.40	92.15	83.98	88.00	84.93

Previous studies have compared how a ResNet-34 model trained on Casia-WebFace [45] has performed for different races using the RFW dataset [44]. In terms of verification accuracy, the model performs the best on Caucasian faces (92.15%) and performs the worst on East Asian faces (83.98%). The results are summarized in Table 3.1.

3.3.2 Augmented CASIA-WebFace

To enlarge the dataset, additional images were synthesized by varying the pose, shape and expression of existing images.

Pose

To synthesize variations of pose for a face image, researchers first applied a landmark detector to pinpoint certain facial features. Given these landmarks, one can then estimate the six degrees of freedom pose for the face using correspondences between detected landmarks in a 2D space and points labeled on a 3D generic face model. After creating the face model, one can render novel pose variations, as shown in Figure 3-8.

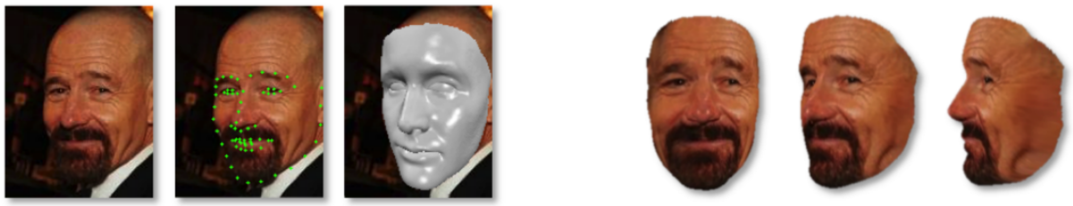

Figure 3-8: Adding pose variation by synthesizing new viewpoints.

Note, these new faces are rendered on a black background as the original background is not preserved during rendering.

3D Shape Variation

3D shape variation is similar to pose synthesis, where given a 2D face, one can project the face onto a 3D model to create new photos at different angles. However, instead of simply choosing different angles, researchers chose 10 different 3D face shapes to project onto, greatly expanding the number of images per identity.

Expression

For expression, researchers synthesized expression variations specifically reducing deformations around the mouth. Given the 2D detected landmarks, one can fit images to 3D generic face models with expressions such as *mouth-opened*, *mouth-closed*, and *smile*. Some slight artifacts are created with this method, however it does not alter the general facial appearances and are less pronounced than noise often present in large image databases.

Final Model

The final trained model with the augmented dataset has a 98.06% accuracy on LFW and a 100% Equal Error Rate.

3.4 Evaluation Metric

In order to most closely resemble the results of the human study of recognition on degraded datasets, we chose to work with an open-set evaluation metric. More precisely, we were interested in how the network clusters images based on its encoding.

3.4.1 Recognition Percentage

The images were passed into the ResNet-101 model and encoded, which transforms an image into a vector within a vector space that preserves high quality clustering for similar images that are within the training distribution. An overly-simplified two-dimensional caricature is shown in Figure 3-9 to help the reader visualize the encoding

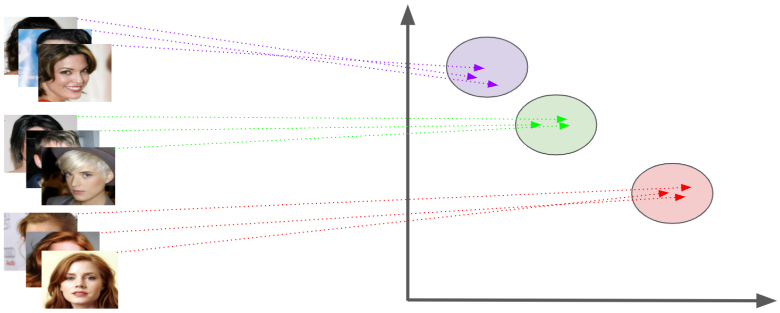

Figure 3-9: Identities Encoded to 2D Vector Space Generalization

with ideal theoretical clustering. From these encodings, we were able to assess the accuracy and consistency of the network for each degraded dataset.

Classifying an image

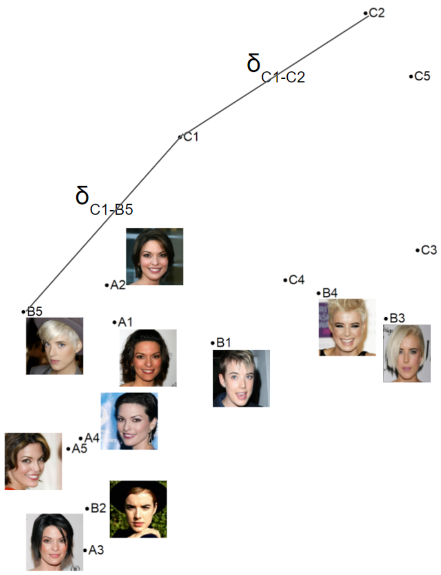

Figure 3-10: Individual Image Classification: In this example, we calculate the distance between $C1$ and $B5$ and $C1$ and $C2$. $C1$'s identity classification would be C, $A1$'s would be B, $B1$ would be B and so on.

For a given image, we calculated the Euclidean distance from the image to all of the other images of that individual and averaged the result. Then we took that same image and calculate the Euclidean distance to all images of another individual, and calculated the average of those distances. We continued this process until we calculated the average Euclidean distance between the initial image and all images of all the other individuals as shown in Figure 3-9. If, on average, the image is closer to the images in its own class, then it is classified correctly (i.e. if a particular image of Brad Pitt is on average closer to the other images of Brad Pitt than it is to photos of other celebrities, then that photo will be classified as Brad Pitt). A photo is misclassified when the image is, on average, closer to the images of another celebrity. For example, if a particular image of Brad Pitt is, on average, closer to images of Barack Obama, it would thus be misclassified as Barack Obama. In summary, the identity that an image is closest to (on average) determines the classification of the image, where the identity is the group of all images of that identity excluding the one being assessed.

"Recognizing" an Identity

After finding the identity classification of each image of a certain individual, we determine an identity as "recognized" when at least 75% of the images of the individual are classified correctly according to the ground truth. This threshhold measure of 75% arises from research standards from brain and cognitive science, on which much of this research is based [46].

Chapter 4

Experiment and Results

4.1 Experiment 1: Degradations' Effects on the Network

The goal of this study was to compare the task performance of neural networks to developmentally-typical adult humans. We compared our results to the previous human trials where humans identified the name of a celebrity based on an image of their face [46]. We designed our computational experiment as an analogy to this celebrity face recognition task by evaluating the ability of a neural network to encode images of celebrities in a vector space that keeps images of each celebrity closely clustered together with ideally non-overlapping clusters that represent each celebrity.

4.1.1 Testing Dataset

Our final base dataset, or $D_{\text{color}}(0)$, for this analysis had 97 unique celebrity identities, each with 19-25 face images. We took these images from the CelebAMask-HQ dataset [24], which is a large-scale image dataset with 30,000 high-resolution face images selected from the larger CelebA-HQ dataset. For our experiment, we wanted to analyze how well the trained model groups the vector encodings of images by identity, so we needed identities with enough images that a cluster quality score could be computed; we thus only kept identities with 19 to 25 unique images of them.

4.1.2 Degraded Datasets

We created 3 sets of degraded datasets from $D_{\text{color}}(0)$. We created a grayscale dataset ($D_{\text{grayscale}}(0)$), a 21.6° hue shifted dataset ($D_{21.6^\circ}(0)$), and 180° hue shifted dataset ($D_{180^\circ}(0)$) by applying grayscale and hue shift filters to $D_{\text{color}}(0)$. Examples of an image from each dataset is shown in Figure 4-1. The 21.6° hue shift transforms images to have a yellow tint which still retains lots of natural human skin colors while the 180° hue shift transforms images to have a blue tint which does not resemble human skin tones and is not naturalistic. The 21.6° was chosen since it is the same hue shift used in the human study [46] which will be used later in the discussion.

We then blurred every image in each of the datasets using a Gaussian blur radius of r and a standard deviation of $\frac{r}{2}$ to created new blurred datasets $D_*(r)$. In total, we generated 60 datasets: $D_{\text{color}}(r)$, $D_{\text{grayscale}}(r)$, $D_{21.6^\circ}(r)$ and $D_{180^\circ}(r)$, each with $r = 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150$.

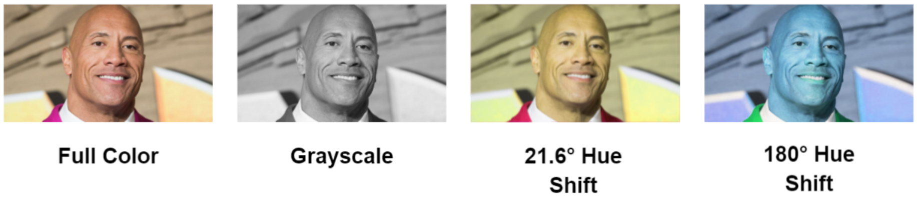

Figure 4-1: Dwayne "The Rock" Johnson in full color, grayscale, 21.6° hue shift, and 180° hue shift conditions.

4.1.3 The Network's Performance on Different Levels of Blur

Figure 4-2 summarizes our analysis of the recognition robustness of ResNet-101 on degraded datasets. The graph shows how the ResNet-101's "Recognition" performance changes over different blur radii for full color images, grayscale images, and hue shifted images (both 21.6° and 180°). Additionally, Figure 4-3 focuses on blur radii of 0 to 60 pixels and shows, with error bars, how the recognition performance changes for each color degradation over different blur radii. At different blur levels, the neural net displays interesting "recognition" performances for each of the degradations – these are discussed in the following subsections.

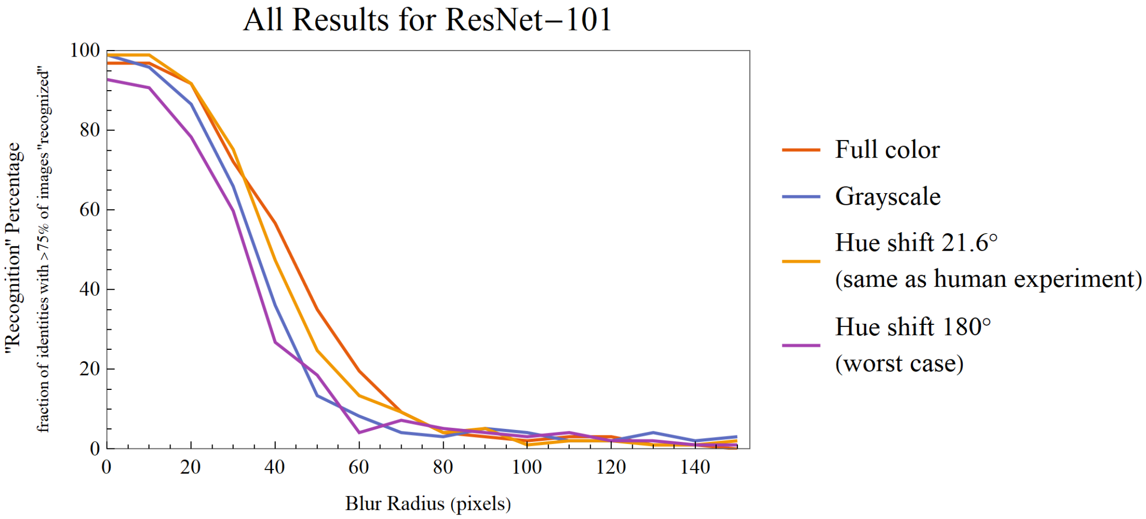

Figure 4-2: Line graph of all results for ResNet-101 comparing recognition percentage vs blur radius.

0 to 35 pixel radius blur

Figure 4-4 shows an example of images blurred from 0 to 35 pixels. For this range of blur, we see that the model's performance on full color, grayscale and the 21.6° hue shifted images is statistically the same. Additionally, its performance on the 180° hue shift was significantly worse than on full color, for this range of blur. This behavior is similar to the human study, where human subjects performed similarly well on color, hue shift and grayscale images at a high resolution, with an exception for the worst color hue shift.

36 to 80 pixel blur radius

Figure 4-5 shows an example of images blurred from 35 to 80 pixels. In this range of blur, the performance on full color is still statistically the same as on the 21.6° hue shift. We also see that the performance on grayscale and 180° hue shifted images are statistically the same. However, a difference now is that the grayscale performs significantly worse than the full color images.

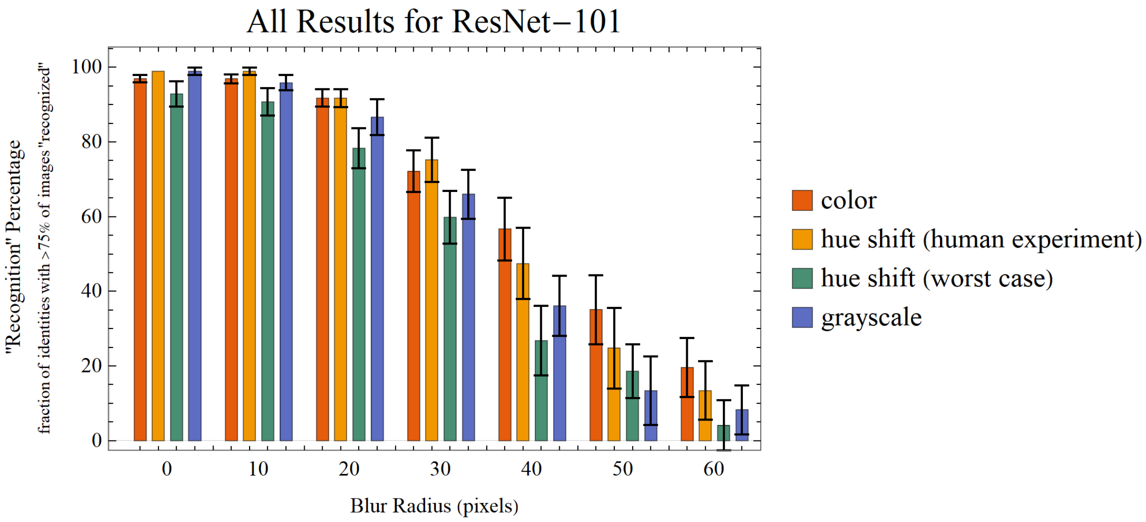

Figure 4-3: Bar chart of all results for ResNet-101 comparing recognition percentage vs blur radius.

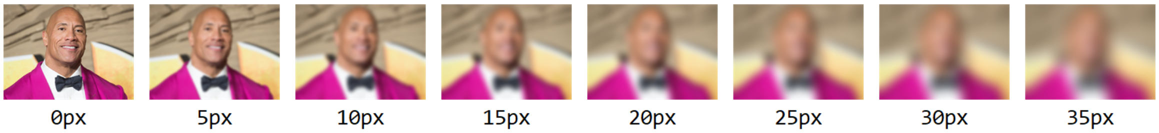

Figure 4-4: Dwayne "The Rock" Johnson blurred from 0 pixels to 35 pixels with increasing intervals of 5 pixels.

81 and beyond pixel blur radius

Figure 4-6 shows an example of images blurred from 80 to 170 pixels. At this level of blur, the model performs identically on all of the degradations, between 0% and 5%.

4.1.4 Comparing Humans and ResNet-101

Next, we aimed to compare our network results to the human results from Sinha 2002 [46].

Converting Human Results to Gaussian Blur

The first step in standardizing our results with those from Sinha 2002 [46] was to scale “cycles between the eyes” to Gaussian blur radius. One cycle is equal to 2 pixels, so this is essentially a measure of the distance between the eyes in a given image. For

Figure 4-5: Dwayne "The Rock" Johnson blurred from 35 pixels to 80 pixels with increasing intervals of 5 pixels.

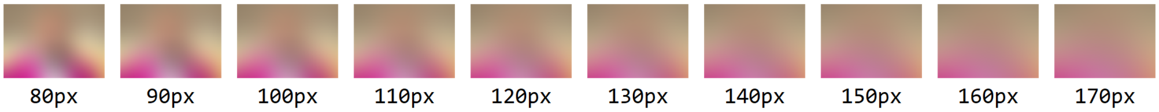

Figure 4-6: Dwayne "The Rock" Johnson blurred from 80 pixels to 170 pixels with increasing intervals of 10 pixels.

the human experiment, the researchers scaled down the images so that there was a certain number of cycles between the eyes and used Photoshop to blur images so that it minimized the distance between the scaled-up image and the blurred image. For this study, we found the average cycles of the eyes for all of the faces in our dataset. Celeb-A-Mask-HQ provides masks for different parts of the face, and we used the masks for the right and left eyes to find the number of pixels between the center of the eyes. The next step was to apply the same amount of Gaussian blur, corresponding to cycles between the eyes as done in the human study. To do this, we first resized our images to smaller images and then enlarged them to get the correct cycles between the eyes. Then we applied different radii of Gaussian blur to find the radius which minimized the image distance of the enlarged version of the face as seen in Figure 4-7. This follows the same methodology as in Sinha 2002 [46] and allows us to directly compare the blur from the human study to our results.

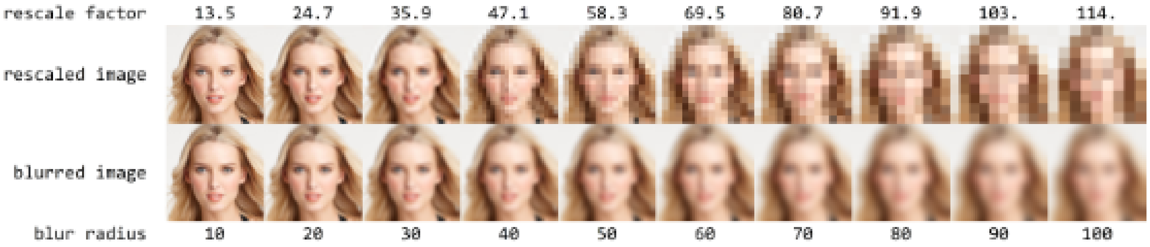

Figure 4-7: Re-scaled images vs corresponding image distance minimizing radius Gaussian blur.

Comparison between Humans and ResNet-101

In Figure A-1, we directly compared each degradation from the human study to the corresponding degradation from our results. For full color, 21.6° hue shift and grayscale images, the recognition performances display the same characteristics and are monotonically decreasing. We also see that humans performed better than the trained model. It's difficult to directly compare the recognition accuracies as the methods of quantifying accuracy are not the same. Additionally, we can characterize the model's recognition performance curve and the human subjects' curve as decreasing logistic functions as shown in Figure A-1. We see that for full color and grayscale images, the human logistic function is essentially shifted about 5 to 10 pixels to the right of the model's curve. The curves for humans and for the model does not align nearly as well for the 21.6° hue shift. This suggests that for full color and grayscale images, the ResNet-101 model trained on augmented data can serve as an accurate tool which resembles humans' ability to recognize.

4.2 Experiment 2: Degradations' Effects on Different Races and Genders

The goal of this experiment was to compare the performance of the network for different races.

4.2.1 Testing Dataset

We noticed that many of the datasets that are publicly available contain lots of noise, so we handcrafted a dataset of 102 individuals with 20 front facing photos each. Similar to the previous experiment, we needed enough images of each identity in order to calculate the cluster quality score. Additionally, we balanced the dataset by race and gender and selected celebrities in between the age range of 20 and 55. In terms of race, we chose to work with Black, East Asian and White identities.

4.2.2 Degraded Datasets

Instead of creating 3 sets of degraded datasets, we expanded the range of hue shifted values by choosing from $36i^\circ$ where i ranges from 0 to 10. This gives us tints from the full color spectrum as shown in Figure 4-8. Additionally, we created a grayscale dataset as in the original experiment. After collecting all of these faces, we cropped the images tightly around the face.

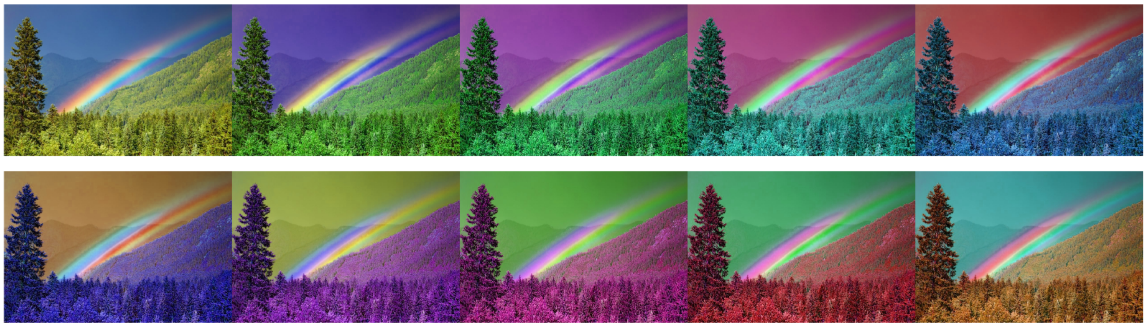

Figure 4-8: Hue shifts ranging from 0° to 324°, shifted by 36° from each other.

4.2.3 The Network's Performance on Different Races

The results from this experiment was surprising given the prior research done in this field. Usually White identities perform better overall followed by East Asian and then Black identities. However, the results for the different degradations did not follow this trend.

Full Color Results

We found that overall, East Asian identities performed the best, with Black identities performing second best and White identities performing the worst at full color as blur increases (also shown in Figure 4-9). In terms of race and gender, at full color and between 0px to 60px of blur, we found that Black and East Asian men perform the best overall while White and Black women perform the poorest in terms of recognition as shown in Figure B-1. After 60px of blur, Black women perform the poorest in terms

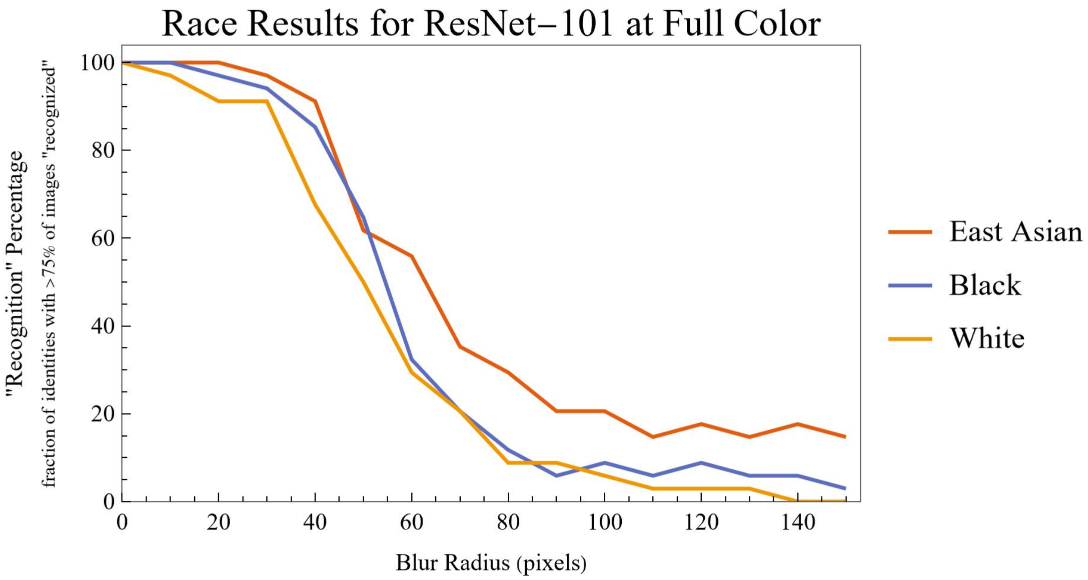

Figure 4-9: Line graph of ResNet-101's performance on each race as blur increases for full color.

of recognition (recognition goes to 0% by 80px of blur) and East Asian men perform the best in terms of recognition.

Grayscale Results

For grayscale, East Asian identities continued to have the best performance throughout all levels of blur, followed by Black identities and then White identities as shown in Figure 4-10. At around 80px of blur, Black and White identities perform similarly and at 100px of blur all of the race perform poorly. Within each gender and race, White women perform the poorest until 60px of blur and East Asian men perform the best until 60px of blur as shown in Figure B-2. After 80px of blur, there is enough noise to not be able to accurately report the results.

Hue shift results

Unsurprisingly, throughout all of the hue shifts, the same pattern as the full color and grayscale results occur as shown in Figure 4-11. Throughout all of the hue shifts, East Asian men, Black men and East Asian women were the top 3 performers, usually in this order as shown in Figure B-3. At hue shift of 216°, East Asian women surpassed

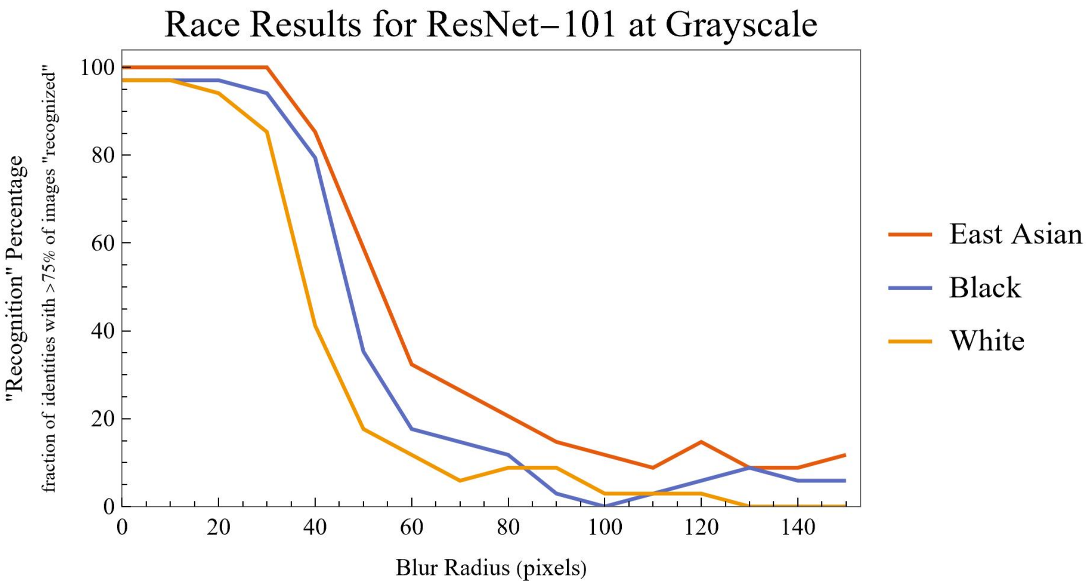

Figure 4-10: Line graph of ResNet-101's performance on each race blur increases for grayscale.

Black men at lower levels of blur (below 50px of blur), but overall the performances stayed in the same order (East Asian men, Black men, East Asian women) for these subgroups. White women performed the poorest for all hue shifts at low levels of blur, usually followed by White men and then Black women.

Worst performing hue shifts per race

We wanted to see if there was any significance in the hue shifts that made each race perform the worst. For White, Black and East Asian identities, the hue shifts that resulted in the worst performance was 108°, 216°, and 108° respectively. The worst case tints are also shown in Figure 4-12. This follows similar results to experiment 1, where drastic hue shifts, like changing natural skin tones to green or blue, resulted in the worst performance of the network.

How does hue shift affect recognition?

In seeing the results over all of the hue shifts for the different demographic subgroups, we found that the accuracy differed wildly for some subgroups, and not as much for others. For the hue shifts, we calculated the standard deviation of the accuracy for

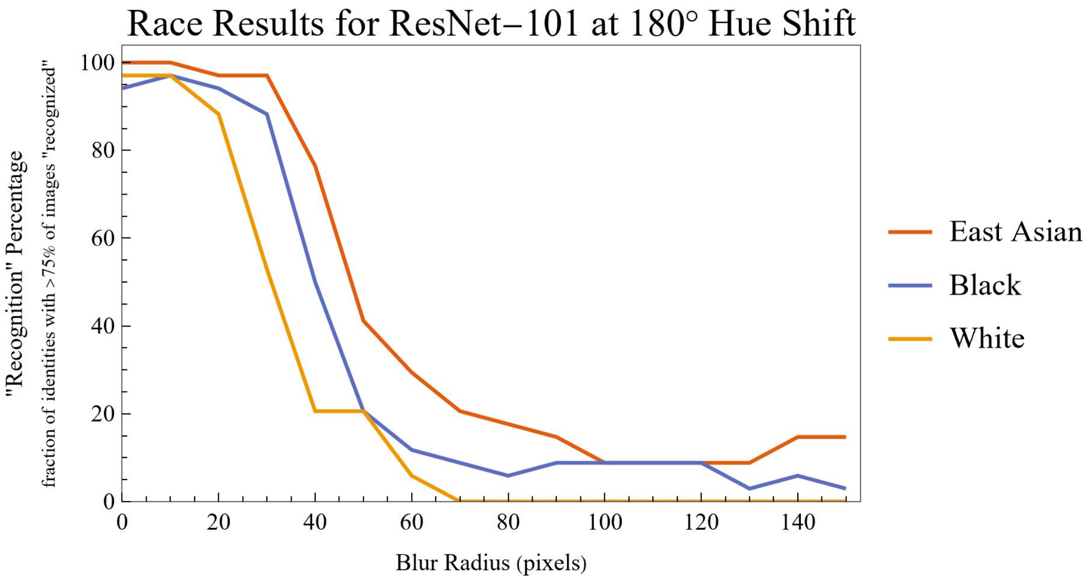

Figure 4-11: Line graph of ResNet-101's performance on each race blur increases for full color, grayscale and hue shift of 180°.

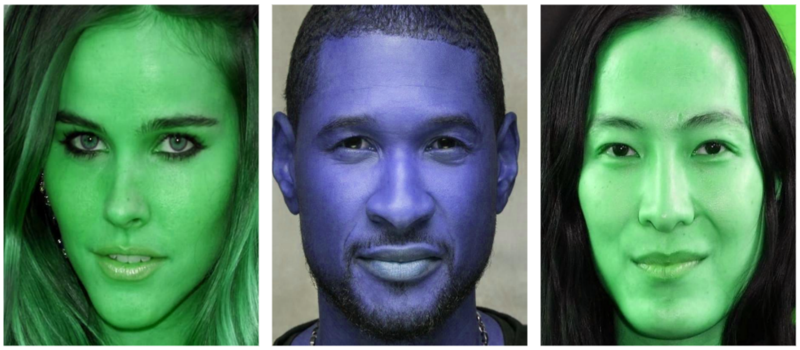

Figure 4-12: The worst case hue shifts for White, Black and East Asian identities.

each race for all of the hue shifts at each point in blur as shown in Figure 4-13. All of the identities' standard deviations peak 40px to 50px of blur and then decrease significantly after that. White and Black identities' standard deviations of recognition performance peak at around 17%, while East Asian identities peak at around 12% indicating that East Asian identities were slightly less affected by hue shift in general. Overall, it seems that hue does not affect recognition performance at low levels of blur ($< 20\%$) but rapidly becomes more important in detecting faces at 30px to 40px of blur. After around 60px of blur, the standard deviation decreases again, indicating that the network has an equally poor performance for all of the hue shifts. This is also reinforced by how the performance curves level off at around 0% to 20% after 60px of blur as shown in Figure 4-9.

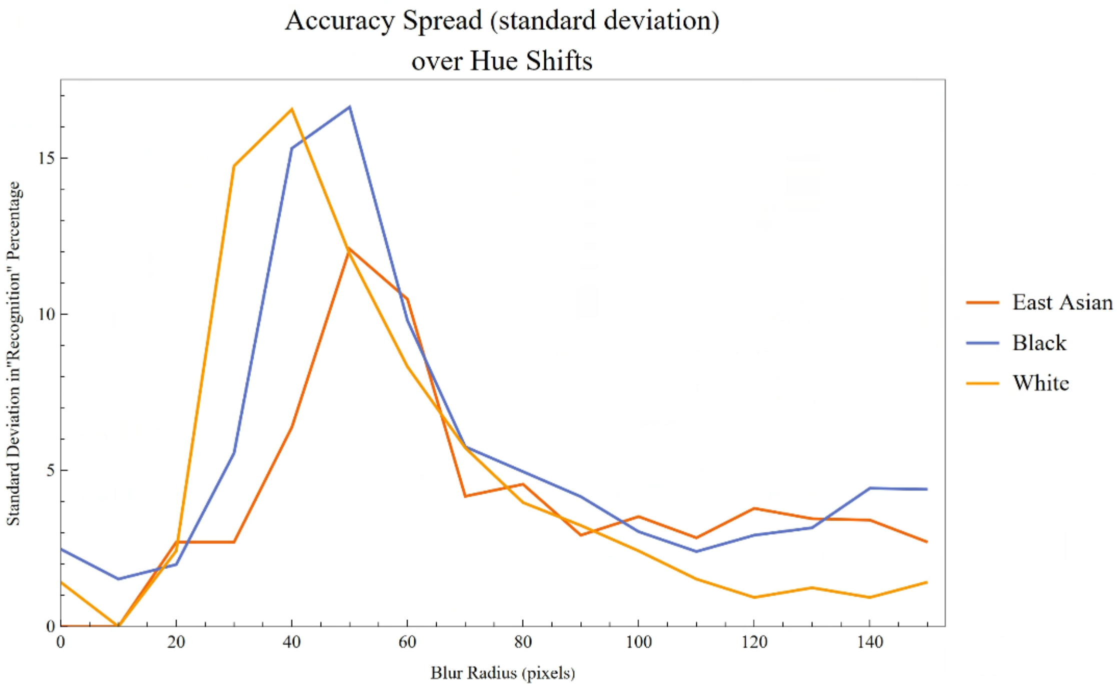

Figure 4-13: The standard deviation in performance over all of the hue shifts for each race.

4.3 Discussion and Conclusions

4.3.1 Experiment 1: Network's Overall Performance

The neural network used in this experiment was trained on full color images, so we expected to see the model perform worse on any kind of hue shift and grayscale degradation. However, that is not exactly reflected in the results, as the 21.6° hue shifted images never significantly under-performed in comparison to the full color images as shown in Figure 4-3. For humans, Sinha hypothesized that color may contribute to recognition primarily by facilitating low-level image analysis tasks, such as segmentation of different parts of the face, rather than providing diagnostic information, like eye color [46]. For the model, the same applies for the 21.6° hue shift. However, the 180° hue shift did significantly affect the model's ability to recognize, which contradicts the idea that color (even if hue shifted) is needed for low level tasks. However, this suggests that natural color is important to the network's recognition performance, and that simply having a colored image does not necessarily contribute to better facial recognition. This also suggests that the original color of the faces are represented in each image's vector encoding in some way.

The similarities in the human trials and the model trials for full color, grayscale and 21.6° hue shift also leads us to wonder how humans would perform on facial recognition tasks that utilized 180 degree hue shifted images. It might be the case that naturalistic color is important for humans as well in recognizing faces, and further work could include a human trial along different levels of hue shift.

The idea that naturalistic color is important at higher levels of blur has further implications for downstream usage of these networks. In the security aspect of facial recognition, a simply hue shift or grayscale degradation on low resolution data has the ability to significantly affect recognition performance. Some examples of when this could occur is trying to determine the identity of someone from low resolution black and white security camera footage.

There are some limitations to our study. First, the diversity of this testing dataset is limited: the dataset was mostly Caucasian faces with an even split between male and female genders. Additionally, the human experiment used in comparison with our results was performed in 2002 with limited data and limited participants. It would be ideal to collect more human data from a wider range of participants using the same dataset we used for the face recognition system. In terms of neural network architectures, further work could include a wider range of neural networks such as FaceNet [37] and CLIP [34]. Finally, similar work with different degradations (such as line drawings of human faces as shown in Figure 2-9) would be worth studying with a similar methodology as the one defined in this paper.

4.3.2 Experiment 2: Network's Performance on Demographic Subgroups

The neural network used in this experiment was trained on mostly White faces (84.5% White), so we expected the network to perform the best on White identities. However, the performance curve for White identities was the poorest overall and best for East Asian identities, which consisted of 1.6% of the dataset. This could be due to a multitude of factors. The first is that there simply was not enough data per demographic subgroup. There are not many publicly available datasets that have around 20 facial images per identity and are grouped in this manner. On top of that, there are not many publicly available datasets which are diverse and have these qualities and are diverse.

Another reason for why the network performed in this order for the different races might be the contrast of hair color against skin tone. For both Black and East Asian folks, the color of their hair is much darker than their skin tone, which adds an extra high contrast facial feature to look for at high levels of blur. On the other hand, for White identities, their hair color is a bit lighter and can sometimes get blurred away. An example of the three women of each race at 40px blur is shown in Figure 4-14, and one can see how prominently eyebrows stand out against skin tone. This idea might also explain why men in general also have a higher accuracy score as their eyebrows tend to be thicker and more prominent than eyebrows of women, allowing the facial feature to survive blurring. This idea that eyebrows play an integral role in recognition is a similar phenomenon among humans, as mentioned in Section 2.2.2. The next step in this experiment would be to create this dataset of faces without eyebrows and test the network's performance on the dataset and see how each race and gender performs accordingly. From there, it would be insightful to compare those results with those of humans to pinpoint similarities and differences.

Part of the experiment in addition to characterizing the network's performance was simply collecting diverse data. Next time, we would like to collect more diverse identities across the world or expand upon some of the work that has already been

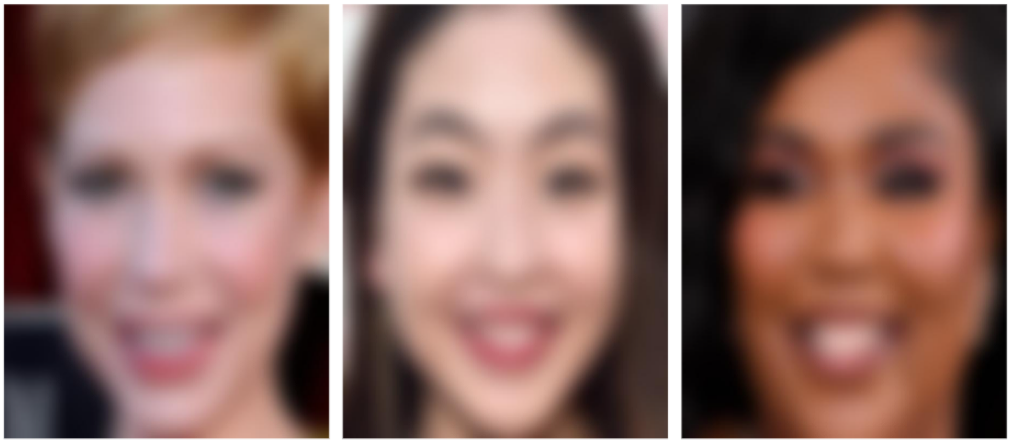

Figure 4-14: Example images of each race at the critical 40px blur.

done. Many datasets generally include Indian identities as well and it would have been interesting to see the recognition performances on this race in addition.

Chapter 5

Conclusion

The project has presented an overview of the transformation tolerance of current state of the art systems and the current demographic biases that exist in these systems today (Chapter 2) as well as tested a state of the art system on common degradations amongst different races (Chapter 3 and 4). This project aimed to link two similar but independent fields, neuroscience and artificial intelligence, to draw stronger conclusions about each other and about face recognition technology.

In our first experiment, our main results are that natural human tones are important to the ResNet-101 model's ability to recognize and group human faces together at high levels of blur, which resembles the features necessary for humans to recognize faces. In our second experiment, we found a demographic bias leaning towards East Asian identities and against White identities at high levels of blur, which sways against popular demographic bias research today.

These results have implications on current widely available networks; simple degradations of hue and blur have the ability to destroy a network's ability to recognize faces at similar accuracy levels as humans. Numerous industries already are relying on facial recognition, from security to criminal justice, even though there are simple ways to destroy the credibility of these systems. Lastly, if the ResNet-101 model is in fact analogous to a human's ability to recognize faces, further work with humans and different hue shifts should be performed to understand how color truly plays a role in human face recognition.

Appendix A

Human vs Network's Accuracy Performance

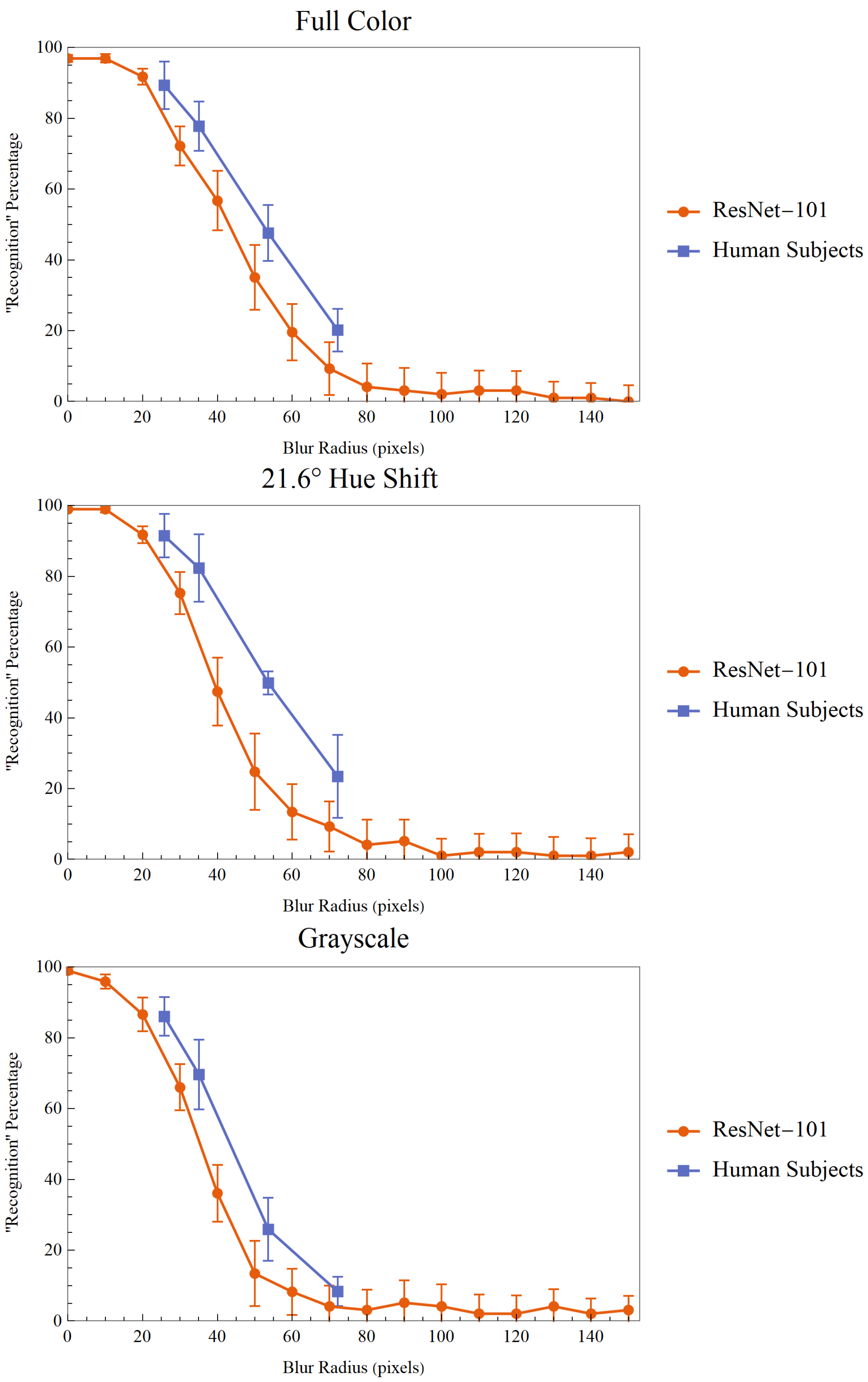

Figure A-1: ResNet-101 vs Human results for full color, 21.6° hue shifted and grayscale images Gaussian blurred at different levels.

Appendix B

Network's Accuracy Performance Comparing Race and Gender

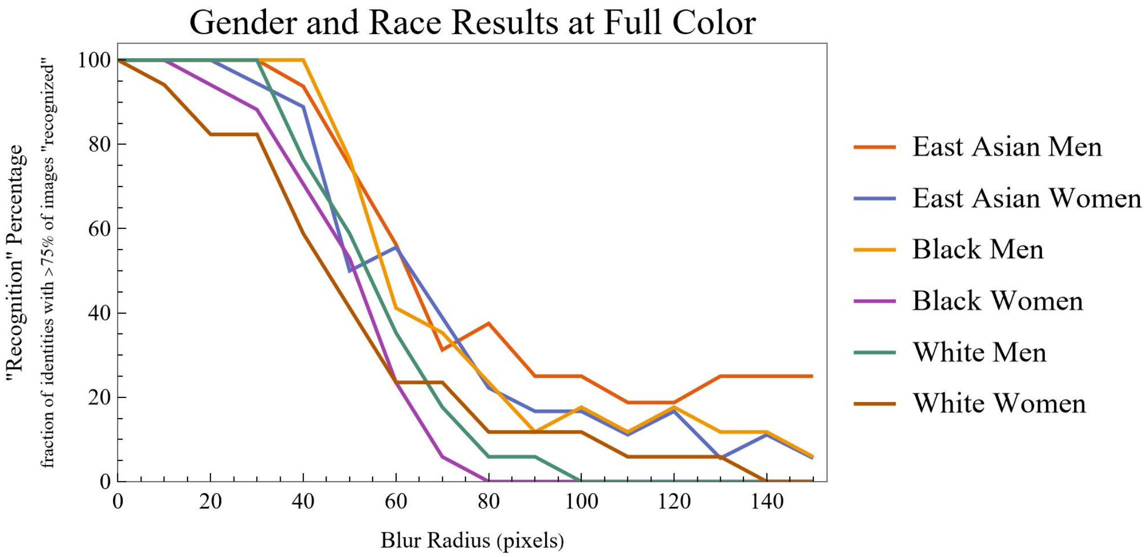

Figure B-1: ResNet-101's performance for all races and genders at full color.

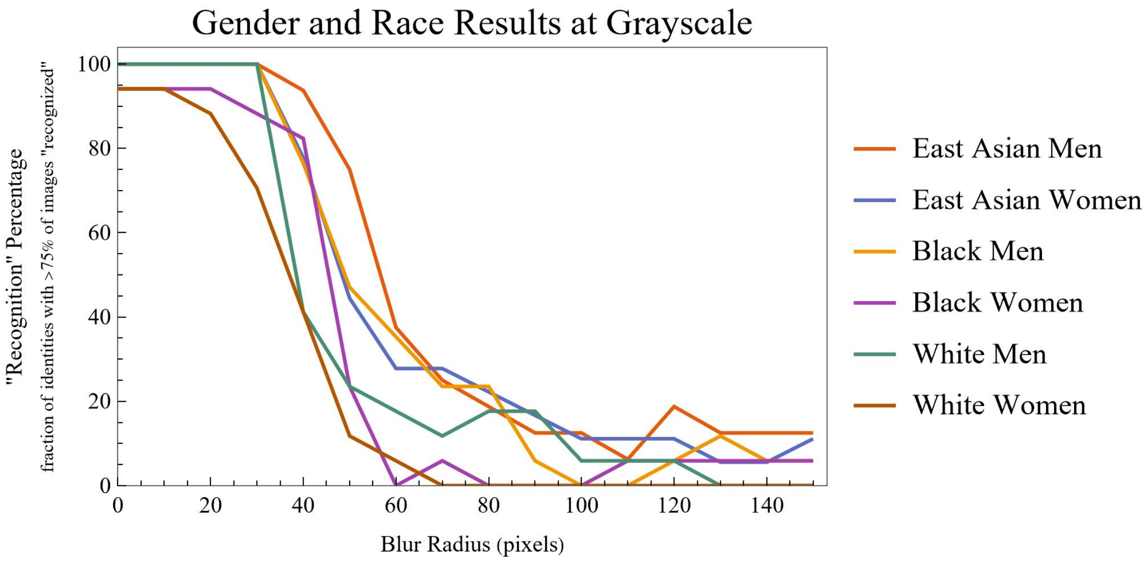

Figure B-2: ResNet-101's performance for all races and genders at grayscale.

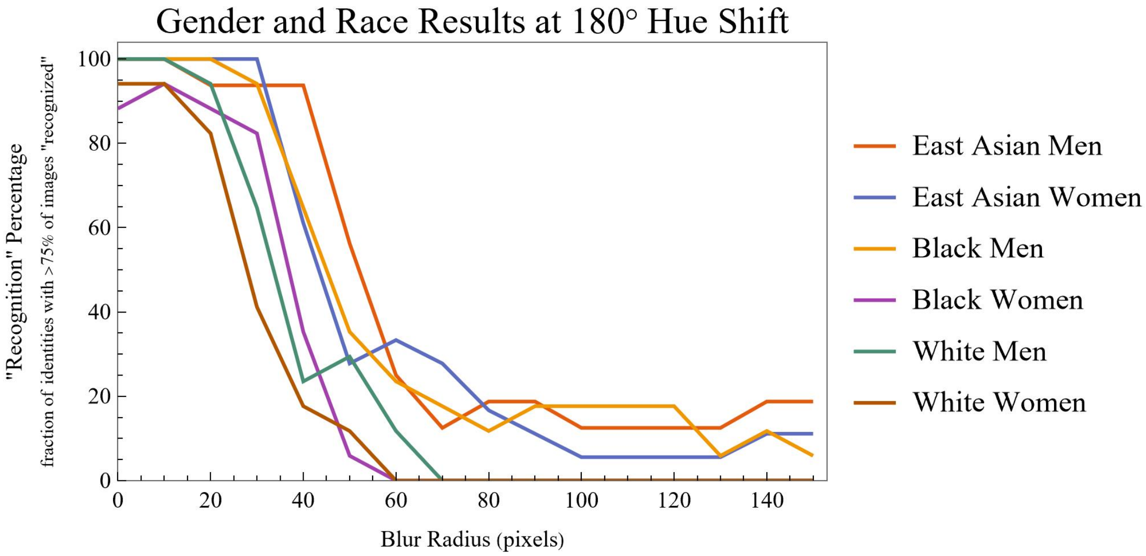

Figure B-3: ResNet-101's performance for all races and genders at a 180° hue shift.

Appendix C

Network's Accuracy Performance Comparing Different Degradations Per Race

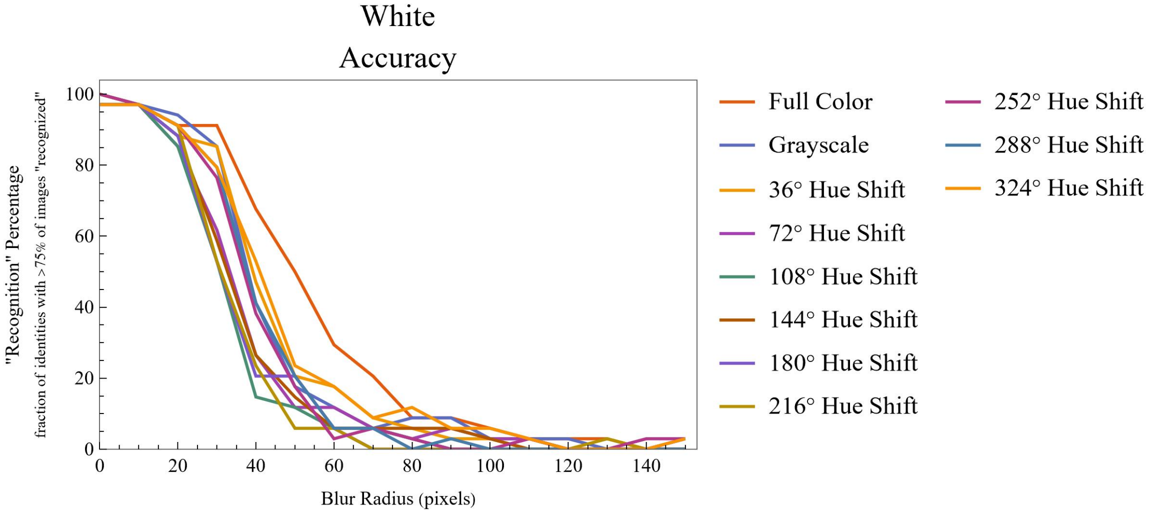

Figure C-1: ResNet-101's performance for all of the degradations for White identities.

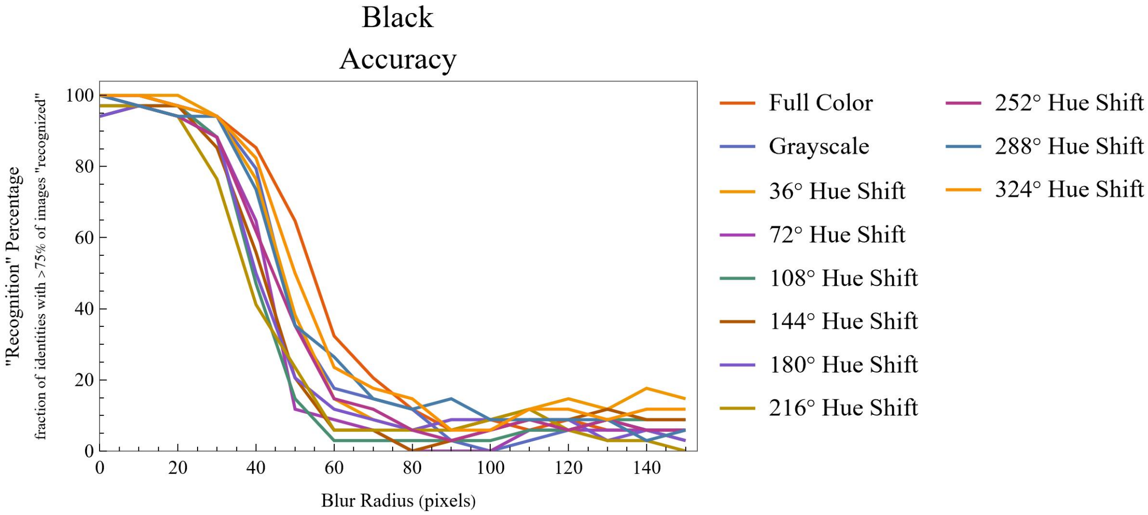

Figure C-2: ResNet-101's performance for all of the degradations for Black identities.

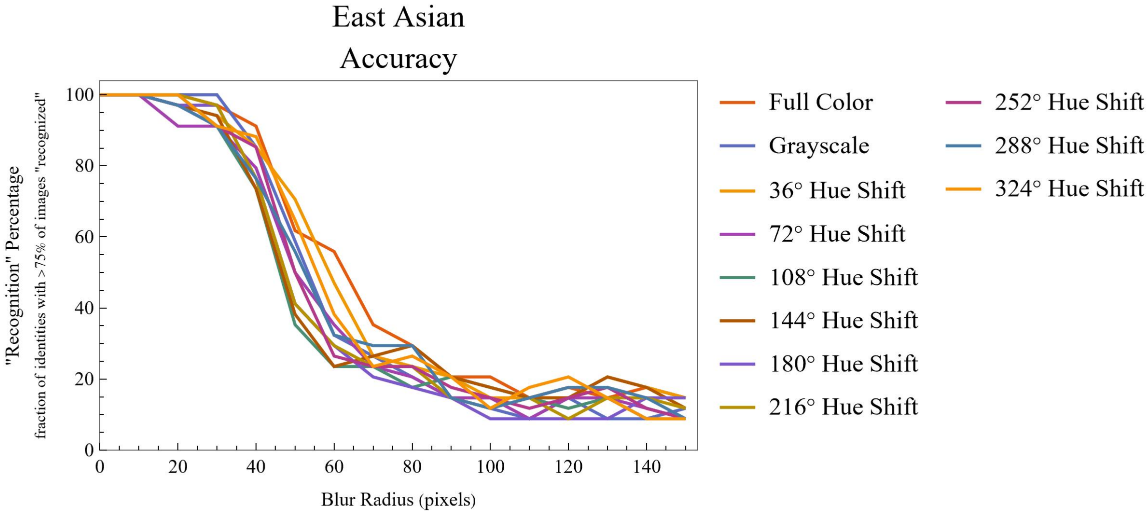

Figure C-3: ResNet-101's performance for all of the degradations for East Asian identities.

Bibliography

[1] Amazon Rekognition. https://aws.amazon.com/rekognition/. Accessed: 2021.

[2] Flickr demographics. https://www.similarweb.com/website/flickr.com/demographics. Accessed: 2022.

[3] Google Cloud Vision API. https://cloud.google.com/vision. Accessed: 2021.

[4] Microsoft Face Service. https://azure.microsoft.com/en-us/services/cognitive-services/face/. Accessed: 2021.

[5] Facial recognition is everywhere. here's what we can do about it., Jul 2020.

[6] Mazida A. Ahmed, Ridip Dev Choudhury, and Kishore Kashyap. Race estimation with deep networks. *Journal of King Saud University - Computer and Information Sciences*, 34(7):4579–4591, 2022.

[7] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[10] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2017.

[11] 2014 6:10 pm UTC Cyrus Farivar Jun 9. First chicago robber caught via facial recognition gets 22 years, Jun 2014.

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, 2016.

[13] Mohammad Haghighat and Mohamed Abdel-Mottaleb. Low resolution face recognition in surveillance systems using discriminant correlation analysis. pages 912–917, 05 2017.

[14] Drew Harwell. Fbi, ice find state driver's license photos are a gold mine for facial-recognition searches, Aug 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[16] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, October 2008. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie.

[17] Isabelle Hupont and Carles Fernández Tena. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7, 2019.

[18] Jamal Hussain Shah, Muhammad Sharif, Mudassar Raza, Marryam Murtaza, and Saeed-Ur-Rehman. Robust face recognition technique under varying illumination. *Journal of Applied Research and Technology*, 13(1):97–105, Feb 2015.

[19] Wolfram Research, Inc. Mathematica, Version 12.2. Champaign, IL, 2020.

[20] Christopher Kanan and Garrison W. Cottrell. Color-to-grayscale: Does the method matter in image recognition? *PLOS ONE*, 7(1):1–7, 01 2012.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2017.

[23] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

[24] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[25] G. Little, S. Krishna, J. Black, and S. Panchanathan. A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose angle and illumination angle. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii/89–ii/92 Vol. 2, 2005.

[26] Chengjun Liu and Harry Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11(4):467–476, April 2002. Funding Information: Manuscript received April 12, 2001; revised November 26, 2001. C. Liu was supported in part by the New Jersey Institute of Technology under SBR Grant 421270. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christine Guillemot. C. Liu is with the Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: liu@cs.njit.edu). H. Wechsler is with the Department of Computer Science, George Mason University, Fairfax, VA 22030 USA (e-mail: wechsler@cs.gmu.edu). Publisher Item Identifier S 1057-7149(02)03548-0.

[27] Iacopo Masi, Anh Tu an Trãn, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision (ECCV)*, October 2016.

[28] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. Diversity in faces, 2019.

[29] Parmy Olson. Faces are the next target for fraudsters, Jul 2021.

[30] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[31] Skand Vishwanath Peri and Abhinav Dhall. Disguisenet : A contrastive approach for disguised face verification in the wild, 2018.

[32] P. Jonathon Phillips. A cross benchmark assessment of a deep convolutional neural network for face recognition. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 705–710, 2017.

[33] P. Jonathon Phillips, J. Ross Beveridge, Bruce A. Draper, Geof Givens, Alice J. O'Toole, David Bolme, Joseph Dunlop, Yui Man Lui, Hassan Sahibzada, and Samuel Weimer. The good, the bad, and the ugly face challenge problem. *Image Vision Comput.*, 30(3):177–185, mar 2012.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[35] Ryan J Reilly. The fbi hasn't arrested hundreds who joined the capitol mob on jan. 6. just ask this maga comedian., Jul 2022.

[36] Javid Sadr, Izzat Jarudi, and Pawan Sinha. The role of eyebrows in face recognition. *Perception*, 32:285–93, 02 2003.

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[39] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[40] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 1988–1996, Cambridge, MA, USA, 2014. MIT Press.

[41] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, page 1701–1708, USA, 2014. IEEE Computer Society.

[42] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.

[43] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, Mar 2021.

[44] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 692–702, 2019.

[45] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch, 2014.

[46] Andrew W Yip and Pawan Sinha. Contribution of color to face recognition. *Perception*, 31(8):995–1003, 2002. PMID: 12269592.

[47] Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Global-local gcn: Large-scale label noise cleansing for face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7728–7737, 2020.