Article

PERCEPTION

Perception
0(0) 1–15

DOI: 10.1177/0301006617713633
journals.sagepub.com/home/pec
SAGE

Identifying Computer-Generated Portraits: The Importance of Training and Incentives

Brandon Mader
Dartmouth College, Hanover, NH, USA

Martin S. Banks
University of California, Berkeley, CA, USA

Hany Farid
Dartmouth College, Hanover, NH, USA

Abstract
The past two decades have seen remarkable advances in photo-realistic rendering of everything from inanimate objects to landscapes, animals, and humans. We previously showed that despite these tremendous advances, human observers remain fairly good at distinguishing computer-generated from photographic images. Building on these results, we describe a series of follow-up experiments that reveal how to improve observer performance. Of general interest to anyone performing psychophysical studies on Mechanical Turk or similar platforms, we find that observer performance can be significantly improved with the proper incentives.

Keywords
computer graphics, photo-realistic, photo forensics

Introduction

Advances in computer graphics have yielded stunningly photo-realistic images. Of particular challenge is the rendering of portraits of people with photo-realistic skin, hair, and facial features. Previous work, dating from 2007 through 2016, found that human observers' ability to distinguish computer-generated from photographic portraits of people has been slowly declining (Farid & Bravo, 2007, 2012; Fan, Ng, Herberg, Koenig, & Xin, et al., 2012; Holmes, Banks, & Farid, 2016). Previous work has also found that it is difficult to computationally discriminate between computer-generated and photographic images (Conotter et al., 2014; Dang-Nguyen et al., 2012a; Dang-Nguyen et al., 2012b; Dehnie et al., 2006; Dirik et al., 2007; Gallagher & Chen, 2008; Khanna et al., 2008; Lalonde &

Corresponding author:
Hany Farid, Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA.
Email: farid@dartmouth.edu

Efros, 2007; Lyu & Farid, 2005; Ng et al., 2005; Wang & Moulin, 2006; Wen et al., 2007). Building on our earlier work in Holmes et al. (2016), we show that observer accuracy may be better than previously thought. In particular, we show that with proper training, feedback, and incentives, observer performance on distinguishing computer-generated from photographic images can be significantly improved.

The accuracy with which observers can distinguish computer-generated from photographic images has both legal and scientific consequences. On the legal front, the prosecution of child pornography crimes can be complicated if the images in question cannot be reasonably proven to be of real children (the reason for this is that the legal statute classifies computer-generated child pornography differently than real photographic images; *Ashcroft v. Free Speech Coalition*, 2002). On the scientific front, quantitative measures of photo-realism provide valuable information for the advancement of even more photo-realistic rendering.

We begin by briefly describing our previous results (Holmes et al., 2016) that provide a baseline accuracy for observer accuracy in discriminating computer-generated and photographic images. We then describe two new experiments that present the conditions that lead to greater accuracy. And lastly, we describe two additional experiments designed to determine what cues observers use to perform this discrimination task.

Methods

Using techniques as described in Holmes et al. (2016), we collected 30 high-quality computer-generated images and 30 matching photographic images (Figures 1–3). These 60 images constitute the testing images. We collected an additional 10 computer-generated images and 10 matching photographic images (Figure 4). These images constitute the training images.

All computer-generated and photographic images were matched for brightness and contrast (in luminance/chrominance space) to ensure that observers could not classify images based on systematic differences in a low-level image statistic—see Holmes et al. (2016) for details on how this image matching was performed.

We recruited participants from Amazon's Mechanical Turk on-line work force. After a Mechanical Turk user opted to participate in our experiment, they were given a brief summary of the task and were asked to consent to the terms of the experiment.

After an image was presented, the observer was asked to make a judgment as to whether it was computer generated (CG) or photographic and whether the person in the image was female or male. Observers saw one image at a time and the matched CG and photographic images were randomly distributed throughout the block. To ensure that observers did not rush their judgment, the experimental program did not allow an observer to make their selection for 3 seconds after image onset. After this delay, the observer could click a button to indicate their choice: "male/CG," "male/photographic," "female/CG," or "female/photographic." The order in which images were presented was randomized. Each participant was paid for their time. The observer's ability to correctly identify the person's gender in each image was used as a means of discarding the results from observers who responded randomly. A threshold of 95% accuracy on this gender-identification task was set prior to data collection.

Results

In our original study (Holmes et al., 2016), we performed two experiments that are briefly reviewed here. In the first experiment—with no training or feedback—observers classified an image as CG or photographic. Observer performance was assessed as a function of image resolution. The second experiment was similar, except this task was preceded by a short

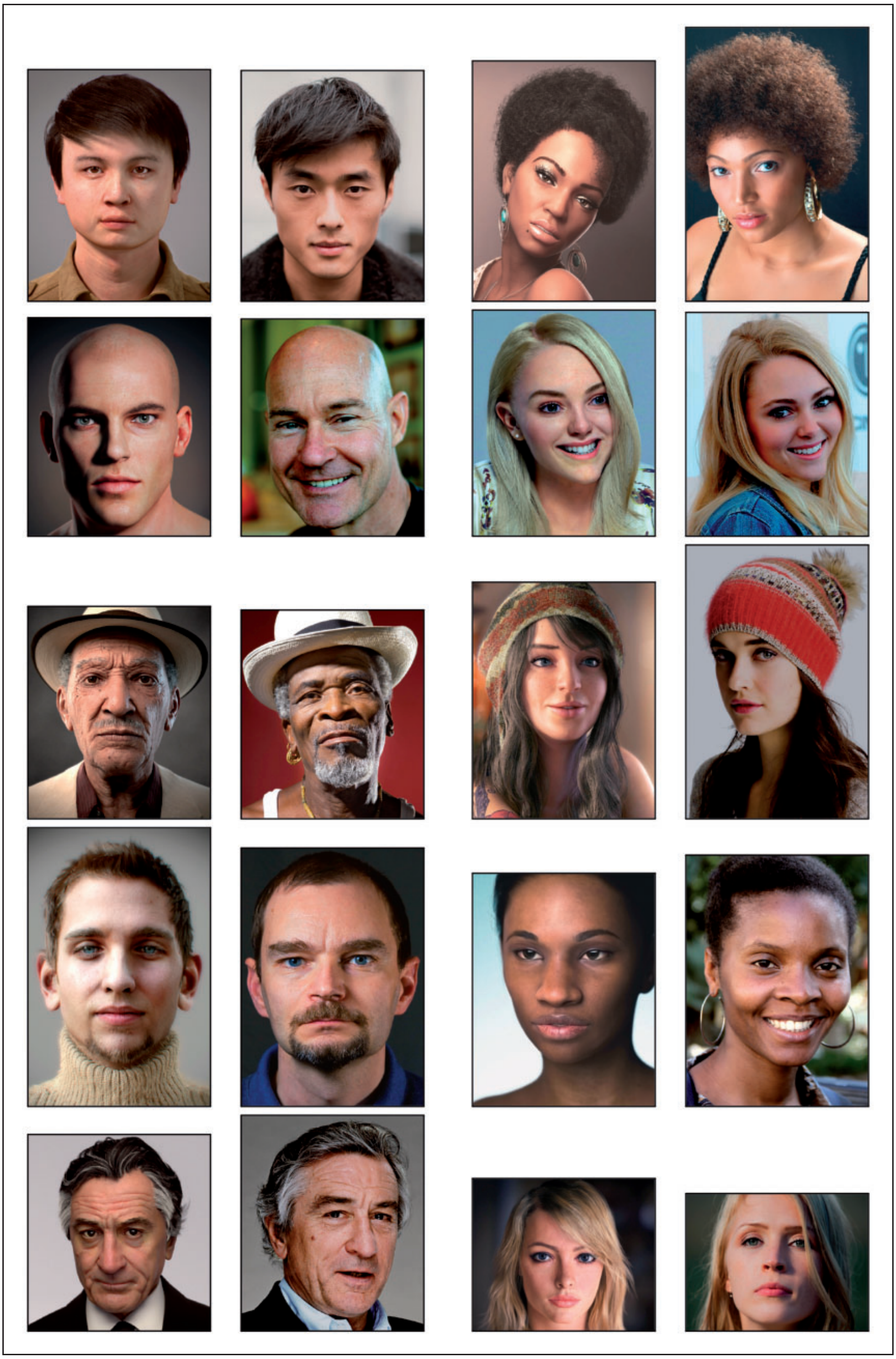

Figure 1. Computer-generated images (first and third columns) paired with their photographic matches (second and fourth columns). See Figures 2 and 3.

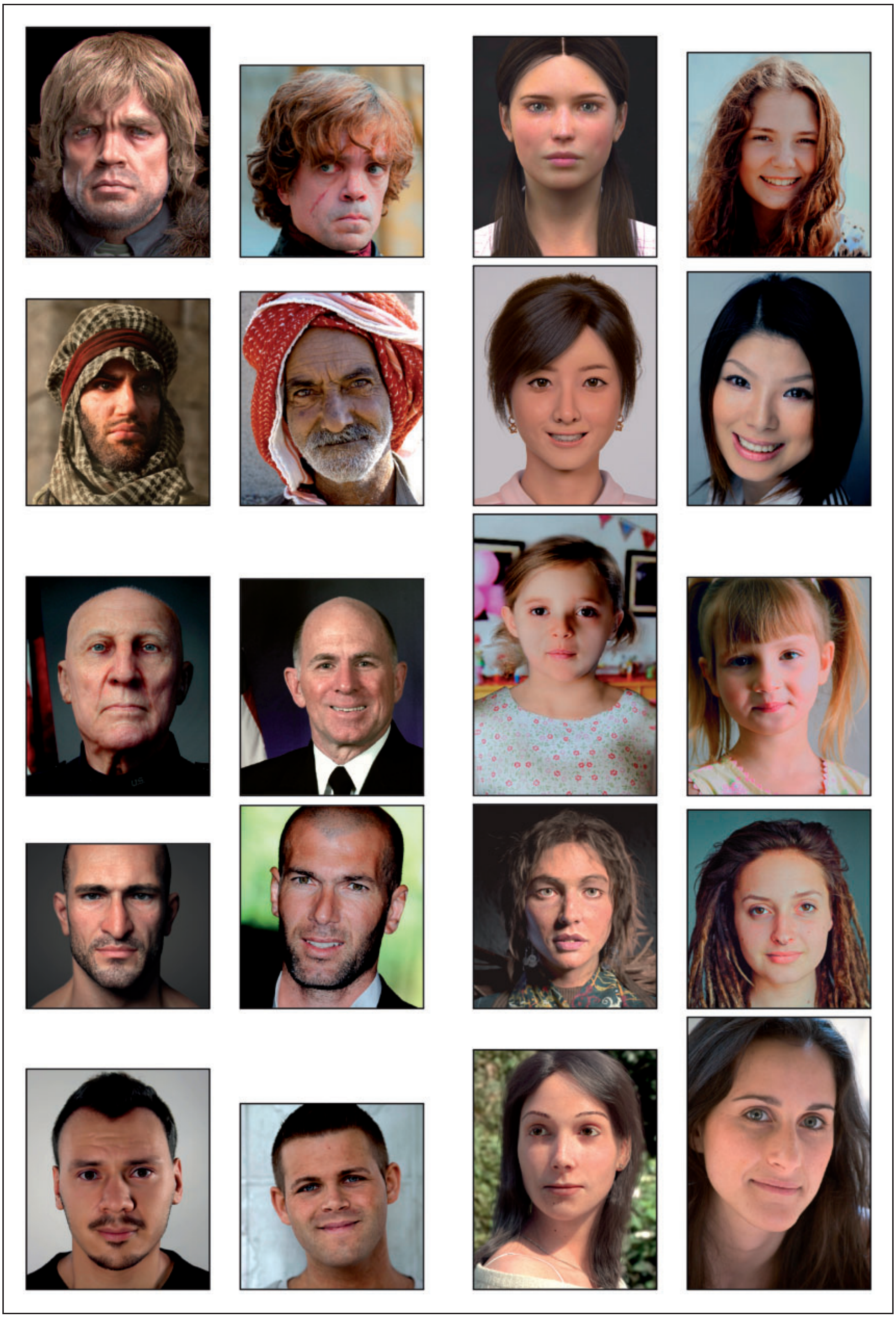

Figure 2. Computer-generated images (first and third columns) paired with their photographic matches (second and fourth columns). See Figures 1 and 3.

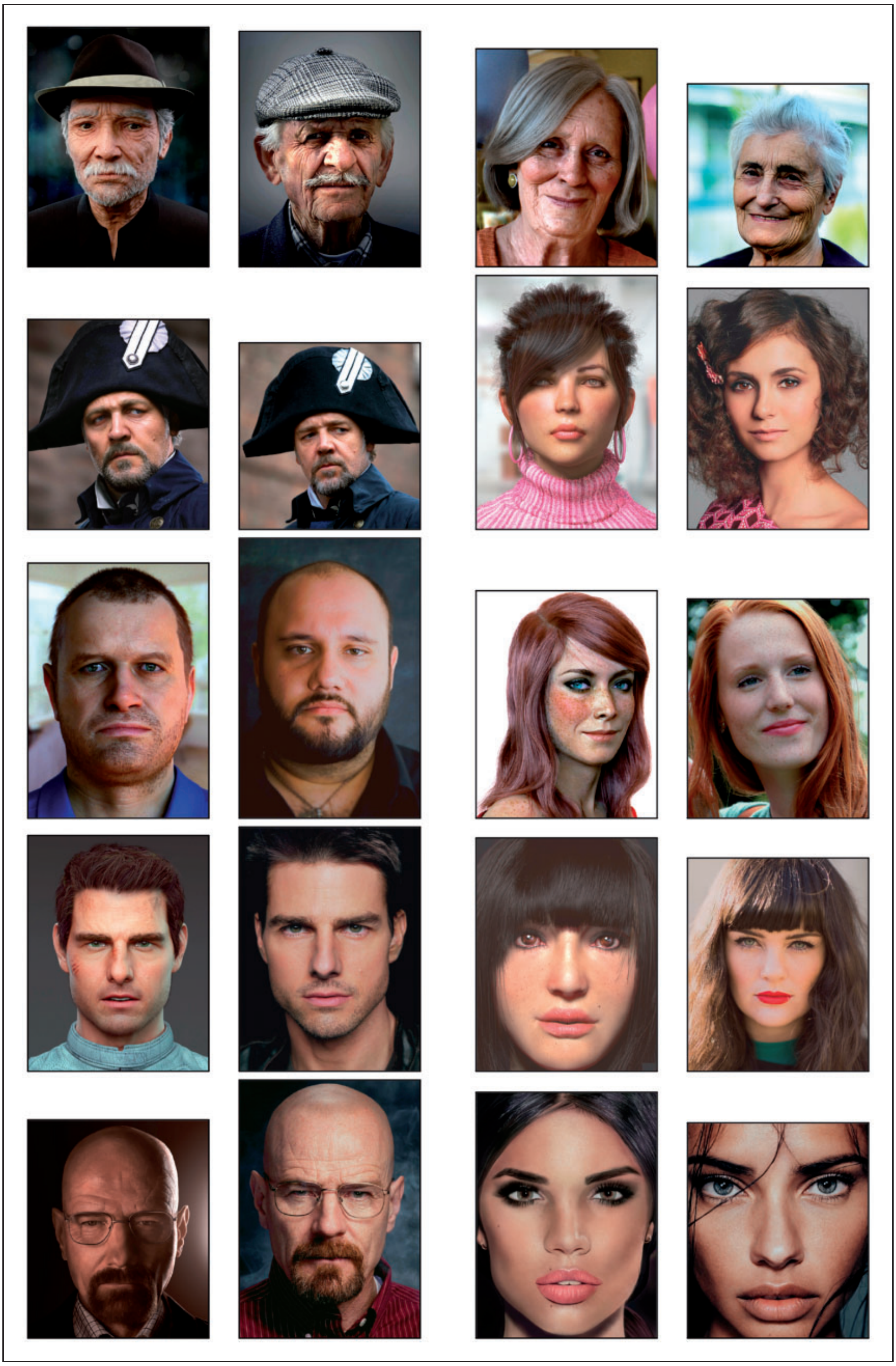

Figure 3. Computer-generated images (first and third columns) paired with their photographic matches (second and fourth columns). See Figures 1 and 2.

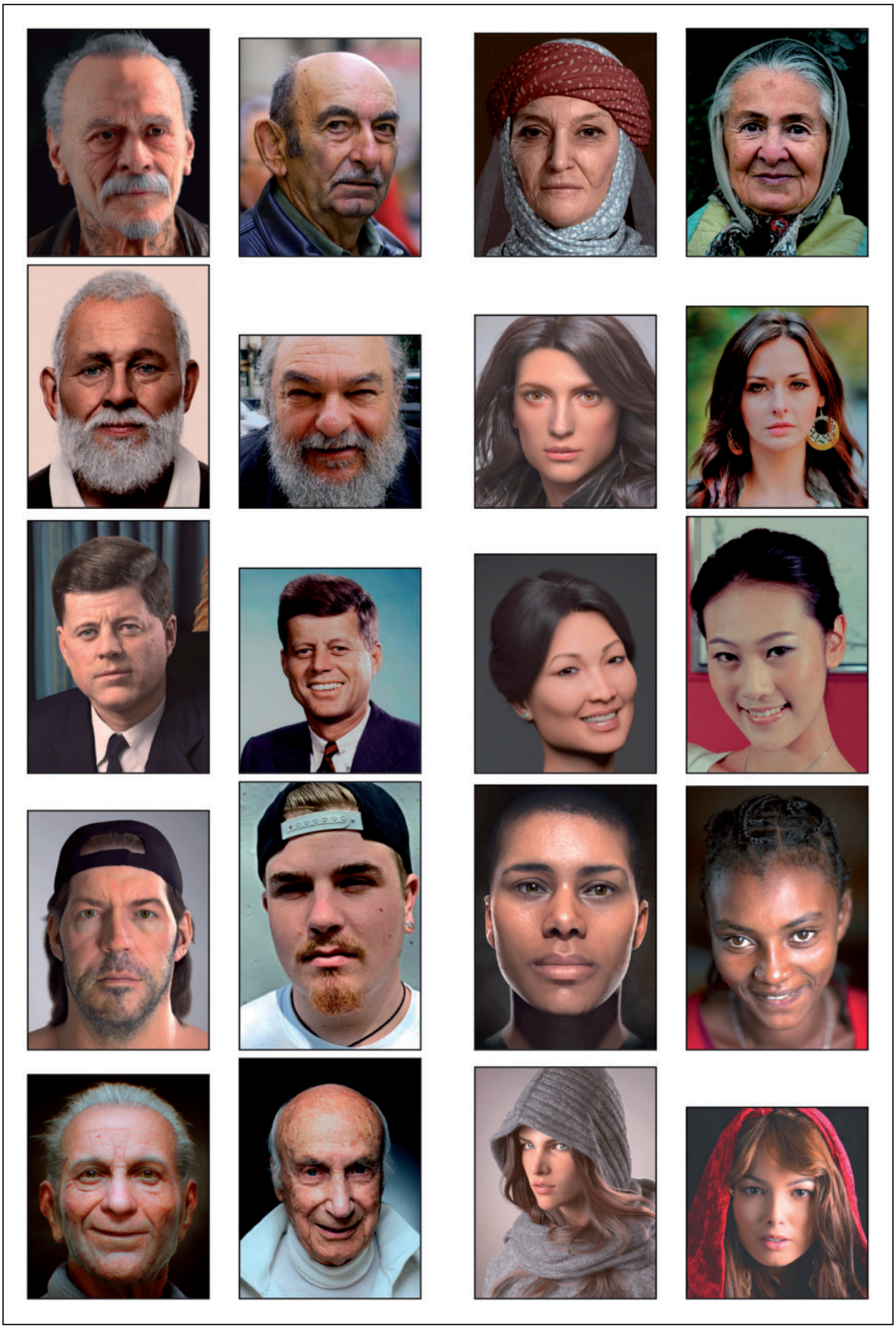

Figure 4. Training computer-generated images (first and third columns) paired with their photographic matches (second and fourth columns).

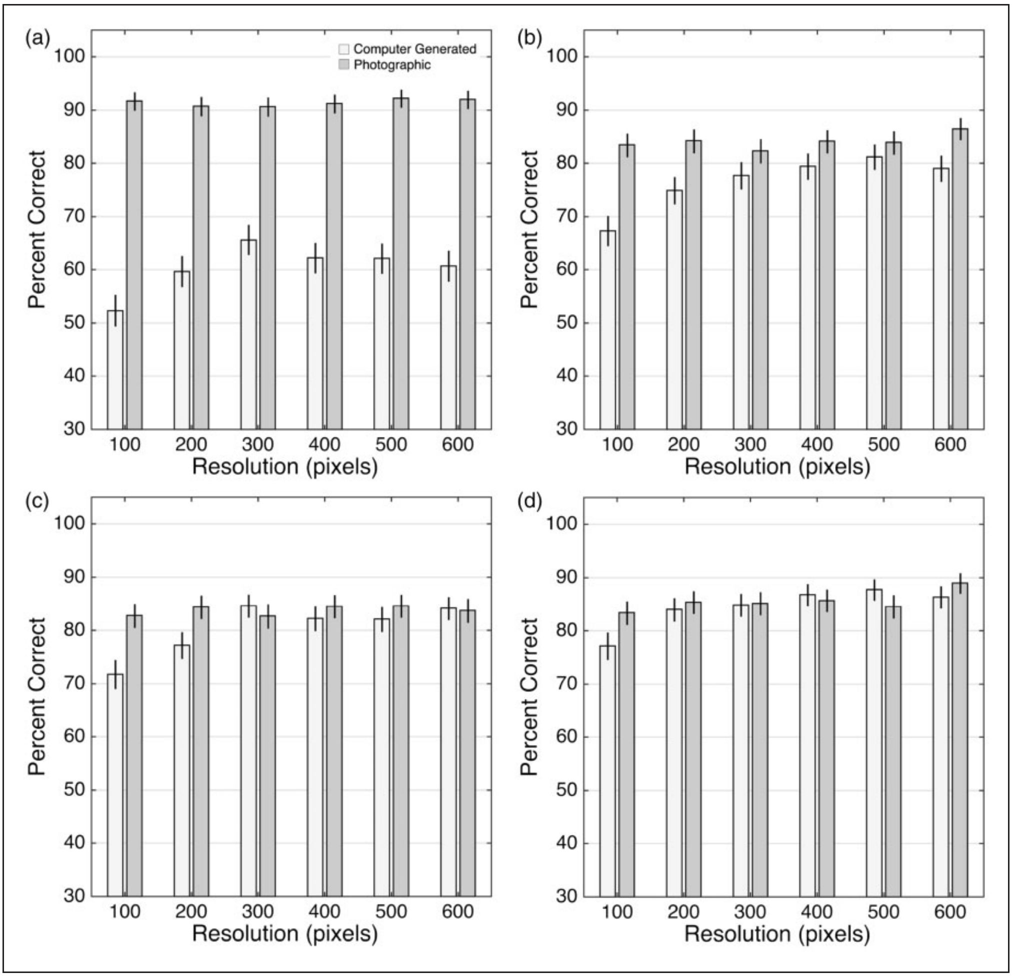

Figure 5. Percent correct for (a) no training or feedback (Holmes et al., 2016); (b) training but no feedback (Holmes et al., 2016); (c) training and feedback; and (d) training, feedback, and incentives. Error bars represent 95% confidence intervals.

training session in which observers were shown 20 examples of computer-generated and photographic images (Figure 4). These 20 distinct training images were selected and processed in the same way as the full set of 60 images. Each image was labeled as photographic or CG and observers were forced to view the image for a minimum of 3 seconds before proceeding to the next training image. Following the training stage, observers saw the same 60 images used in the first experiment. During this testing stage, no feedback was provided. In each experiment, we collected responses from 250 Mechanical Turk participants. Participants were paid $0.50 for their time. Shown in Figure 5(a) and (b) is the average observer accuracy in correctly identifying computer-generated and photographic images as a function of image resolution. Percent correct response is not necessarily the best measure of observer performance because an observer might have a bias to say that most images are photographic (or CG) for all images. In that case, overall percent correct is reduced due to bias. The appropriate way to assess observer performance independent of bias is provided by signal-detection methods as described next.

Table 1. d' and β for (a) no training or feedback (Holmes et al., 2016); (b) training but no feedback (Holmes et al., 2016); (c) training and feedback; (d) training, feedback, and incentives; and (e) masked eyes.

	(a)		(b)		(c)		(d)		(e)	
Resolution	d'	β	d'	β	d'	β	d'	β	d'	β
100	1.43	2.29	1.52	1.75	1.70	1.71	1.85	1.54	1.08	1.38
200	1.59	2.22	1.80	1.71	1.92	1.60	2.18	1.40	1.44	1.24
300	1.78	2.23	1.84	1.54	2.09	1.29	2.19	1.38	1.34	1.23
400	1.72	2.15	2.02	1.53	2.10	1.49	2.27	1.28	1.54	1.31
500	1.75	2.36	2.04	1.45	2.10	1.47	2.23	1.19	1.52	1.32
600	1.65	2.27	2.00	1.62	2.14	1.40	2.35	1.46	1.54	1.45
Average	1.65	2.25	1.87	1.60	2.01	1.49	2.18	1.37	1.41	1.32

Shown in Table 1(a) and (b) are these results expressed as d' and β.[1] d' represents observer sensitivity (independent of bias) and β represents observer bias. In a two-alternative task such as ours, d' is a measure of sensitivity and if there is no bias, $d' = 1$ corresponds to an overall accuracy of 76%. A value of $\beta < 1$ corresponds to a bias to classify an image as CG, $\beta > 1$ corresponds to a bias to classify an image as photographic, and $\beta = 1$ corresponds to no selection bias. In our previous work, we pooled all of the observer data and computed a single d' and β (Holmes et al., 2016). Here, we compute the average individual d' and β, thus allowing us to perform statistical tests for significance.

In summary, we previously showed that observers with no training have a clear bias for classifying an image as photographic (Holmes et al., 2016). Across all six resolutions, the average β was 2.25. Resolution had only a relatively small effect on performance from d' of 1.65 (at the maximum resolution of 600) to d' of 1.43 (at the minimum resolution of 100). A small amount of training was surprisingly effective at reducing the bias and increasing the sensitivity. Across all resolutions, the average β fell from 2.25 to 1.60, and the average d' increased slightly from 1.65 to 1.87.

There are three rationales for testing multiple resolutions in each of these two experiments: (a) provide a direct comparison to our original study (Holmes et al., 2016); (b) in the legal setting, it is common for jurors to be asked to assess the realism of images of varying resolution from small thumbnail-sized to full-resolution images; (c) by looking at how performance changes over resolution, we may be able to infer something about what content observers do or do not rely on when classifying images as CG or photographic.

We hypothesize that because a small amount of training was so effective at reducing bias, most of our observers were simply unaware as to the high quality of modern computer-generated images. One of the goals of the present paper is to further assess this hypothesis.

Feedback

Because a small amount of training had such a dramatic increase in performance, we wondered whether our previous study had fully characterized the photo-realism of computer-generated content. Understanding the limits of human performance is important in the legal setting where it is critical to understand observer sensitivity and bias, and for the computer-graphics community where it is valuable to understand the successes and failures of photo-realistic rendering.

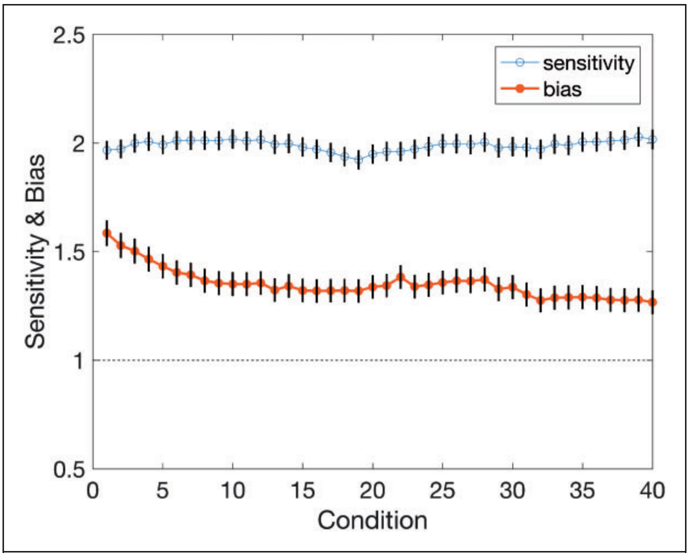

Figure 6. Sensitivity d' (blue open circles) and bias β (red filled circles) computed over a sliding window of 20 images for Experiment 3 (training and feedback). The error bars correspond to a 95% confidence interval.

Building upon our two previous experiments, we attempted to further explore the limits of human performance on this task. To do so, we provided observers with the same training as in the second experiment, and we provided feedback during testing. After each response in this third experiment, the image was annotated with the correct classification (photographic or CG). To make sure that the observer noted the feedback, he or she had to click the button corresponding to the correct answer. The correct answer was displayed in red if the observer's response was incorrect and in green if it was correct.

Shown in Figure 5(c) is the average observer accuracy in correctly identifying computer-generated and photographic images as a function of image resolution. Shown in Table 1(c) are these results expressed as d' and β. With training and feedback, sensitivity (d') increased and bias (β) decreased. As compared to the condition with training but no feedback, average d' increased from 1.87 to 2.01 and β fell from 1.60 to 1.49.

Interestingly, we observed that over the course of the experiment, observer sensitivity was relatively constant and bias steadily decreased as observers received more feedback. In particular, shown in Figure 6 is the sensitivity and bias averaged over a sliding window of 20 trials. With this sliding window, the first data point is the average response over trials 1 through 20, the second data point is the average response over trials 2 through 21, and so forth. The average sensitivity (the open (blue) data points) fluctuates from a minimum of 1.92 to a maximum of 2.03 but shows no discernible trend over time. In contrast, the bias (the filled [red] data points) starts out at a high of 1.59 and gradually decreases to 1.24. These results imply that the sort of feedback provided in this experiment helps observers reduce their bias but does not help them improve their sensitivity.

Incentives

In this fourth experiment, we attempted to further improve observer performance by incentivizing our observers with both a sense of imperative, a financial reward,

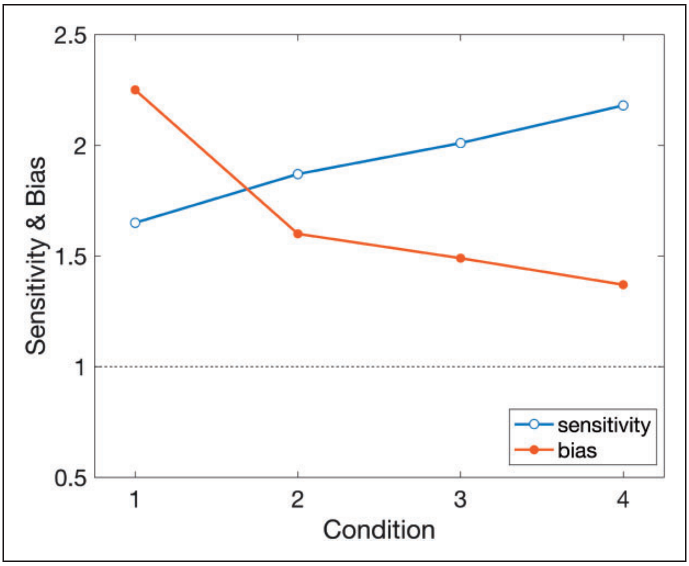

Figure 7. Sensitivity d' (blue open circles) and bias β (red filled circles) for each of four conditions: (a) no training and no feedback; (b) training and no feedback; (c) training and feedback; (d) training, feedback, and incentives. The error bar for each condition is smaller than the data point.

and a sense of competitiveness. Specifically, we added the following text to the instructions:

> Child pornographers in the United States can escape justice by claiming that their images are computer generated. Your performance in this study will help to evaluate if jurors can reliably determine if an image is computer generated or not. We ask, therefore, that you try your best at this task.

We also increased the payment from \$0.50 to \$1.00 and observers were told at the beginning of the experiment that they would receive a \$4.00 bonus if their overall accuracy was 90% or higher (their accuracy was shown after each trial). And lastly, to tap into their competitiveness, observers were told on each trial if their performance was below or above average performance from our previous observers.

Shown in Figure 5(d) is the average observer accuracy in correctly identifying computer-generated and photographic images as a function of image resolution. Shown in Table 1(d) are these results expressed as d' and β. The addition of incentives led to an increase in sensitivity (d') and reduction in bias (β). As compared to the condition with training and feedback, average d' increased from 2.01 to 2.18 and β fell from 1.49 to 1.37.

Although we cannot say which of the incentives was most critical (imperative, financial, or competitiveness), the introduction of incentives considerably improved observer performance. We imagine that in a legal setting, where the stakes are quite high, jurors, prosecutors, and judges would be equally incentivized.

Shown in Figure 7 is the average sensitivity (d') and bias (β) for each of the four conditions. We performed a four-way ANOVA to determine if there was a main effect on d' and β for the four conditions (no training/no feedback; training/no feedback; training/feedback; training/feedback/incentives). The ANOVA, however, failed Levene's test for homogeneity of variance for both d' and β. We therefore performed a nonparametric permutation analysis which revealed a significant ($p < 0.05$) effect of condition for both d' and β.

Follow-up Wilcoxon tests (one-sided) revealed the following pattern of significance for training (T), feedback (F), and incentives (I). The p values were corrected for multiple comparisons using the Holm method.

Conditions	d'	β
no T/no F v. T/no F	<0.001	<0.001
no T/no F v. T/F	<0.001	<0.001
no T/no F v. T/F/I	<0.001	<0.001
T/no F v. T/F	<0.001	0.002
T/no F v. T/F/I	<0.001	<0.001
T/F v. T/F/I	<0.001	0.021

In summary, training, feedback, and incentives significantly increase sensitivity and decrease bias.

Single Resolution

In the fifth experiment, we tested observers on just one resolution using the testing plus feedback paradigm (Feedback section). We hypothesized that observers may use different strategies to classify high-, medium-, and low-resolution images. If this were true, then we would expect that more training/feedback at one resolution would improve performance.

A new set of observers was trained and tested on the full set of 20 training and 60 testing images at one resolution of 600, 400, or 200 pixels in size. The 300 observers were divided into three pools with 100 observers responding to each resolution. At a resolution of 600, the average sensitivity and bias were 2.09 and 1.04, as compared to 2.14 and 1.40 in Experiment 3 (Feedback section). At a resolution of 400, the average sensitivity and bias were 2.09 and 1.22, as compared to 2.10 and 1.49 in Experiment 3. And, for a resolution of 200, the average sensitivity and bias were 1.88 and 1.37, as compared to 1.92 and 1.60.

A two-sided Wilcoxon test reveals that d' for the variable-resolution is higher ($p < .001$) and that there was no statistically significant effect on β ($p = .25$). It seems, therefore, that observers do not use a specialized resolution-based strategy.

We were surprised by this finding and by the general observation that across all conditions, there was not a more dramatic effect of resolution on performance. We hypothesize that this is a result of the fact that observers commonly see faces at different distances and so perhaps have adapted to reasoning about faces at a range of resolutions. To test this theory, we ran a control condition in which images were down-sized to each of six resolutions and then re-sized back to their original resolution of 600 pixels. In this fixed-resolution condition, the images had the same spatial frequency content as the variable-resolution condition. If our theory was correct, then we would have expected to see lower performance on this condition. However, the average d' and β in the fixed-resolution condition was 2.03 and 1.59 as compared to 2.05 and 1.50 in the variable-resolution condition. A two-sided Wilcoxon tests reveals no statistically significant difference in d' or β between these conditions. Contrary to our hypothesis, observer performance was not lower in the fixed-resolution condition. Although somewhat surprising, our results are consistent with previous studies that showed that images as small as 32×32 pixels provide enough information to identify objects and the semantic category of real-world scenes (Torralba, 2009).

Figure 8. Original (top) and masked eyes (bottom) photographic (Columns 1 and 3) and computer-generated (Columns 2 and 4) images.

Masked Eyes

In the sixth and final experiment, we attempted to determine what cues observers may use to differentiate photographic from computer-generated images. We hypothesized that highly realistic computer-generated eyes may play a role in assigning animacy to a photo (see, for example, Looser & Wheatley, 2010). To test this hypothesis, we placed a black bar over both eyes in each of the 20 training and 60 testing images (Figure 8).

Shown in Table 1(e) are these results expressed as d' and β. Across each resolution, sensitivity (d') decreased but there was almost no bias ($\beta \approx 1$). As compared to the previous condition (with training and feedback, Feedback section), the average d' decreased from 2.01 to 1.41 and β fell slightly from 1.49 to 1.32. A nonparametric permutation analysis reveals a significant ($p < 0.05$) effect of condition for both d' and β. This suggests that the eyes are helpful in distinguishing photographic from computer-generated faces.

To make sure that the differences in performance were in fact due specifically to the eyes being masked out, we repeated this experiment with a mask of the same size placed randomly in each image. In this condition, the average d' and β were 1.98 and 1.70, similar to the original unmasked condition (Experiment 3). It would seem then that the eyes do play an important role in assigning animacy: Their absence removes bias and also makes the task considerably more difficult. A more in-depth examination found no consistent benefit afforded by different facial features (Holmes, 2016).

Familiar Faces

Several of the faces in our testing data set were of famous actors (Robert DeNiro, Pete Dinklage, Russel Crowe, Tom Cruise, Bryan Cranston, and AnnaSophia Robb). Because it has been established that familiar and unfamiliar faces are processed differently (Johnston & Edmonds, 2009), we wondered if these differences transferred to our task of differentiating computer-generated from photographic images. We separately analyzed performance between the seven pairs of familiar faces from the remaining 23 pairs of unfamiliar faces. In the training, feedback, and incentives condition, observers performed better across all resolutions. Averaged across all resolutions, observer accuracy on familiar CG and photo faces was 81.6% and 90.4% ($d' = 2.48$ and $\beta = 1.21$) as compared to performance on unfamiliar faces of 84.4% and 84.0% ($d' = 2.18$ and $\beta = 1.36$). A nonparametric permutation analysis reveals a significant ($p < 0.05$) effect of familiarity on both d' and β.

As with other aspects of face perception and recognition, observer's performance with familiar faces is somewhat better.

Discussion

Modern-day computer graphics is capable of rendering highly photo-realistic imagery. Despite these advances, we find that with the appropriate training and incentives, observers are able to reliably distinguish between computer-generated and photographic portraits: depending on the resolution of the images, observer accuracy ranges from 85% to 90%. We were struck by the impact of incentivizing observers with a sense of imperative and competitiveness along with a financial reward. The effect of these incentives is particularly important to anyone performing psychophysical studies on Mechanical Turk or similar platforms (Buhrmester, Kwangand, & Gosling, 2011; Horton, Rand, & Zeckhauser, 2011; Paolacci, Chandler, & Ipeirotis, 2010).

It could be argued that the increase in performance in the training condition is a result of observers noting that the number of computer-generated and photographic images were balanced (they were not told the distribution in any of the conditions). If this were the case, then we would not have expected to see a continuing improvement in the feedback and incentive conditions. While there may have been some benefit to knowing the distribution of images, we believe that the more likely explanation for the improved performance is that our observers simply learned differentiating characteristics through training and feedback.

Precisely what image features or properties that observers are using is still not clear. We expect that by better understanding how observers are performing, we can train observers to perform even better at this task and, paradoxically, improve the photo-realism of computer-generated content.

Acknowledgements

The authors thank Paul Whalen for his suggestion to explore the impact of observer incentives and Emily Cooper for her assistance with the statistical analysis.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by NSF BCS-1354029.

Note

1. Denote the hit-rate H as the accuracy of correctly identifying a computer-generated image and the false-alarm rate F as 1 minus the accuracy of correctly identifying a photographic image. Observer sensitivity d' is $z(H) - z(F)$, where $z(\cdot)$ corresponds to the z-transform. Observer bias β is $g(z(H))/g(z(F))$, where $g(\cdot)$ is a zero-mean, unit-variance normal distribution.

References

Ashcroft v. Free Speech Coalition, 535 U.S. 234 (2002).

Buhrmester, M., Kwangand, T., & Gosling, S. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 1, 3–5.

Conotter, V., Bodnari, E., Giulia, B., & Farid, H. (2014). Physiologically-based detection of computer generated faces in video. In *IEEE international conference on image processing* (pp. 248–252). New York, NY: IEEE.

Dang-Nguyen, D.-T., Boato, G., & De Natale, F. G. B. (2012a). Discrimination between computer generated and natural human faces based on asymmetry information. In *IEEE European signal processing* (pp. 1234–1238). New York, NY: IEEE.

Dang-Nguyen, D.-T., Boato, G., & De Natale, F. G. B. (2012b). Identify computer generated characters by analysing facial expressions variation. In *IEEE workshop on information forensics and security* (pp. 252–257). New York, NY: IEEE.

Dehnie, S., Sencar, H., & Memon, S. (2006). Digital image forensics for identifying computer generated and digital camera images. In *IEEE international conference on image processing* (pp. 2313–2316). New York, NY: IEEE.

Dirik, A. E., Bayram, S., Sencar, H. T., & Memon, N. (2007). New features to identify computer generated images. In *IEEE international conference on image processing* (Vol. 4). New York, NY: IEEE.

Fan, S., Ng, T.-T., Herberg, J. S., Koenig, B. L., & Xin, S. (2012). Real or fake? Human judgments about photographs and computer-generated images of faces. In *SIGGRAPH Asia 2012 technical briefs* (pp. 17:1–17:4). New York, NY: ACM.

Farid, H., & Bravo, M. J. (2007). Photorealistic rendering: How realistic is it? *Journal of Vision*, *7*, 9.

Farid, H., & Bravo, M. J. (2012). Perceptual discrimination of computer generated and photographic faces. *Digital Investigation*, *8*, 226–235.

Gallagher, A. C., & Chen, T. (2008). Image authentication by detecting traces of demosaicing. In *IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 1–8). New York, NY: IEEE.

Holmes, O. (2016). *How realistic is photorealistic?* (Dartmouth Computer Science Technical Report TR2015-786).

Holmes, O., Banks, M. S., & Farid, H. (2016). Assessing and improving the identification of computer generated portraits. *ACM Transactions on Perception*, *13*, 7:1–7:12.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*, 399–425.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*, 577–596.

Khanna, N., Chiu, G. T. C., Allebach, J. P., & Delp, E. J. (2008). Forensic techniques for classifying scanner, computer generated and digital camera images. In *IEEE international conference on acoustics, speech and signal processing* (pp. 1653–1656). New York, NY: IEEE.

Lalonde, J. F., & Efros, A. A. (2007). Using color compatibility for assessing image realism. In *IEEE international conference on computer vision* (pp. 1–8). New York, NY: IEEE.

Looser, C., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science*, *21*, 1854–1862.

Lyu, S., & Farid, H. (2005). How realistic is photorealistic? *IEEE Transactions on Signal Processing*, *53*, 845–850.

Ng, T.-T., Chang, S.-F., Hsu, J., Xie, L., & Tsui, M.-P. (2005). Physics-motivated features for distinguishing photographic images and computer graphics. In *ACM international conference on multimedia* (pp. 239–248). New York, NY: ACM.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical Turk. *Judgment and Decision Making*, *5*, 5.

Torralba, A. (2009). How many pixels make an image? *Visual Neuroscience*, *26*, 123–131.

Wang, Y., & Moulin, P. (2006). On discrimination between photorealistic and photographic images. In *IEEE international conference on acoustics, speech and signal processing* (Vol. 2). New York, NY: IEEE.

Wen, C., Shi Yun, Q., & Guorong, X. (2007). Identifying computer graphics using HSV color model. In *IEEE international conference on multimedia and expo* (pp. 1123–1126). New York, NY: IEEE.