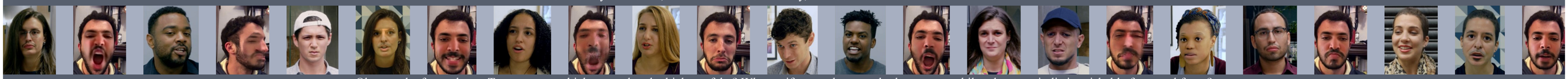
The Uncanniness of Face Swaps

Ethan Wilson[1] (ethanwilson@ufl.edu), Aidan Persaud[1], Nicholas Esposito[1], Sophie Jörg[2], Rohit Patra[1], Jenny Skytta[3], Frederick Shic[3], Eakta Jain[1]

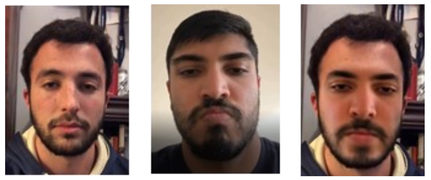
[1]University of Florida, [2]Clemson University, [3]Seattle Children's Research Institute



Observe the faces above. Try to guess: which are real and which are fake? What artifacts make some look uncanny while others are indistinguishable from real faces?

Introduction

- Face swapping can paste an **actor's** face onto the **original** subject in a video, creating very realistic fake media.
- We aim to understand how different dataset properties influence the perception of face swaps by measuring through the lens of uncanniness.



(a) (b) (c)

Figure 1: (a) **Original** subject; (b) **Actor's** face; (c) Face swapped result.

Background

- Emerging work on the perception of face swaps has investigated perceived artifacts [1] and sincerity of emotions [2].
- The perception of computer generated faces and digital avatars have previously been investigated using the concept of the uncanny valley effect [3].
- The uncanny valley effect is the hypothesis claiming that as objects/robots/characters approach a human likeness, viewer appeal increases. But, there is a "valley" present near human realism where objects appear uncanny/eerie/revolting.

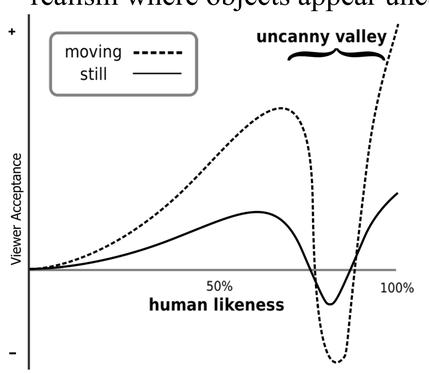


Figure 2: A visualization of the uncanny valley effect.

We follow a similar approach to Ho and MacDorman [3].

We incorporate established metrics for measuring the uncanny valley effect of avatars to evaluate the visual quality and viewer acceptance of face swaps.

Design

We designed a perceptual experiment that measures uncanniness through five distinct bipolar adjective pairs based on Ho and MacDorman's work around uncanniness of avatars [3].

Experiment 1: Evaluating many faces via a standard benchmark.

Participants (N=39) viewed 40 short clips (20 real, 20 face swapped) taken from the FaceForensics++ DeepFakeDetection dataset [4]. Every video was rated on a 7 point Likert scale for each bipolar adjective pair. This experiment quantified the uncanniness of face swaps relative to real faces.

Experiment 2: Viewer responses to controlled manipulations.

Face swap stimuli were generated using DeepFaceLab [5], the most popular open-source deepfake tool, on two facial video recordings. Stimuli corresponded to distinct dataset degradations. Participants (N=28) viewed all stimuli and rated them on the same uncanniness metrics as before. This experiment identified multiple data manipulations that can be found in real world scenarios and assessed their potential impact on uncanniness.

	FaceForensics++		Our Data	
Adjective Pairs	**Original**	**Face Swap**	**Original**	**HQ Face Swap**
Real/ Synthetic	2.11 ± 0.99	5.20 ± 0.90	1.56 ± 1.66	4.36 ± 1.58
Agreeable/ Repulsive	2.18 ± 0.95	4.49 ± 0.96	1.60 ± 1.15	3.68 ± 1.31
Unremarkable/ Unusual	2.26 ± 0.90	4.80 ± 0.97	1.76 ± 1.36	4.00 ± 1.41
Plain/ Weird	2.26 ± 0.90	4.88 ± 0.98	1.80 ± 1.38	4.24 ± 1.48
Ordinary/ Uncanny	2.30 ± 1.04	4.74 ± 0.99	1.80 ± 1.61	4.36 ± 1.68
Average	**2.22 ± 0.93**	**4.82 ± 0.91**	**1.70 ± 1.24**	**4.13 ± 1.30**

Table 1: Viewer responses to five bipolar adjective pairs designed to measure uncanniness. On Face Forensics, viewers compared 20 real to 20 manipulated faces. Results on our data compare the unaltered original clip and a high quality (HQ) face swap using the best possible parameters.

Results

(Exp. 1) Face swaps are perceived to be more uncanny than their original counterparts.

(Exp. 2) Multiple data manipulations significantly increase uncanniness, including:

- Deficient facial expressions **(1)**
- Decreased dataset size **(2)**
- Mismatched facial expressions **(3)**
- Decreased character face resolution **(4)**
- Mismatched facial orientations **(5)**

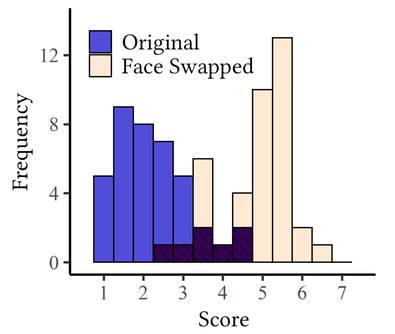


Figure 3: Distribution of responses for FaceForensics++ faces. Higher values represent a higher average uncanniness score.

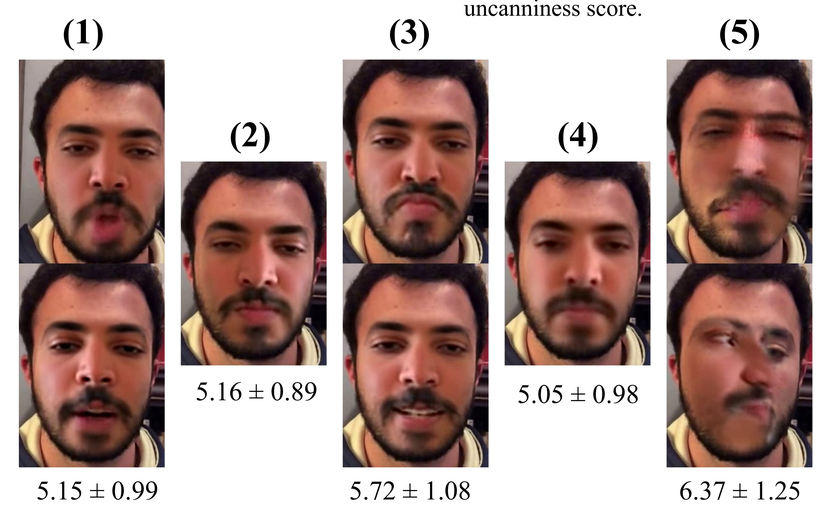


Figure 4: Visual results from Exp. 2 of the same pose shown in Figure 1. Participants viewed all stimuli and rated videos on the five bipolar adjective pairs seen in Table 1. All shown deficiencies were statistically significant when compared to the high quality face swap (4.13 ± 1.30) with Bonferroni correction of $\alpha = 0.0042$.

Conclusion

- Our work identifies some dataset deficiencies that negatively impact face swaps which can easily be encountered in future application spaces.
- Our experiment was a wide-breadth and low-depth initial investigation into the human perception of face swaps.
- As face swapping technology improves, the requirement for a large directory of training data also improves. The technology is now at the point where data constraints are the limiting factor of high fidelity face swapping.

[1] Wöhler, L., Henningson, J. O., Castillo, S., & Magnor, M. (2020, October). PEFS: A validated dataset for perceptual experiments on face swap portrait videos. In International Conference on Computer Animation and Social Agents (pp. 120-127). Springer, Cham.

[2] Wöhler, L., Zembaty, M., Castillo, S., & Magnor, M. (2021, May). Towards understanding perceptual differences between genuine and face-swapped videos. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

[3] Ho, C. C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect. International Journal of Social Robotics, 9(1), 129-139.

[4] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1-11).

[5] Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., ... & Zhang, W. (2020). DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv preprint arXiv:2005.05535.