Transformation Tolerance of Machine-based Face Recognition Systems

Ashika Verma[1,2,3], Kyle Keane[1,2], Alyssa Unell[1,3], Anna Musser[1], and Pawan Sinha[1]

[1]**MIT Brain and Cognitive Sciences Department, Sinha Lab**
[2]**MIT Schwarzman College of Computing, Quest for Intelligence**
[3]**MIT Department of Computer Science and Engineering**

ABSTRACT

Face recognition is widely acknowledged to be a very complex visual task for both humans and computers. Previous studies which analyze human robustness of facial recognition have revealed that the human ability to recognize faces becomes worse as the blur levels of face images increases, and that color is important for facial recognition at high blur levels. 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 direct comparison to previous human-subject results. We created a full-color, 21.6° hue shifted, 180° hue shifted and grayscale datasets and Gaussian blurred each dataset at different intensities and compared how AI systems perform relative to humans. We found that the pre-trained ResNet network performed similar to humans: as blur increases, the performance of the network declines under all color degradations, and that naturalistic color is important in for facial recognition at higher levels of blur.

Keywords: facial recognition, computer vision

INTRODUCTION AND RELATED WORK

From unlocking personal devices to criminal justice applications, facial recognition is a commonly used biometric authentication methods in the world today. As such, vulnerabilities within facial recognition systems have real world implications. Facial recognition systems have been used to establish probable cause for arrests, such as in cases involving identity theft [10], passport fraud [11], and assault whereby the actions of assailants were filmed and uploaded to YouTube [12]. The implementation of facial 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 facial recognition systems [7]. As such, it can be seen that facial recognition, when accurate, can be a useful tool for increasing the reach of our judicial system. Given the ubiquity and power of facial recognition systems, it is paramount that such systems be able to perform robustly under real world conditions 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.

Since the introduction of AlexNet in 2012 [9], the number of deep learning methods that recognize faces has exploded, with networks such as ResNet [4], VGGNet [17], and FaceNet [16] serving as the baseline for high accuracy recognition networks [20]. Several companies with substantial investment in AI technologies, such as Google, Facebook and Baidu, have declared success on the profoundly important task of facial recognition, with Facebook developing a face classifier (DeepFace [19]) with an accuracy of 97.35% on the famous benchmark dataset ”Labeled Faces in the Wild (LFW)” ([19][5]). 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 [2], and Microsoft Face Service [3].

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

Figure 1. Example of low quality surveillance footage.

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. 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 originate from an uncontrolled environment, the accuracy of facial recognition systems decrease [7]. 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 [10]. The impact of adversarial attacks and physical facial disguises (such as wigs or makeup) on pre-trained networks has been seen to decrease accuracy of facial recognition in pre-trained networks [18]. Other work has focused on the role of illumination and lighting on the accuracy of the facial recognition models [12] [6]. Some work has been done regarding the impact of color cues on recognition of faces by humans ([21]) and, similarly, the impact of related cues on the recognition of objects by pre-trained networks [8]. No work has been performed on the union of these questions, diving into 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.

To help characterize current vulnerabilities within facial recognition systems, we devised an experiment based off of a cognitive science study tested the impact that naturalistic image degradations have on humans. 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 over different 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 [21]. 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 [21]. 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.

Our experiment is designed to explore the effects of these color shifts on neural networks' ability to identify faces. In the same way that color appears to play a role in face segmentation for humans, we investigated whether color plays a similar role in how a face is encoded within a neural network. To accomplish this end, images from the CelebAMask-HQ dataset ("full color" images) were altered to produce a set of "grayscale" images and "hue shifted" images (21.6°) to mirror the degradations that have been explored with human recognition ability, as described earlier. We also created an additional "hue shifted" set of images with a 180° hue shift which was not used in the human study. These three sets of images and the original set of color images were used to create testing sets with different levels of blur. We then used a pre-trained facial recognition neural network to create vector embeddings which map faces from the dataset to a compact Euclidean space. We evaluated the distances between the points in the space to quantify the cluster quality of images that correspond to each unique individual in the dataset From there we compared the performance of the network on each of the datasets and compared them with each other and with human results.

This work represents an initial attempt to directly link the rich field of human psycho-physics with the cutting-edge of computer vision. 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, like recognizing faces.

METHODS

Overview

Our goal was to evaluate the performance and robustness of a current state-of-the-art facial recognition architecture as it "recognized" faces in a dataset which had undergone naturalistic image degradations. More specifically, we used the ResNet-101 architecture to encode degraded images as a vector and characterized the vector encodings to reflect the accuracy of the network architecture for these degradations. In this section we describe the technical details of each stage of the analysis process.

Investigated Facial Recognition System

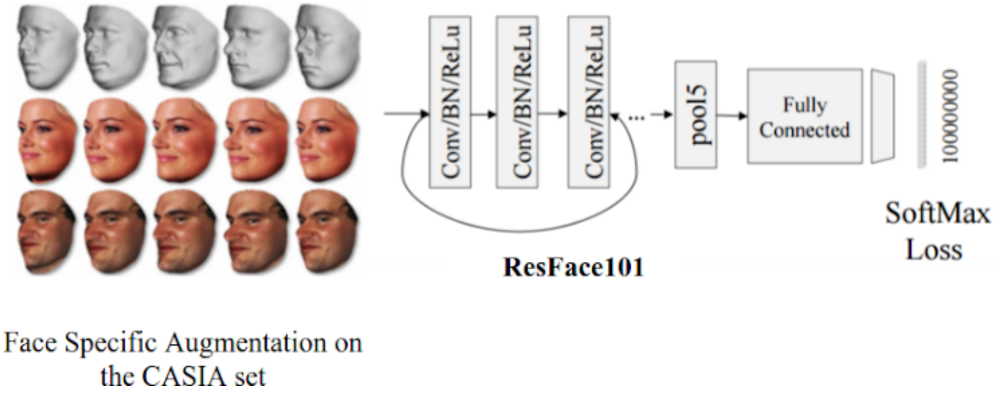

Figure 2. Summary graphic of ResNet-101 Trained on Augmented CASIA-WebFace Data.

The architecture of the model we evaluated throughout the experiment was the ResNet-101 as shown in Figure 2. It 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 model was trained on the Augmented CASIA-WebFace dataset [14], which is a collection of 494,414 facial photographs of 10,575 subjects. Additionally, a far greater per-subject appearance was achieved by synthesizing pose, shape and expression variations from each single image. This model has a 98.06% accuracy on LFW and a 100% Equal Error Rate. The network is available for download on the Wolfram Neural Net Repository [7].

Dataset

We evaluated the ResNet-101 architecture with the CelebAMask-HQ dataset [11]. The original dataset is a large-scale image dataset which has 30,000 high-resolution face images selected from the CelebA and then CelebA-HQ datasets [13]. We chose this dataset due to its high image per identity rate, which was essential for us to test recognition ability. Additionally, each of the masks of CelebAMask-HQ were manually-annotated with important facial features such as eyes and ears, which was used in later calculations.

The aim of the experiment was to analyze how well this trained model grouped the vector encodings of images by identity, thus celebrities with fewer than 19 unique photos were removed from the dataset. We also removed all photos that were not clearly of the individual's face (i.e. photos where the face was obscured by hair, turned away, etc.). This resulted in 97 unique identities for further analysis. We called this final down-sampled set of high-resolution color images $D_{\text{color}}(0)$.

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 3. Hue is one of the main properties of color and is represented quantitatively by an angle on a color wheel. The 21.6° hue shift changes images to have a yellow tint which still retains lots of natural human skin colors while the 180° hue shift changes 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 [21] 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.

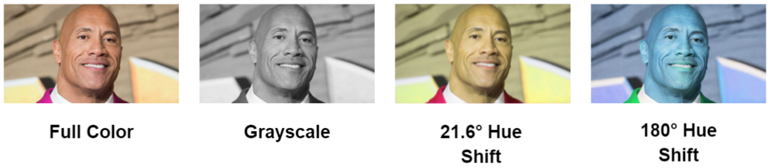

Figure 3. Dwayne 'The Rock' Johnson in full color, grayscale, 21.6° hue shift, and 180° hue shift conditions.

Recognition Percentage

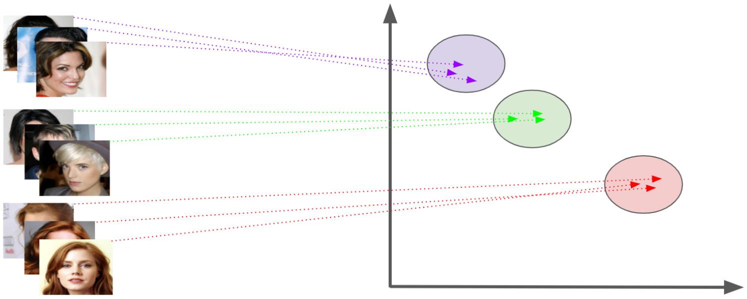

Figure 4. Identities Encoded to 2D Vector Space Generalization

The images were passed into the ResNet101 encoder described earlier, 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 4 to help the reader visualize the encoding 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

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

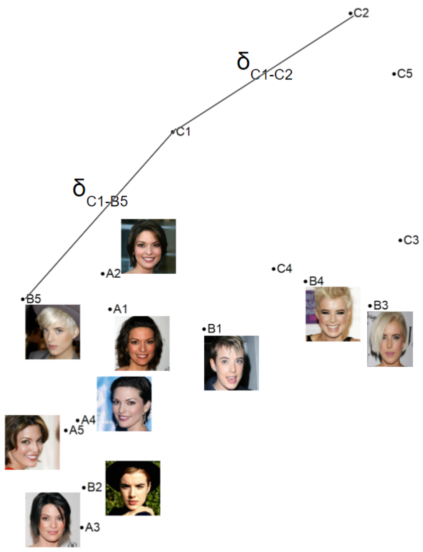

Figure 5. Individual Image Classification: In this example, we calculate the distance between $C1$ and $B5$ and $C1$ and $C2$. $C1$'s identity classification would be C, $A1$'s would be B, $B1$ would be B and so on.

individuals [Fig. 4]. 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 measure of 75% arises from research standards from brain and cognitive science, on which much of this research is based [21].

Evaluating the Network

To finish the evaluation of the network on a dataset, we found the percentage of identities which were correctly “recognized”. From here we calculated how well the network performed over different blur degradations to generalize the robustness of the network with respect to blur.

RESULTS

The Network’s Performance on Different Levels of Blur

Figure 6 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 7 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.

0 to 35 pixel radius blur

Figure 8 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

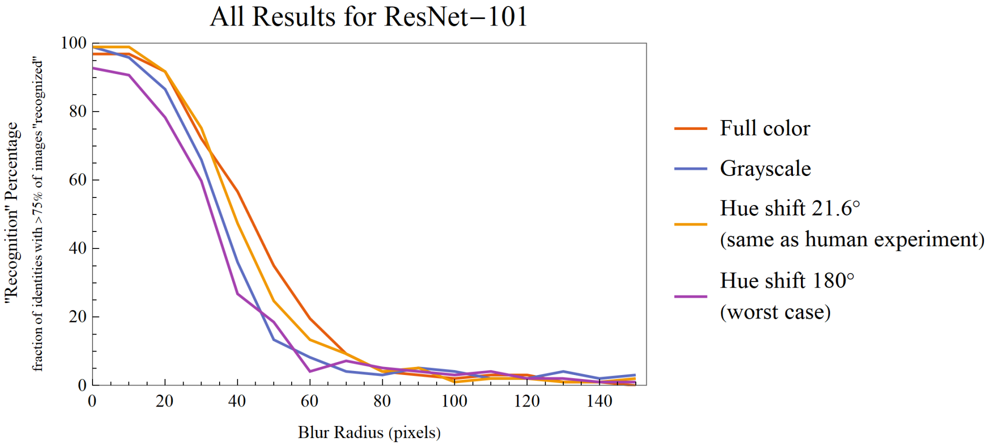

Figure 6. Line graph of all results for ResNet-101 comparing recognition percentage vs blur radius.

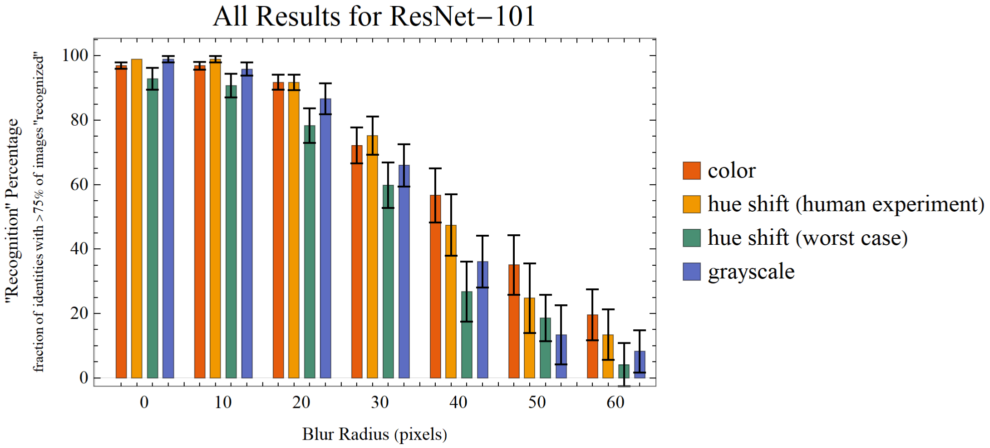

Figure 7. Bar chart of all results for ResNet-101 comparing recognition percentage vs blur radius.

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 9 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.

81 and beyond pixel blur radius

Figure 10 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%.

Comparing Humans and ResNet-101

Next, we aimed to compare our network results to the human results from Sinha 2002 [21].

Converting Human Results to Gaussian Blur

The first step in standardizing our results with those from Sinha 2002 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 the human experiment, the researchers scaled down the

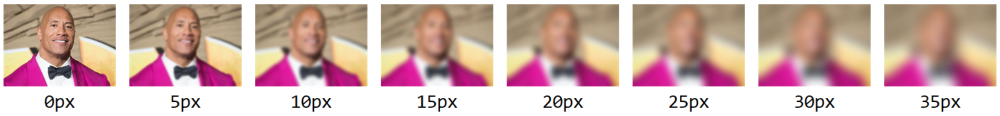

Figure 8. Dwayne ”The Rock” Johnson blurred from 0 pixels to 35 pixels with increasing intervals of 5 pixels.

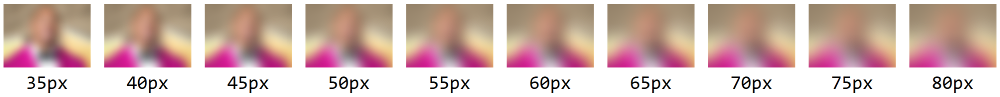

Figure 9. Dwayne ”The Rock” Johnson blurred from 35 pixels to 80 pixels with increasing intervals of 5 pixels.

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 11. This follows the same methodology as in Sinha 2002 [21] and allows us to directly compare the blur from the human study to our results.

Comparison between Humans and ResNet-101

In Figure 12, 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 12. 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.

Discussion

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 7. 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 [21]. 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.

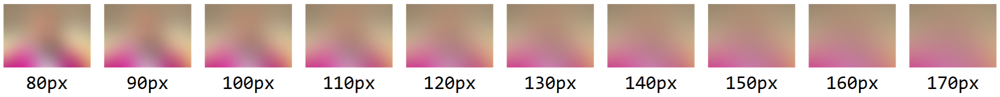

Figure 10. Dwayne ”The Rock” Johnson blurred from 80 pixels to 170 pixels with increasing intervals of 10 pixels.

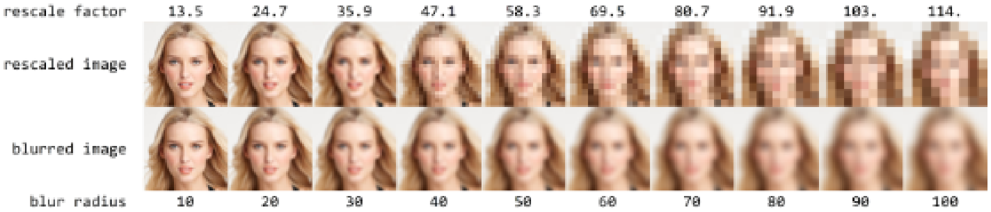

Figure 11. Re-scaled images vs corresponding image distance minimizing radius Gaussian blur.

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 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 [16] and CLIP [15]. Finally, similar work with different degradations (such as line drawings of human faces) would be worth studying with a similar methodology as the one defined in this paper.

CONCLUSION

In summary, 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. 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.

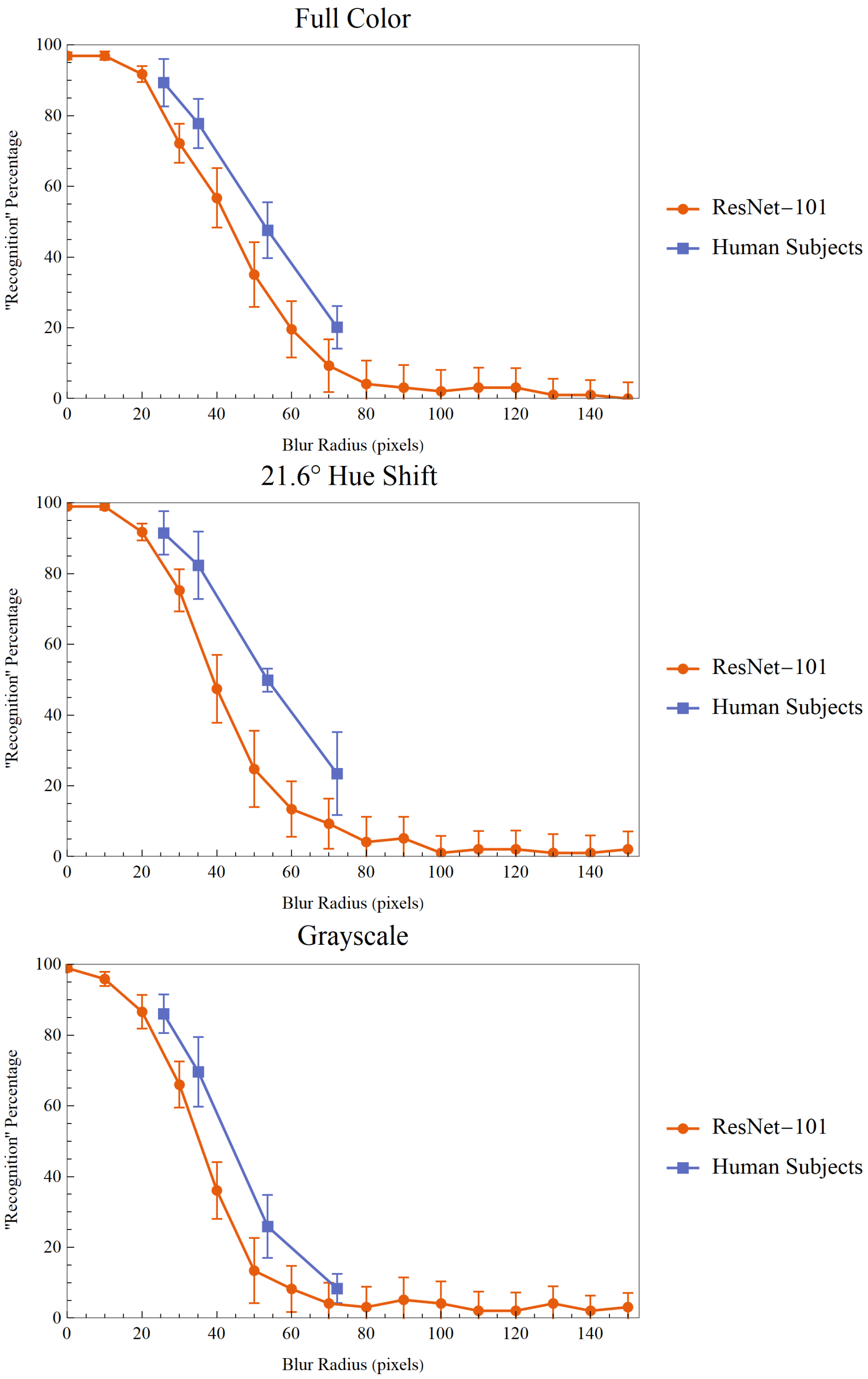

Figure 12. ResNet-101 vs Human results for full color, 21.6° hue shifted and grayscale images Gaussian blurred at different levels.

REFERENCES

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

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

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

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

[5] 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.

[6] 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.

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

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

[9] 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.

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

[11] 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.

[12] 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.

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[14] 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.

[15] 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.

[16] 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.

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

[18] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. On the robustness of face recognition algorithms against attacks and bias, 2020.

[19] 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.

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

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