Perceptual Discrimination of Computer Generated and Photographic Faces

Hany Farid and Mary J. Bravo

Department of Computer Science, Dartmouth College
Department of Psychology, Rutgers University

Abstract

Modern day computer graphics are capable of generating highly photorealistic images resulting in challenging legal situations. For example, as a result of a 2002 U.S. Supreme Court ruling, computer generated child pornography is protected speech, while pornographic photographs depicting an actual child remains illegal. The ability to distinguish between protected and illegal material assumes that law enforcement agents, attorneys, jurors, and judges can reliably distinguish between computer generated and photographic imagery. We describe a series of psychophysical experiments that used images of varying resolution, JPEG compression, and color to explore the ability of observers to distinguish computer generated from photographic images of people. The results allow us to assign a probability that an image that is judged to be a photograph is, in fact, a photograph.

Keywords: Digital Forensics, Photorealism

1. Introduction

The past few decades have seen tremendous advances in computer graphics rendering software and hardware. These advances have led to remarkable theatrical releases that blur the line between reality and fantasy. At the same time, this technology has resulted in challenging legal situations. Most notably, the 1996 Child Pornography Prevention Act expanded the prohibition on child pornography to include not only pornographic images of actual children but also any computer generated (CG) images that simulate a minor engaging in sexually explicit conduct. This ruling was subsequently challenged, and in 2002 the U.S. Supreme Court found that portions of the

CPPA were overly broad and infringed on the first amendment (*Ashcroft v. Free Speech Coalition*). This new ruling classified computer generated child pornography, the creation of which does not involve an actual child, as protected speech. The ability to distinguish between protected (CG) and illegal (photographic) material has, therefore, become essential.

Previous work has considered computational techniques for distinguishing CG from photographic images [1, 2, 3, 4, 5, 6]. These techniques exploit a variety of low-level statistical differences to classify images as either CG or photographic. While these techniques have met with some success, they are vulnerable to variations in resolution, image quality (SNR), compression quality, and color, which can dramatically alter the underlying statistical measurements.

On the other hand, the human visual system is remarkable at reasoning about images and videos of humans and human faces [7, 8]. We previously conducted a perceptual study to test the ability of human observers to discriminate CG and photographic images [9]. This study compared performance for images of people, man-made objects, and natural environments for CG images generated between the years 2000 and 2006. We found that human observers were able to reliably distinguish between CG and photographic images. Building on this earlier work, the current study focuses exclusively on images of people, it updates the CG images to include images rendered between 2007 and 2010, and it explores the impact of the variations in image quality that arise in real-world settings: (1) resolution, (2) JPEG compression, and (3) color vs. grayscale. We describe a series of experiments that probe the reliability of observers to judge that a photograph is in fact a photograph. We use this measure of performance because in child pornography cases, the legality of an image depends on whether it is a photograph.

2. Methods

2.1. Images

We downloaded thirty CG images from two popular computer graphics websites (www.forums.cgsociety.org and www.creativecrash.com). We received written confirmation from the website editors that the posted images were solely computer generated (i.e., did not contain any photographic components). The people depicted in these images vary in age, gender, race, pose, and lighting. In order to avoid bias in the selection of images, we downloaded all images of people that were of sufficient resolution. While

Figure 1: Shown are paired CG (left) and photographic (right) images. See also Figure 2.

most of the images exemplified cutting-edge photorealism, there was some variability in quality.

As our control set, we downloaded thirty high-resolution photographic images that closely matched the CG images in terms of age, gender, race, and pose. These images were downloaded from a variety of websites. The content and context of these websites virtually guaranteed that these images were photographic in nature.

The background was manually deleted from each CG and photographic image so that observers could only use the rendered or photographed person to make their judgments.

Because the sizes of the CG and photographic images varied significantly, each image was cropped to a square aspect-ratio and down-sampled so that the area of the person depicted was $122,600$ pixels (approximately 350×350

Figure 2: Shown are paired CG (left) and photographic (right) images. See also Figure 1.

pixels). This down-sampling had the added benefit of largely removing any JPEG artifacts in the original JPEG image. The image was then saved as a JPEG with quality 100 (on a scale of 100 (best) to 1 (worst)).

Each CG and photographic image was then color adjusted to match the brightness (mean) and contrast (variance) of each luminance and chrominance channel. Denote a CG image as $f_g(x, y, c)$ and a photographic image as $f_p(x, y, c)$, where c corresponds to the luminance (Y), chrominance (Cr), or chrominance (Cb) channel. The brightnesses were matched by adjusting the mean of each channel as follows:

$$f_g(x, y, c) = f_g(x, y, c) - \mu_g(c) + \mu(c) \quad (1)$$

$$f_p(x, y, c) = f_p(x, y, c) - \mu_p(c) + \mu(c) \quad (2)$$

where $\mu_g(c)$ and $\mu_p(c)$ are the means of the c^{th} channel of the CG and pho-

Figure 3: A CG image at six different resolutions. For point of reference, the third image from the right is of the size of a typical thumbnail.

tographic images, and $\mu(c)$ is the average brightness across all 60 images ($\vec{\mu} = [150\ 121\ 137]$). The contrasts were then matched by adjusting the variances of each channel as follows:

$$f_g(x,y,c) = \sqrt{\frac{\sigma}{\sigma_g}}(f_g(x,y,c) - \mu(c)) + \mu(c) \quad (3)$$

$$f_p(x,y,c) = \sqrt{\frac{\sigma}{\sigma_p}}(f_p(x,y,c) - \mu(c)) + \mu(c) \quad (4)$$

where σ_g and σ_p are the variances of the c^{th} channel of the CG and photographic image, and σ is the average variance across all 60 images ($\vec{\sigma} = [5527\ 69\ 109]$). After color adjusting in the luminance/chrominance space, the images were converted back to their original RGB color space. This brightness and contrast matching ensured that observers could not classify images based on systematic differences in simple low-level image statistic (as they could if, for example, the CG images generally had a higher contrast than their photographic counterparts).

The cropped, masked, down-sampled, and brightness and contrast adjusted images are shown in Figures 1 and 2, where the paired CG and photographic images are shown side-by-side.

Figure 4: A CG image compressed at six different qualities: worst (left) to best (right).

2.1.1. Image Manipulations

In addition to testing the ability of human observers to classify images as CG or photographic, we also wanted to consider the impact of subjecting images to basic degradations in quality that typically arise in real-world settings. In particular, we considered the effect of (1) resolution, (2) JPEG compression, and (3) color vs. grayscale.

Shown in Figure 3 is a CG image at its original resolution (1.000), and down-sized by a factor of 0.025, 0.050, 0.100, 0.250, and 0.500 along each dimension. Because the images were initially scaled to match the area of the person depicted, the absolute size of these images depends on content. Across all 60 CG and photographic images, the average size in pixels are: 13×13, 27×27, 54×54, 109×109, 218×218, and 436×436.

Shown in Figure 4 is a CG image at 0.250 resolution and JPEG compressed with a quality of 10, 15, 25, 50, 75, and 100. The JPEG qualities are specified on a scale of 1 (worst) to 100 (best), as employed by the MatLab JPEG encoder (`imwrite`).

Shown in Figure 5 are three CG and three photographic images in their original RGB color and converted to grayscale as follows:

$$\text{gray} \quad = \quad 0.299\text{R} + 0.587\text{G} + 0.114\text{B}. \tag{5}$$

In the final manipulation, we asked observers to classify images that were inverted, Figure 5. Inversion is known to greatly impair face recognitions [8], but it should have no effect on the discrimination of most image statistics. If performance is unaffected by inversion, then this would suggest that observers are basing their judgment on a low-level cue.

We did not explore the full space of resolution, quality, grayscale, and orientation, as this would have produced a prohibitively large number of stimuli. Instead, we explored slices through this parameter space. Most notably, the JPEG quality was fixed at 100 as the resolution was varied, and the resolution was fixed at 0.250 as the quality was varied.

Figure 5: CG (left) and matching photographic (right) images in color (top), grayscale (middle), and inverted (bottom).

2.2. Psychophysical Setup

Thirty-six observers were recruited from the Introduction to Psychology subject pool at Rutgers-Camden. All observers reported having normal or corrected-to-normal acuity and normal color vision.

The 360 stimuli (6 resolutions $\times$ 1 quality $\times$ 60 images, or 1 resolution $\times$ 6 qualities $\times$ 60 images) were divided into 6 sets such that every image appeared once in a set. Each observer ran only one set (i.e., they saw each CG or photographic person once and at one resolution or one quality). Observers were told that they would see a sequence of images and that their task was to determine whether each image was photographic or computer generated, where CG is defined as an image entirely created using computer software. Observers were also told that the background had been removed from each image and so the blank backgrounds provided no useful information.

The stimuli were displayed on a PowerMac G5 using Psychtoolbox routines [10]. The viewing distance was not constrained and many observers adjusted their viewing distance depending on the size of the image. The observers registered their responses using the "F" and "J" keys on the keyboard. To ensure that observers did not rush their judgment, the experimental program ignored any responses made within 3 seconds of image onset.

No feedback was given and there was a 1 second delay between images. The entire experiment lasted approximately 5 minutes.

A second group of observers was recruited through Amazon's Mechanical Turk. This crowd sourcing utility has become popular among social scientists as a way to quickly collect large amounts of data from human observers around the world [11].

Four hundred and thirty six observers were paid $1.00 each to classify 30 images (partitioned similar to that described above). Given the uncontrolled nature of the data collection, some data filtering was necessary. Approximately 10% of observers were excluded because they responded randomly (missing the easiest images) or because they always pressed the same key on every trial.

3. Results

We characterize reliability as the probability that an image is in fact photographic when an observer believes it to be photographic. We adopt this measure because in a legal setting it is the most critical measure in assessing a defendant's guilt or innocence. This measure can be expressed as the following conditional probability, where R denotes the user response and I the image category (both variables can take on the value of "CG" or "photo"):

$$P(I = \text{photo}|R = \text{photo}). \tag{6}$$

This conditional probability can be estimated directly from the observer responses.

3.1. Resolution

Shown in the first two rows of Figure 6 is the reliability of photograph classification, Equation (6), for the color (RGB) condition at six different resolutions and a JPEG quality of 100. The first row corresponds to our laboratory observers and the second row corresponds to our Mechanical Turk observers. Note first that both sets of observers have fairly similar performance, with the laboratory observers performing slightly better at the higher resolutions. An interesting and somewhat surprising result is that observers consistently perform better at one-half resolution than the full resolution: for example, 90.0% versus 81.4% for the laboratory observers. Pooled across

	0.025	0.050	0.100	0.250	0.500	1.000
RGB	65.4%	58.4%	78.0%	79.8%	90.0%	81.4%
RGB	62.1%	64.8%	73.1%	79.6%	84.7%	79.5%
Gray	54.6%	62.5%	65.6%	78.3%	76.0%	75.3%
Invert	54.9%	66.0%	68.7%	75.6%	73.3%	71.0%

Figure 6: Shown is the reliability of photograph classification, Equation (6), for different resolutions and conditions. The first row (RGB) corresponds to the accuracy for our laboratory observers and the second row (RGB) and remaining rows correspond to our Mechanical Turk observers.

observers, this difference was significant ($z = -2.887$, two tailed $p < 0.0039$). We speculate that performance at the highest resolution is lower because the fine details in computer generated images are so accurate that observers take their presence as evidence of a photographic image. At one-half resolution, however, these details are not visible, and so observers rely on other cues which, interestingly, are more distinguishing.

Even at the lowest resolutions of 0.025 and 0.050, corresponding to an average image size of 13×13 and 27×27 pixels, observers are above chance performance of 50%. This surprising accuracy is consistent with previous studies which showed that images as small as 32×32 pixels provide enough information to identify objects and the semantic category of real-world scenes [12].

Shown in Figure 7 are all thirty CG images, at the 0.250 resolution, ranked in order of perceptual discrimination averaged over all laboratory and Mechanical Turk observers. There is no obvious relationship between gender, age, race, pose, etc. and ease of classification.

In addition to the reliability of judging that an image is a photograph, it is also useful to know the reliability of judging that an image is CG. By replacing "photo" with "CG" in Equation 6, we have estimated the conditional probability that an image is CG if an observers says it is CG, $P(I = \text{CG}|R = \text{CG})$. The reliability of the Mechanical Turk observers, from smallest to largest scale, is 60.25%, 72.51%, 88.11%, 90.48%, 94.42%, and 92.28%. Note first

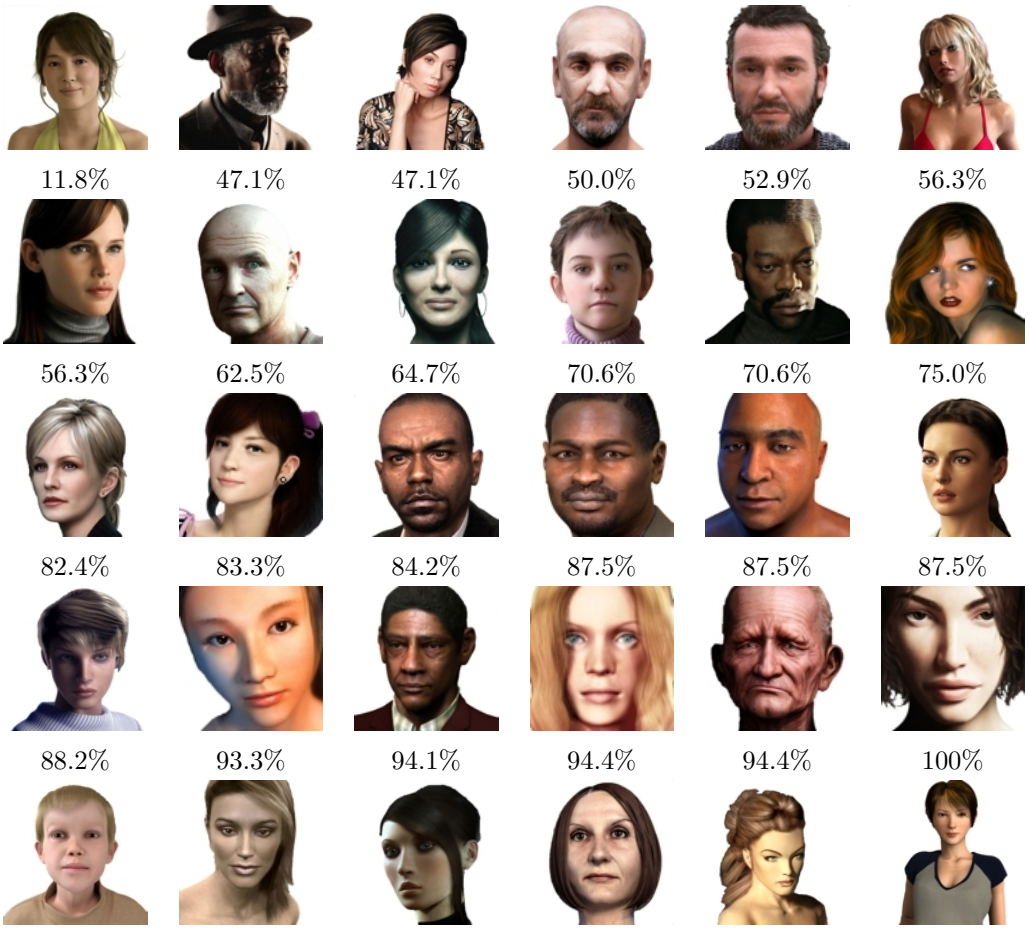

Figure 7: Shown are all thirty CG images ranked (left to right and top to bottom) in order of perceptual discrimination for the 0.250 resolution condition. The value below each image denotes the percentage of trials in which observers correctly classified the image as CG.

	10	15	25	50	75	100
JPEG	72.9%	77.4%	77.3%	76.5%	81.2%	81.4%

Figure 8: Shown is the reliability of photograph classification, Equation (6), for different JPEG qualities.

that observers are highly reliable across resolution. These results are also valuable in interpreting the reliability of photograph classification. If, for example, observers were highly conservative in classifying an image as photo, then they would have a high reliability for photographs but low reliability for CG. This is clearly not the case since observers have similar reliability for judging photographs and CG.

Recall that observers were forced to view each image for a minimum of 3 seconds. The laboratory observers viewed each CG image for an average of 4.80 seconds and each photographic image for an average of 4.51 seconds. The Mechanical Turk observers viewed each CG image for an average of 5.58 seconds and each photographic image for an average of 4.96 seconds.

3.2. JPEG Compression

Shown in Figure 8 is the reliability of the photograph classification, Equation (6), for the JPEG compression condition at a single resolution of 0.250. Note first that the accuracy at the highest JPEG quality of 100 is comparable to the same condition in Figure 6 (second row, fourth column). In addition, performance stays fairly constant through a broad range of compression qualities. At the lowest JPEG quality, the performance degrades to 72.9% from a maximum of 81.4%.

Observers viewed each CG image for an average of 5.25 seconds and each photographic image for an average of 5.28 seconds.

3.3. Grayscale

Shown in the third row of Figure 6 is the reliability of photograph classification, Equation (6), for the grayscale condition. As with the RGB condition, six resolutions and one JPEG quality were considered. At the highest resolutions there is a significant degradation in performance due to the loss of color. The most significant degradation comes at one-half resolution where performance drops from 90.0% to 76.0%.

Observers viewed each CG image for an average of 5.39 seconds and each photographic image for an average of 5.39 seconds.

3.4. Orientation

Shown in the fourth row of Figure 6 is the reliability of photograph classification, Equation (6), for the condition in which the images were inverted. As with the RGB and grayscale conditions, six resolutions and one JPEG quality were considered. At each resolution there is a significant degradation in performance. There are no practical implications of this condition because unlike the, resolution, JPEG compression, and grayscale conditions, it is trivial to control for the orientation of an image. This condition does, however, suggest that observers are employing some high-level perceptual judgment, as opposed to some low-level statistical judgment, which would not have been affected by the orientation of the face.

Similar to previous conditions, observers viewed each CG image for an average of 5.79 seconds and each photographic image for an average of 5.00 seconds.

4. Discussion

Recent advances in computer graphics rendering technologies have made it possible to create remarkably photorealistic imagery. This blurring of the perceptual boundary between real and fake poses a significant forensic challenge because the legality of an image may hinge on whether it is computer generated or photographic. Current computational techniques for image classification are limited, especially for images with low resolution or high image compression. We have explored the ability of observers to reliably determine that an image of a person is a photograph, and therefore illegal if it depicts child pornography. We have considered the impact of resolution, compression quality, and color on performance. The results of these experiments are summarized in a simple and intuitive measure of the probability that an image is photographic when an observer classifies it as photographic.

When observers judged an image as a photograph, they are 85% reliable for color images with medium resolution (between 218×218 and 436×436 pixels in size) and high JPEG quality. This judgement was still quite reliable (75%) for images as small as 54×54, and and even the seemingly impossible images 13×13 pixels in size, supported above chance performance. In general, judgments of grayscale images were less reliable than their color counterparts.

Even fairly significant JPEG compression has a relatively minimal impact on reliability.

It seems very likely that the accuracies reported here are a lower bound on human performance. Our observers were given no training and no incentive to perform well (their reward was independent of their performance). Most decisions were made within 5 seconds. Compared with the types of images encountered in forensic settings, our images were relatively impoverished, containing only a single person depicted from the neck up against a blank background. A full body figure interacting with the environment or other people is far more difficult to render photorealistically. But while observer performance can likely be improved, there is little doubt that with time rendering technologies will also improve.

The discriminations examined in this study currently play a role in deciding the outcome of court cases. So it is essential to have some measure of the reliability of these discriminations. To our knowledge this study provides the first such measure, and so despite the unavoidable limitations described above, these data should have direct and immediate impact on legal practitioners who must evaluate photographic evidence in child pornography cases. In addition, these data should inform policy makers both here and abroad on the practicality of laws that require discriminating photographs from computer generated images.

Acknowledgment

This work was supported by a gift from Adobe Systems, Inc., a gift from Microsoft, Inc. and a grant from the National Science Foundation (CNS-0708209).

References

[1] S. Lyu, H. Farid, How realistic is photorealistic?, IEEE Transactions on Signal Processing 53 (2) (2005) 845–850.

[2] T.-T. Ng, S.-F. Chang, Y.-F. Hsu, L. Xie, M.-P. Tsui, Physics-motivated features for distinguishing photographic images and computer graphics, in: ACM Multimedia, 2005, pp. 239–248.

[3] S. Dehnie, H. T. Sencar, N. Memon, Digital image forensics for identifying computer generated and digital camera images, in: IEEE International Conference on Image Processing, 2006, pp. 2313–2316.

[4] Y. Wang, P. Moulin, On discrimination between photorealistic and photographic images, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2006.

[5] N. Khanna, G. T.-C. Chiu, J. P. Allebach, E. J. Delp, Forensic technqiues for classifying scanner, computer generated and digital camera images, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2008, pp. 1653–1656.

[6] P. Sutthiwan, X. Cai, Y. Q. Shi, H. Zhang, Computer graphics classification based on Markov process model and boosting feature selection technique, in: IEEE International Conference on Image Processing, 2009, pp. 2913–2916.

[7] R. Blake, M. Shiffrar, Perception of human motion, Annual Review of Psychology 58 (2007) 47–73.

[8] P. Sinha, B. Balas, Y. Ostrovsky, R. Russell, Face recognition by humans: 19 results all computer vision researchers should know about, Proceedings of the IEEE 94 (11) (2006) 1948–1962.

[9] H. Farid, M. Bravo, Photorealistic rendering: How realistic is it?, Journal of Vision 7 (9) (2007) [abstract].

[10] D. Brainard, The psychophysics toolbox, Spatial Vision 10 (4) (1997) 433–436.

[11] G. Paolacci, J. Chandler, P. Ipeirotis, Running experiments on Amazon mechanical turk, Judgment and Decision Making 5 (5) (2010) 411–419.

[12] A. Torralba, How many pixels make an image?, Visual Neuroscience 26 (1) (2009) 123–131.