How Realistic is Photorealistic?

Olivia B. Holmes
Senior Honors Thesis
Advisor: Professor Hany Farid

Dartmouth Computer Science Technical Report TR2015-786

June 6, 2015

Abstract

From its inception in 1960, computer graphics (CG) technology has quickly progressed from simple 3-D models to complex, photorealistic recreations of the human face and body. Alongside this innovation, lawmakers and courts in the United States have struggled to define what is illegal, what is "obscene?, and what is protected under the First Amendment with regards to child pornography. What has emerged from this debate is that the laws surrounding child pornography hinge on whether the material in question is photographic or CG. To this end, we measure how reliable the human visual system is in distinguishing CG from photographic images. After establishing a baseline for observer performance in this task as a function of both resolution and contrast, we address the following two questions: (1) is it possible to improve observer performance by isolating select features of the face? and (2) will training observers improve their performance?

Contents

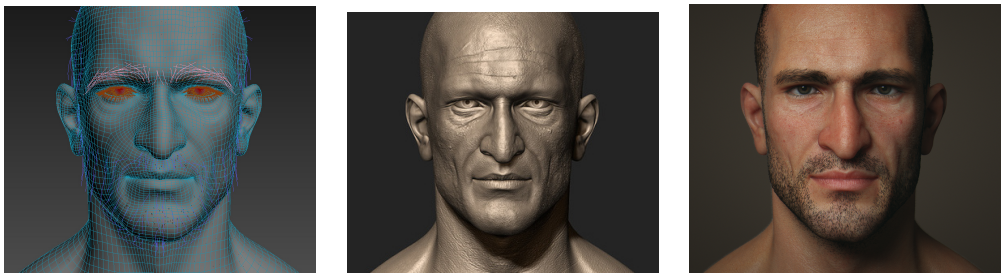

Figure 1: The making of a CG model (from left): wire-frame, skinned, and texture-mapped.

1 Introduction

In 1960, the term computer graphics was coined to describe the newly formed field of digital artistic expression [1]. With time, "CGI" (computer-generated imagery) has become the popular way to refer to any image that is created solely through the use of a computer. Although CGI can trace its roots back to humble black and white geometric shapes rendered on low-resolution monitors, today the best CG images blur the line between what is virtual and what is real.

Making a CG image requires (1) creating a three-dimensional model of the subject(s), (2) adding color and texture, and (3) illuminating the model(s) with a virtual light source (Figure 1). Once the model has been created, it is rendered, a process analogous to taking a picture with a virtual camera.

Savants in computer graphics have mastered this process which has allowed them to create intensely realistic CG images that are easily mistaken for photographs. However, the computer programs that assist artists in creating these images would not be in existence without the now primitive tools like Sketchpad (created in 1963 by Ivan Sutherland [2]) that first allowed for artistic creation via the computer.

Before the advent of "digital art" and CGI, there was "electronic art", a short lived artistic period during the early 1950s defined principally by its use, and manipulation of, sound waves. Ben Laposky, the best known artist for this work, referred to each of his pieces as "oscillons" (Figure 2) because of his use of the audio oscillator to produce the waves in featured in these pieces [3].

The late 1960s introduced basic, two-dimensional human and animal

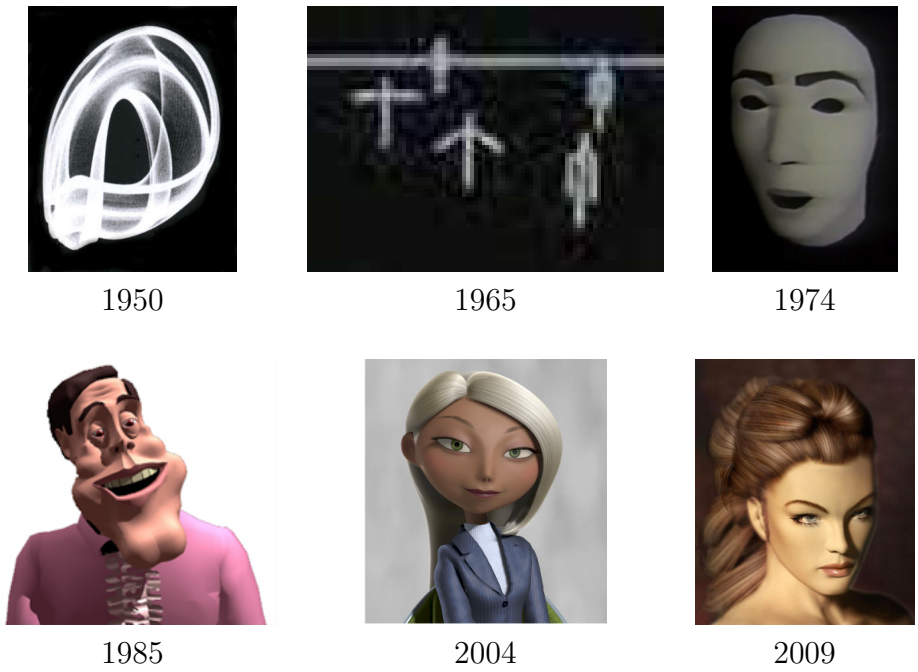

Figure 2: Mapping the progression of CGI from its origins in 1950 to the near-present day (2009).

forms onto the computer screen. In Charles Csuri's "Hummingbird" [4], for example, a hummingbird composed of shaky, disjointed lines flits across the screen. Similarly, in Bell Laboratory's "A Computer-Generated Ballet" [5], a group of spindly stick figures "dance" to and fro (Figure 2). It was not until the mid-1970's that artists began to create graphic models of the human face, initially only producing mask-like figures with little detail or expression (refer to the work of Fred Parke in Figure 2 [6]).

By the 1980s, CGI had progressed in its depiction of the human form through the introduction of cartoon characters, often created to have humorous, exaggerated features (such as the man in the short animated clip, *Tony De Peltrie* [7] in Figure 2). In part because technology placed a limit on the degree of realism that could be achieved by artists, cartoon characters became prolific in computer graphics and animation. The continuous development of this field eventually led to the creation of Pixar's *Toy Story*, the first fully computer animated movie. CGI produced in the early 2000's demonstrates a shift from creating cartoon characters to more realistic hu-

man models, the results of which are evidenced in the images chosen for this study (Figures 3 and 4).

It is important to note that, apart from still imagery, CGI has also infiltrated the medium of video with the help of motion capture technology. Additionally, CG technology has been used to create entire movie sets, such as the mythical worlds of the popular 2009 film, *Avatar*.

Unfortunately, the proliferation of CG has also created complex legal issues surrounding the definition and prosecution of child pornography. In the landmark 1982 case of *New York v. Ferber*, a New York law making child pornography illegal was upheld by the Supreme Court who ruled that doing so was not in violation of the First Amendment [8]. In 1996, Congress passed the Child Pornography Prevention Act (CPPA) which made illegal "any visual depiction including any photograph, film, video, picture or *computer-generated image*" that "is, or appears to be, of a minor engaging in sexually explicit conduct" (emphasis added) [9]. Thus, passing the CPPA was meant, in part, to update the *New York v. Ferber* ruling to respond to technological advances in CGI.

In 2002, however, the CPPA was challenged in the Supreme Court case of *Ashcroft v. Free Speech Coalition* for being overly broad and thus creating an unintentional ban on speech that was, in fact, lawful. The Court ruled that virtual images were to be treated as "protected speech" under First Amendment rights [10]. Unfortunately, one consequence of this act was that it provided anyone accused of possession of child pornography with the defense that the material is computer-generated (regardless of its true origins) and thus, protected.

One year later, in 2003, Congress passed the PROTECT Act which strengthened the provisions against virtual child pornography using the charge of "obscenity" and thus created a further distinction between virtual and photographic material [11]. To determine obscenity, a three-pronged strategy established in the case of *Miller v. California* is employed by the courts. This strategy calls upon the courts to question: "(a) whether 'the average person, applying contemporary community standards' would find that the work, taken as a whole, appeals to the prurient interest. . . (b) whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and (c) whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value." [12]. In 2008, Dwight Whorley became the first person convicted under the PROTECT Act for possession of virtual pornography (among other charges)

that was deemed to be obscene [13].

Out of the continuous discussion surrounding child pornography law in the United States, what is clear is that, since the case of *Ashcroft v. Free Speech Coalition* there is a clear distinction in conviction hinging upon whether the material is photographic or CG. As making this judgement has become of considerable legal importance, it is essential that there is a reliable method in place by which to make this distinction.

Currently, there are two approaches to distinguishing CG from photographic images: computational techniques and human judgement. The statistical models used in computational methods rely on underlying, low-level qualities to classify the image in question [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Although these models have been somewhat successful, they are highly sensitive to changes in characteristics like compression and resolution, both of which are inevitably variable in any real-world situation.

When asked whether one would be able to distinguish a CG from a photographic image, instinctually, many people reply in the affirmative because they believe there is a special "human-ness" present in photographs of real people that cannot be replicated in CG models. Interestingly, there seems to be some truth to this conviction: studies have shown that the human visual system is capable of incredible reasoning powers when presented with a human face [26, 27, 28]. With no viable computational method at hand, it is important to test the strength and reliability of human judgement for this particular task. In other words, can a court place their trust in an ordinary observer (i.e., someone on the jury) to make this critical classification?

The first study to quantify the reliability of observers in performing this task used images rendered between 2007 and 2010 [29]. The authors concluded that observers were fairly reliable in their judgements regarding image type, performing at 85% accuracy for images at medium resolution. A cursory search of current CG images reveals that, astounding strides have been made towards photorealism since 2010. Acknowledging the rapid speed with which CG technology improves, we argue that it is imperative that observers are tested again on more recently rendered images to see if their reliability has changed with time. After establishing a baseline for observer performance in identifying images as CG or photographic as a function of both resolution and contrast, we will address the following:

- Is it possible to improve observer performance by isolating select features of the face?

- Will training observers before they are asked to determine image type help improve their performance?

2 Experiment 1: Resolution

The primary goal of this experiment was to determine how good observers are at identifying images of the human face as CG or photographic. Motivated by a 2010 study [29] that was the first to investigate this problem, the results from Experiment 1 should provide an updated measure of observer reliability.

2.1 Methods

2.1.1 Images

We selected the vast majority of the 30 images used in this experiment from the following popular CG websites: `www.cgsociety.org`, `www.3dtotal.com`, and `www.cgarena.com`. In addition, a few of the CG images were downloaded from the website of a single CG artist (`www.romans3d.ru`). The content and context of these websites virtually guaranteed that these images were computer generated in nature. The primary aim in choosing these images was to select the most photorealistic and recently rendered CG images. Within this already restricted pool of available images, the CG images that were chosen were meant to showcase the diversity in age, race, expression, and accessories (i.e., clothing, glasses) found in real life.

These 30 CG images all had the following qualities: a human face posed facing outward, a resolution of at least 800 pixels (defined as the minimum of its width and height), and a render date between 2013 and 2014 (with the majority created in the latter year). The 30 CG images are composed of 15 male and 15 female faces. Because we planned to test observers on the sex of the person in the image, we tried to choose images where this was easily identifiable.

Thirty "matching" photographic images were also chosen. In selecting the photographic counterpart to a CG image, we looked to match the age, gender, race, pose, and accessories. The 30 photographic images that were chosen were downloaded from several websites (although the majority were found on `www.flickr.com`). The content and context of these websites virtually guaranteed that these images were photographic in nature.

It was important that observers were prevented from classifying images based on underlying low-level cues that might persist across image type. To this effect, we reviewed the backgrounds of all the images. Noting little

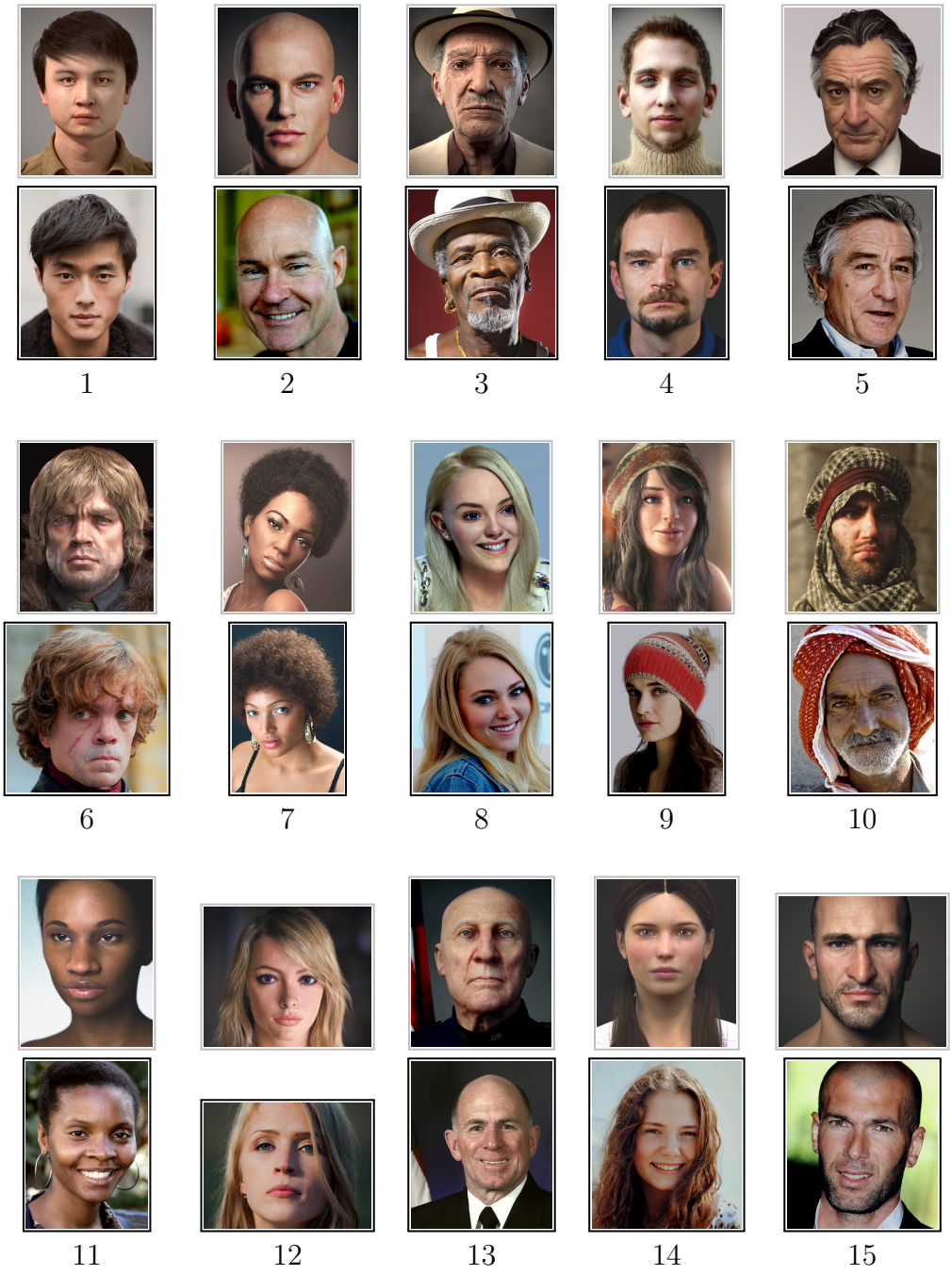

Figure 3: CG images (top, with grey border) paired with their photographic matches (bottom, with black border). See also Figure 4.

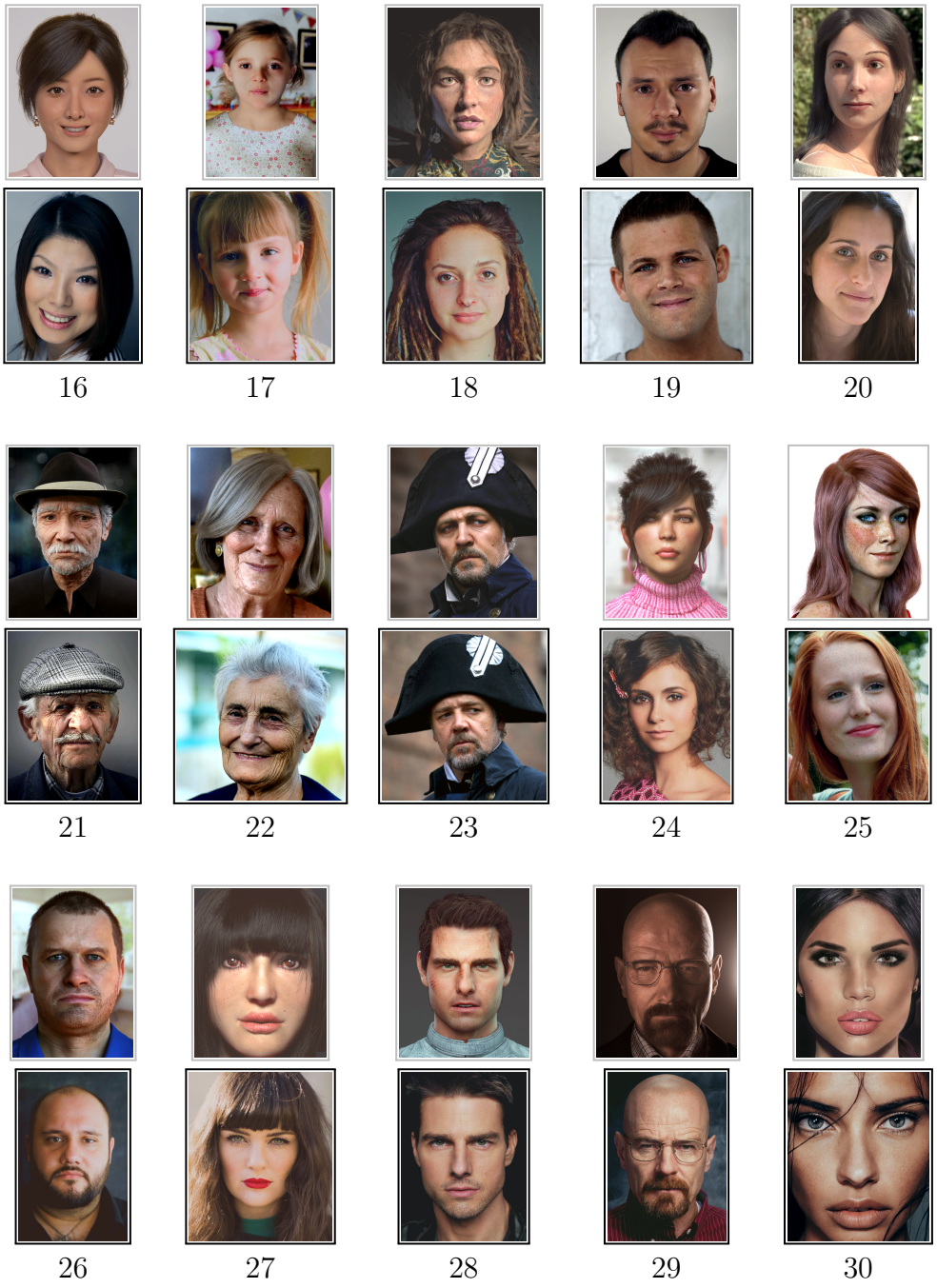

Figure 4: CG images (top, with grey border) paired with their photographic matches (bottom, with black border). See also Figure 3.

difference between image type, we assumed that observers would not be able to classify an image based on the background and thus did not remove it from each image (as was done in [29]). In leaving the image backgrounds intact, we hoped to avoid imposing any artifacts on the image that could impair observer judgement.

The brightness and contrast of an image is another, more subtle cue that an observer could use to classify an image. To resolve this issue, we color adjusted our set of 60 images with respect to brightness (mean) and contrast (variance). Each image was converted from RGB into luminance/chrominance space so that the mean and variance could be calculated for each of the three channels in this space: Y (luminance), Cb (chrominance), and Cr (chrominance). These three channels exhibit more statistical independence than the RGB channels which allowed us to modify each channel individually and then recombine them without creating any artifacts in the process.

In the following equations for color adjustment, the mean is denoted $\mu_c(i)$ for CG images and $\mu_p(i)$ for photographic images where i is one of the three luminance/chrominance channels. The variance is denoted $\sigma_c(i)$ for CG images and $\sigma_p(i)$ for photographic images, where i is similarly the Y, Cb, or Cr channel. Both the mean and variance values were computed over only the facial features of each image to avoid undue impact of particularly dark or light backgrounds.

The mean and variance were then averaged across all 60 images. The average variance and brightness at channel i is denoted as $\sigma(i)$ and $\mu(i)$, respectively. An original CG image, $F_c(i)$, and photographic image, $F_p(i)$, is adjusted to yield color balanced images as follows:

$$\hat{\mathrm{F}}_c(i) = \sqrt{\frac{\sigma(i)}{\sigma_c(i)}}(F_c(i) - \mu_c(i)) + \mu(i) \tag{1}$$

$$\hat{\mathrm{F}}_p(i) = \sqrt{\frac{\sigma(i)}{\sigma_p(i)}}(F_p(i) - \mu_p(i)) + \mu(i). \tag{2}$$

For our set of images, $\vec{\sigma} = [2299, 54, 92]$ and $\vec{\mu} = [106, 118, 141]$. Examples of color adjustment on the original images in our set are shown in Figure 5. Shown in Figures 3 and 4 are the final set of all 60 color-adjusted images.

While the main purpose of this experiment was to quantify how reliable observers are in identifying images as CG or photographic, it is also important

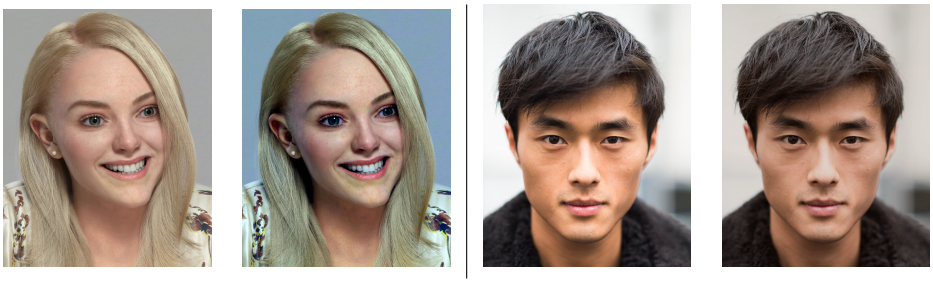

Figure 5: Original images (left) with color-adjusted counterparts (right). The pair of images on the left is an example of more overt changes, whereas the images on the right are an example of more subtle changes.

to see how degradation in quality can affect performance. This aspect of the experiment better allowed us to extend the results into a real-world context where image quality is rarely perfect.

One way in which we can degrade images is by reducing their resolution. To this effect, we reproduced the 60 images at six different resolutions (a partial example of which is shown in Figure 6). This process resulted in 360 total images on which to test observers. Because a square aspect ratio was not enforced on the images, when referring to resolutions of 100, 200, 300, etc. pixels it is implied that this is the maximum dimension of the image.

2.1.2 Psychophysical Setup

We recruited 250 observers to participate in this experiment using Amazon's Mechanical Turk, a crowd-sourcing utility that provides an experimenter with fast access to a large group of human observers.

After a Mechanical Turk user opted to participate in our experiment they were given a brief summary of the task they would be asked to perform and were then asked to consent to the terms of the experiment.

During the experiment, an observer saw all 60 images, one at a time, rendered at a randomly determined resolution. After each image was presented, the observer was asked to make a judgement as to whether the image was CG or photographic and whether the person in the image was female or male. To ensure that observers did not rush their judgment, the experimental program ignored any responses made within 3 seconds of image onset. After the delay, the observer could click a button to indicate their choice: "male/CG", "male/photographic", "female/CG", or "female/photographic".

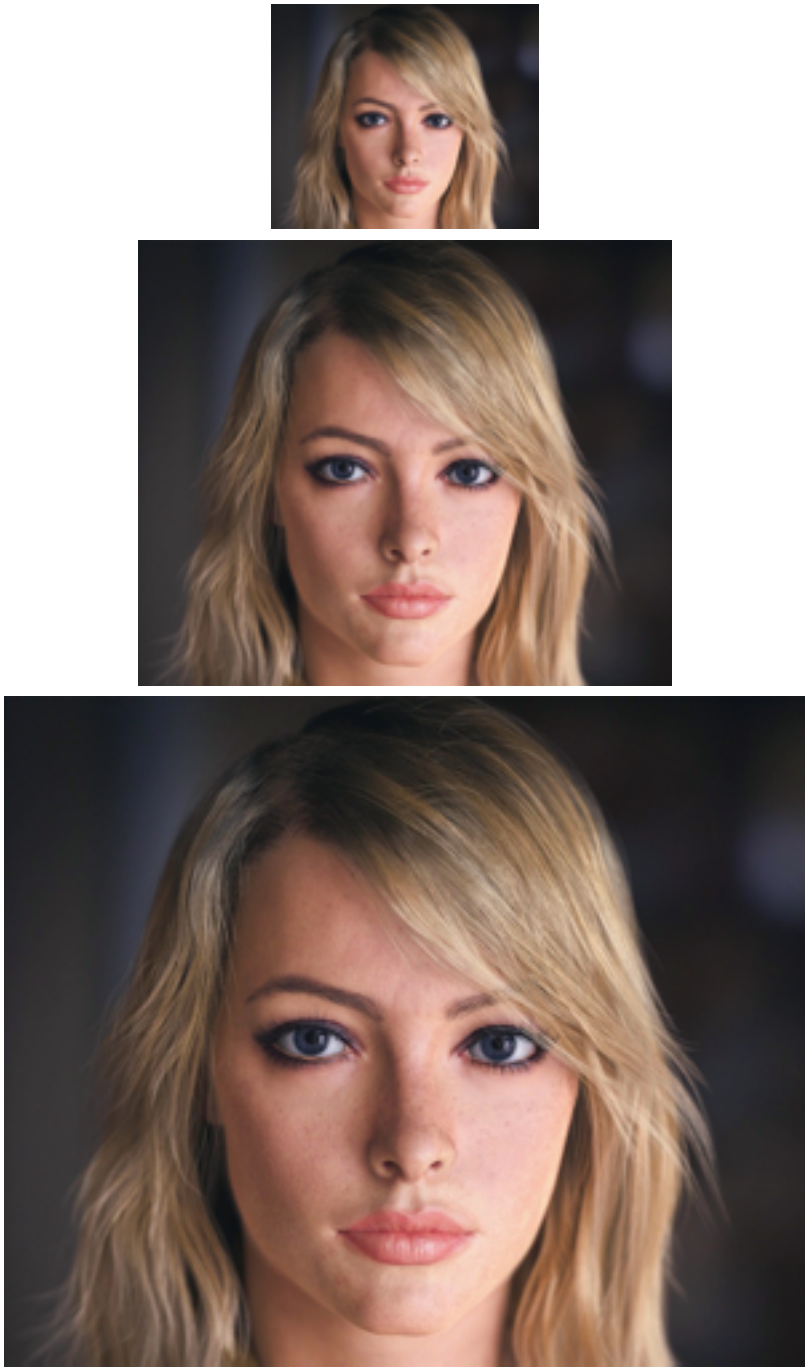

Figure 6: A sample CG image at three of the six resolutions (100, 200, and 300 pixels from top to bottom) used in the experiment.

	(a)		(b)		(c)		
Resolution	**d'**	**beta**	**d'**	**beta**	**Contrast**	**d'**	**beta**
100	1.45	2.61	1.42	1.45	**1**	0.07	0.98
200	1.57	2.34	1.68	1.32	**3**	0.41	1.03
300	1.73	2.21	1.69	1.15	**5**	0.84	1.20
400	1.67	2.40	1.82	1.18	**10**	0.98	1.55
500	1.73	2.62	1.88	1.11	**50**	1.77	2.95
600	1.68	2.60	1.91	1.33	**100**	2.03	5.02

Table 1: D' and beta as a function of resolution for experiments 1 (a), 2 (c), and 4 (b).

An observer viewed a total of 60 images, never viewing the same image twice. Each image that he/she viewed was rendered at a random resolution so that, over 60 trials, the observer saw both CG and photographic images at a variety of sizes. The order in which the 60 images were presented to the observer was randomized.

Each participant was paid $0.50 for their time. No feedback was given to participants as to how well they had performed. The observer's ability to correctly identify the sex of the figure in each image was used as a means of discarding the results of those who had rushed through the experiment.

2.2 Results

Although 250 observers participated in the experiment through Amazon's Mechanical Turk, one participant's results were discarded because that observer's accuracy in determining the sex of the faces in the images was below 95%.

Figure 7 shows the accuracy of observers in correctly identifying CG and photographic images as a function of image resolution. Note that observers seem to be very good at discriminating an image as photographic across all six resolutions. The lowest performance (90.69% accuracy at 300 pixel resolution) differs little from observers' best performance (92.25% accuracy at 500 pixel resolution), suggesting that resolution had little to no effect on correctly identifying photographic images as photographic.

In contrast, observer accuracy in identifying CG images is more variable,

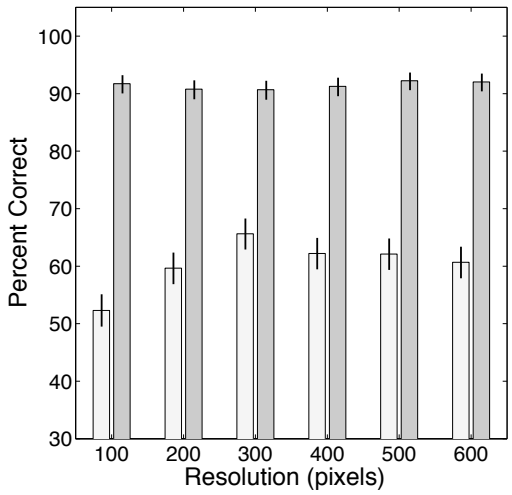

Figure 7: Observer accuracy as a function of resolution. Light grey bars indicate performance on CG images and grey bars, photographic images. Error bars are shown in black. Chance performance is 50%.

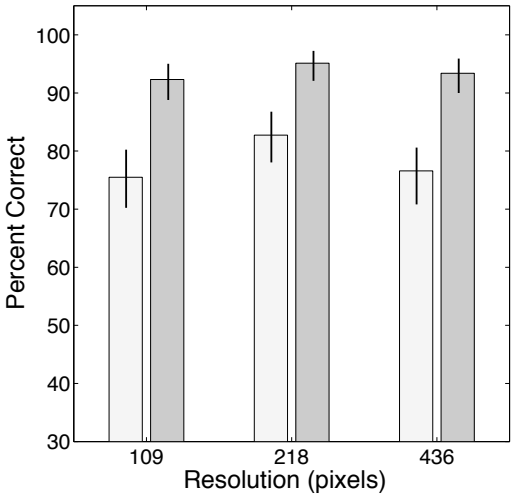

Figure 8: Observer accuracy as a function of resolution in the 2010 study [29]. All CG images used in this study were rendered between 2007 and 2010. Light grey bars indicate performance on CG images and grey bars, photographic images. Chance performance is 50%.

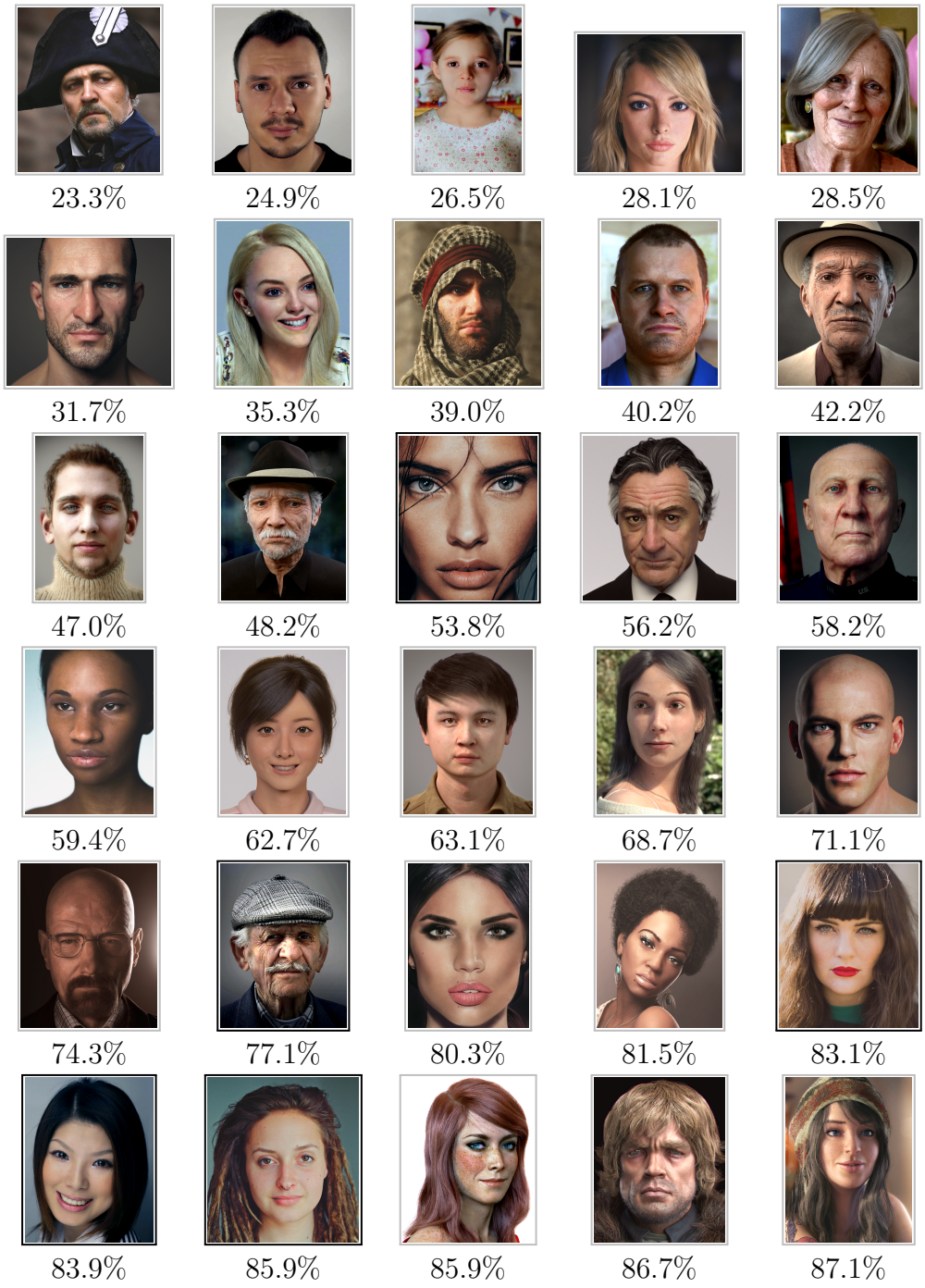

Figure 9: Images in ascending order based on accuracy averaged over all six resolutions. CG images have a grey border, photographic images have a black border. See also Figure 10.

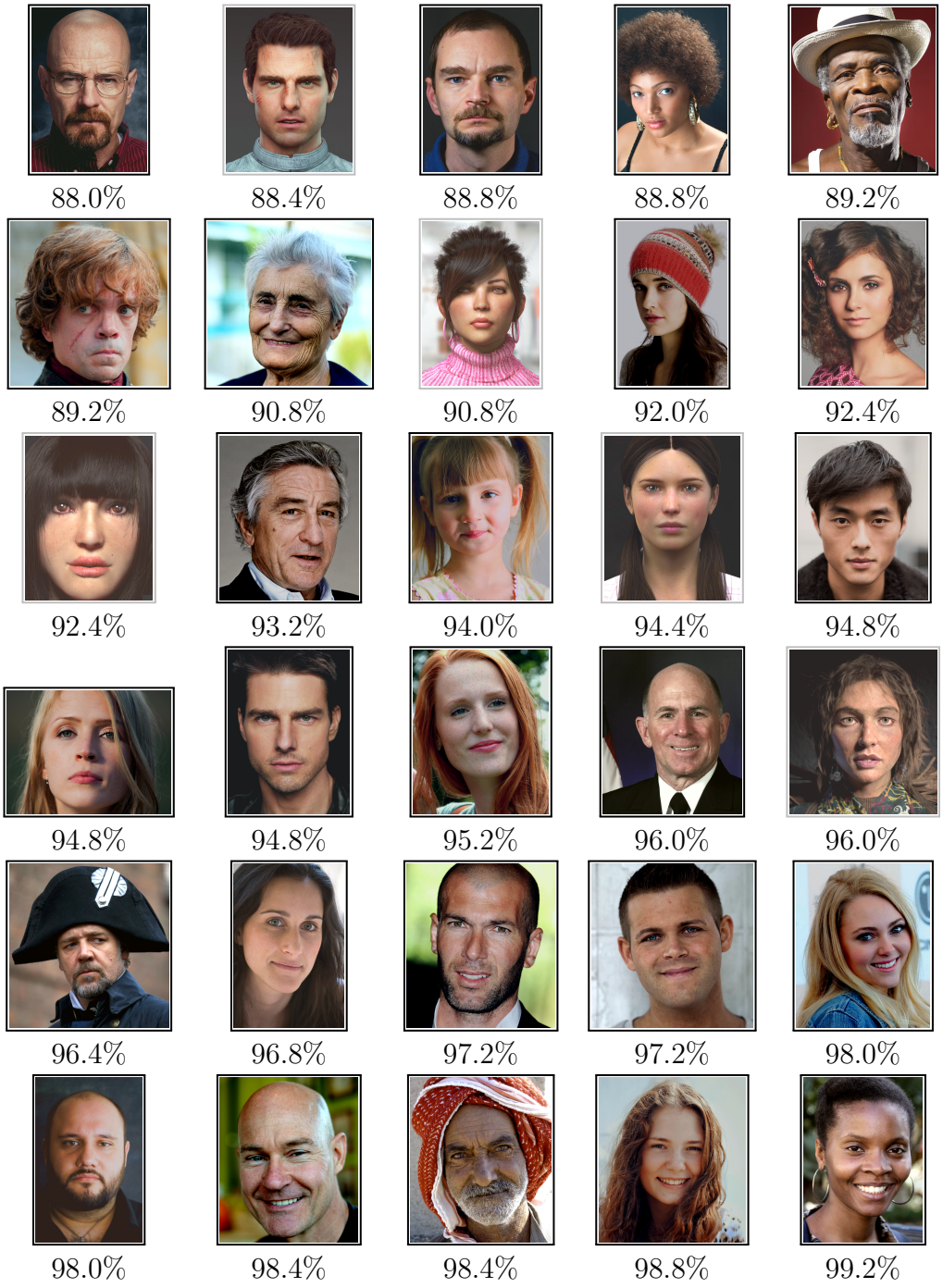

Figure 10: Images in ascending order based on accuracy averaged over all six resolutions. CG images have a grey border, photographic images have a black border. See also Figure 9.

although it seems to plateau at 400 pixels, with accuracy staying between 60.67% and 62.21%. It is clear however, that observer performance is significantly worse for CG than photographic images. This trend is most clearly illustrated at 100 pixel resolution where observer accuracy falls nearly to chance for CG images, yet is maintained at over 90% for photographic images.

Interestingly, observer accuracy for CG images does not peak at the maximum resolution, 600 pixels, but rather at the medium resolution of 300 pixels. This is somewhat puzzling as one might expect that many subtle features that are difficult to render would not be visible at a lower resolution, making these images more difficult to classify. We have yet to develop a satisfactory theory to explain this unusual result.

In Figures 9 and 10, the 60 images are ranked by their average accuracy over all six resolutions (see also Figure 11). Note that there does not appear to be any pattern with regard to gender, race, etc. that would indicate that a particular feature led to easier discrimination.

Shown in Table 1(a) are the percent correct values converted to d' and beta. These values characterize observer sensitivity and bias. A low d' across resolutions reinforces the conclusions already drawn from previous Figures: observers are not reliable in their judgements about image type. The beta for this experiment, although fairly consistent across resolutions, clearly indicates a bias to classify an image as photographic rather than CG. Is this trend due to the fact that observers are not as familiar with CG images? Or perhaps observers do not realize how photorealistic CG images have become? These questions will be addressed to some extent later in this paper.

Figures 12 and 18 separate the 60 images by type (CG/photographic) and then sort them by observer accuracy at 600 pixel resolution. The order of the images is maintained in Figures 13 through 23 which display accuracy for individual images at 100, 200, 300, 400, and 500 pixel resolutions to allow the reader to track the accuracy of one image across several resolutions.

2.3 Discussion

The purpose of this experiment was to update the 2010 study [29] which was the first attempt to quantify the reliability of the average observer at classifying an image as CG or photographic. As stated above, our experiment used many of the same methods as [29] to test observer accuracy with an

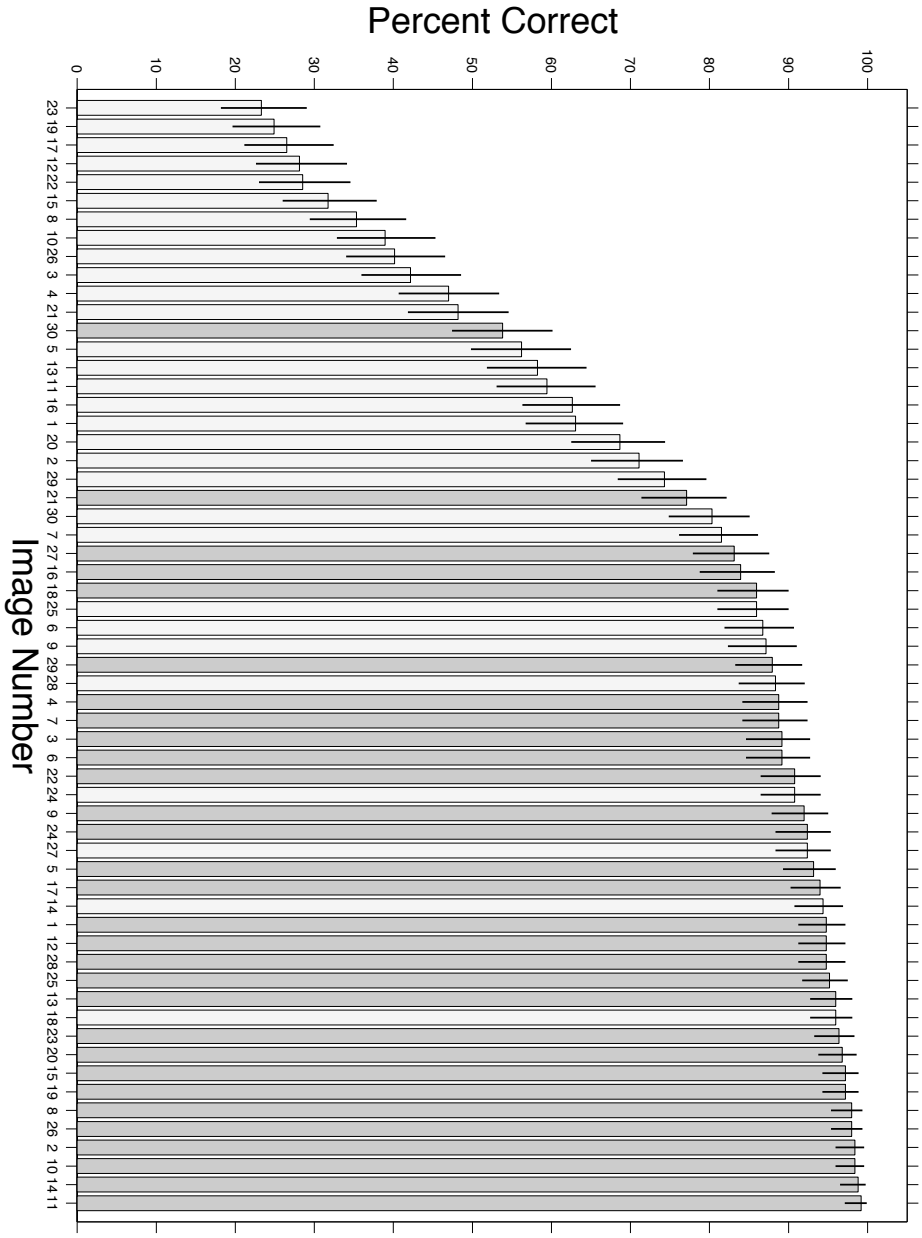

Figure 11: All images sorted by accuracy averaged over all the six resolutions. Note that the y-axis begins at 0. Light grey bars indicate performance on CG images and grey bars, photographic images. Image numbers correspond to the numbers in Figures 3 and 4.

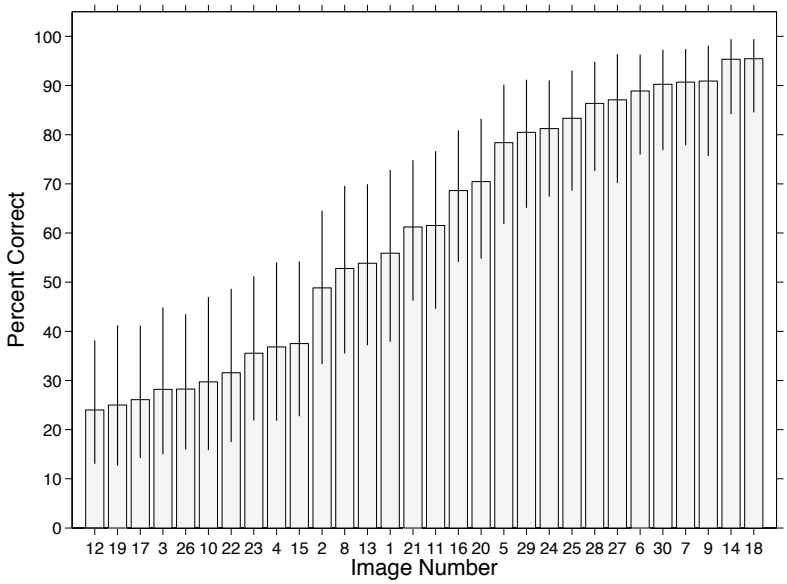

Figure 12: CG images sorted by accuracy at 600 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

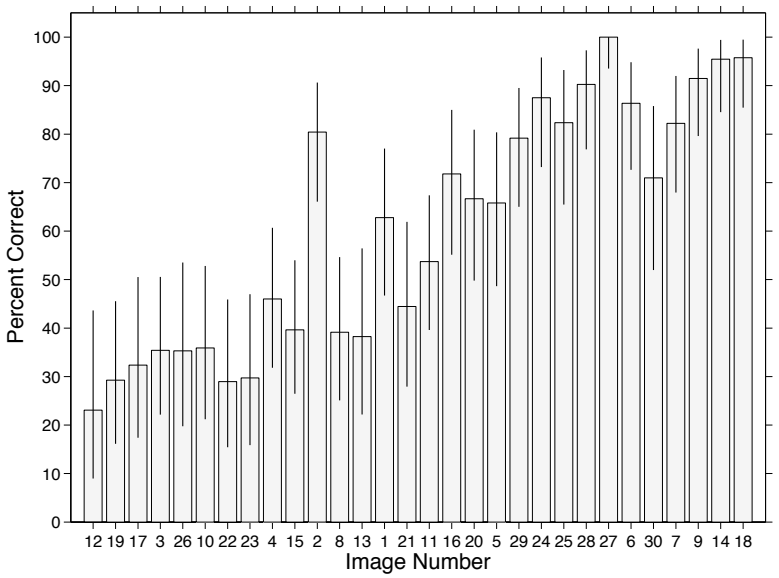

Figure 13: CG images sorted by accuracy at 500 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

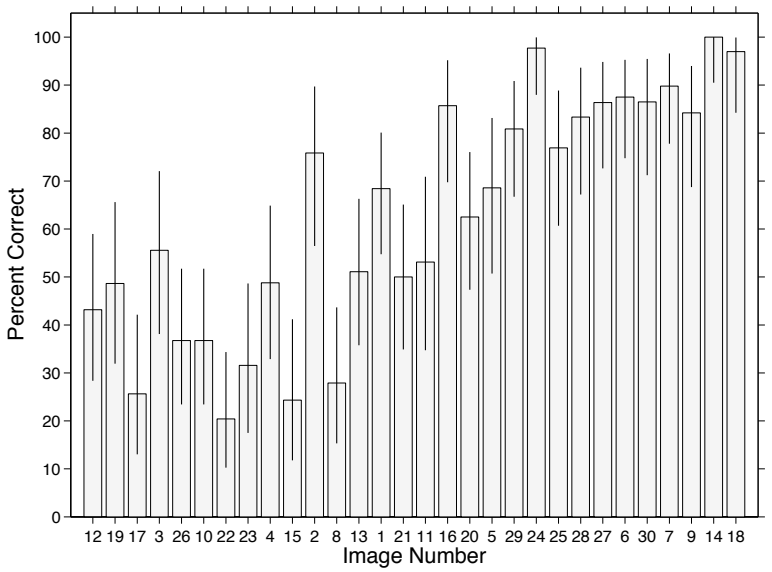

Figure 14: CG images sorted by accuracy at 400 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

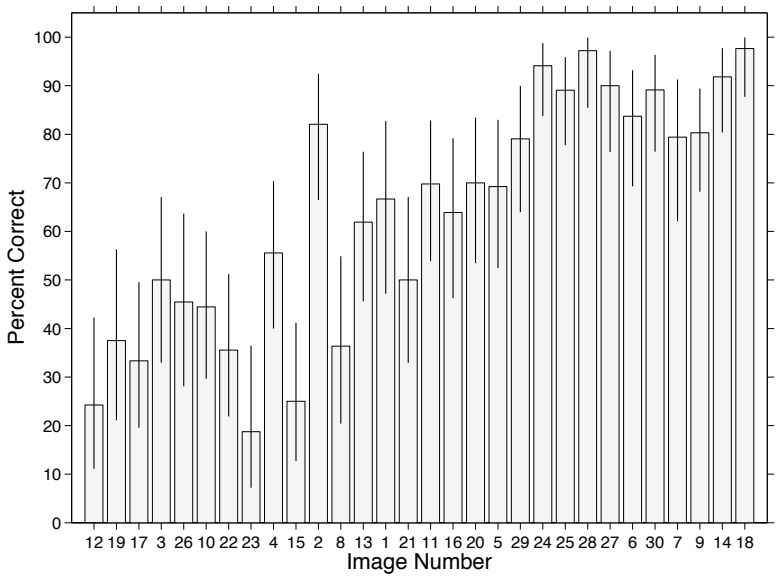

Figure 15: CG images sorted by accuracy at 300 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

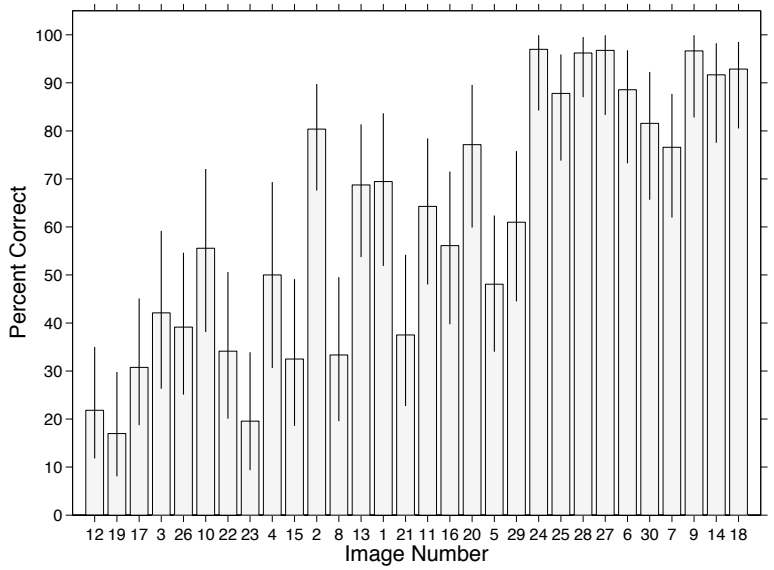

Figure 16: CG images sorted by accuracy at 200 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

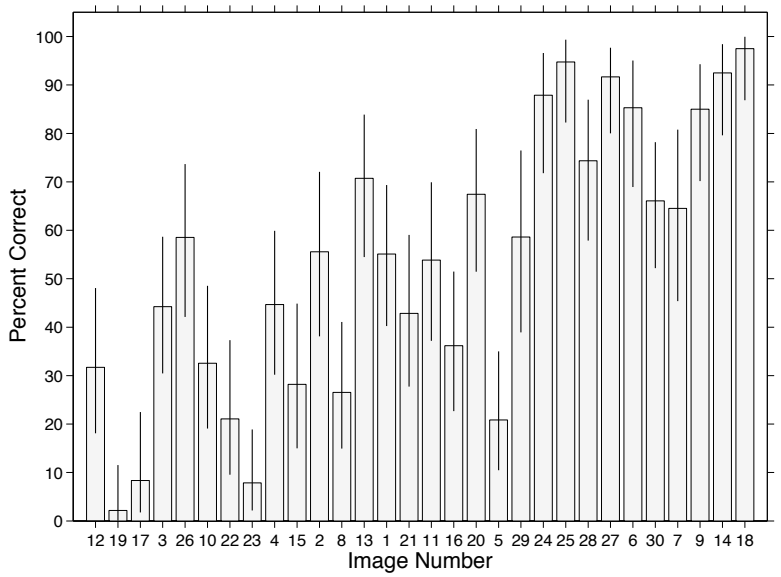

Figure 17: CG images sorted by accuracy at 100 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

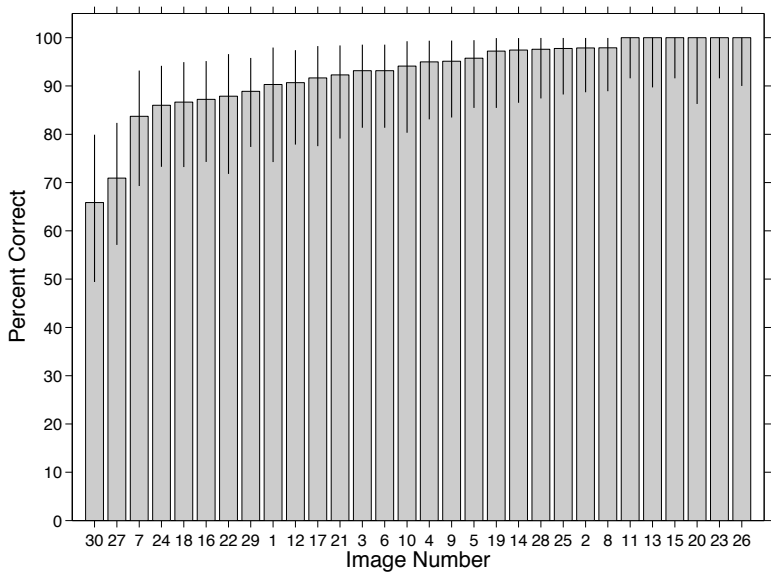

Figure 18: Photographic images sorted by accuracy at 600 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

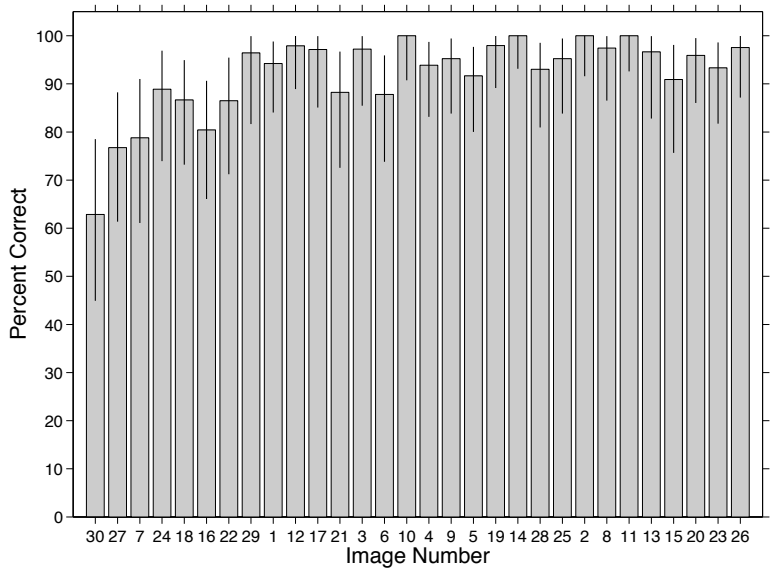

Figure 19: Photographic images sorted by accuracy at 500 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

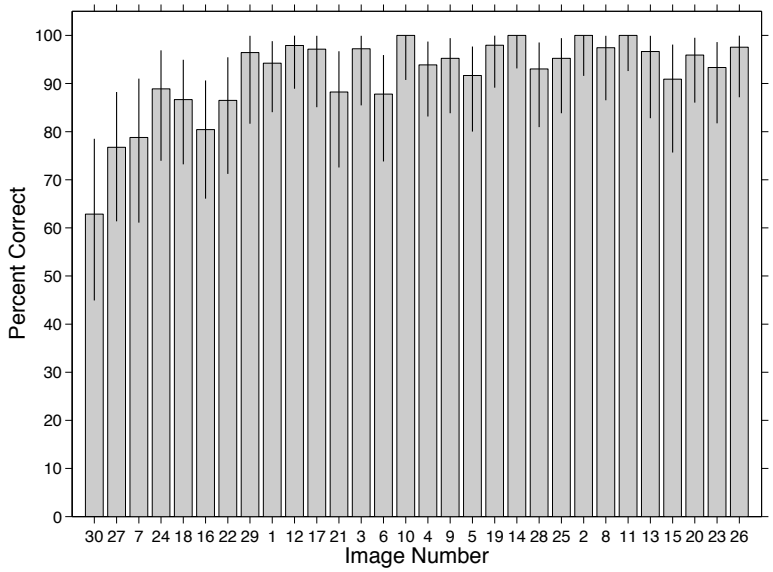

Figure 20: Photographic images sorted by accuracy at 500 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

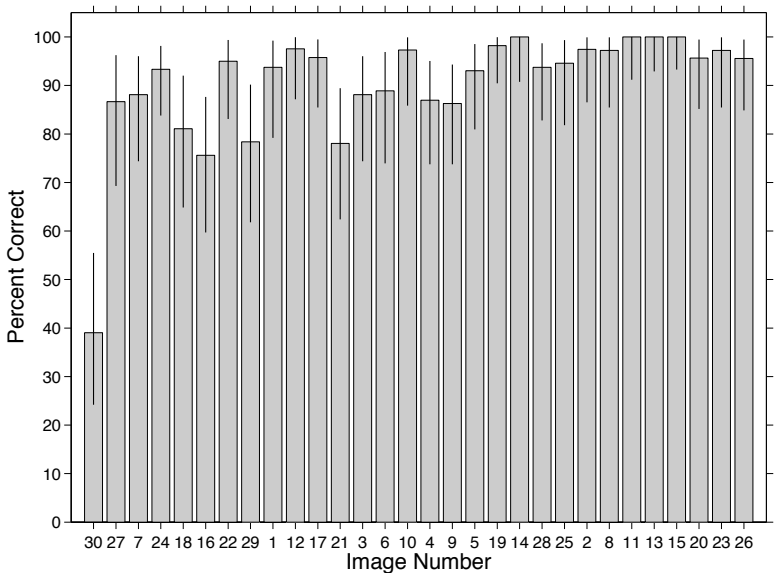

Figure 21: Photographic images sorted by accuracy at 300 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

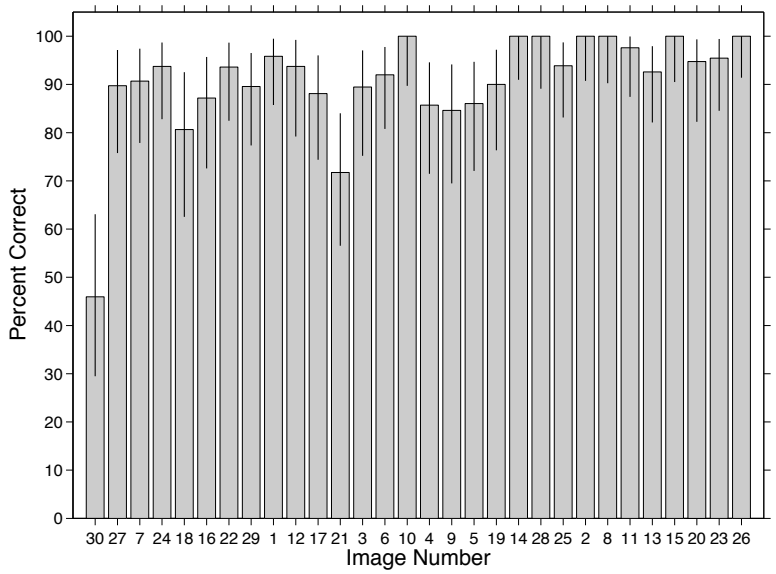

Figure 22: Photographic images sorted by accuracy at 200 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

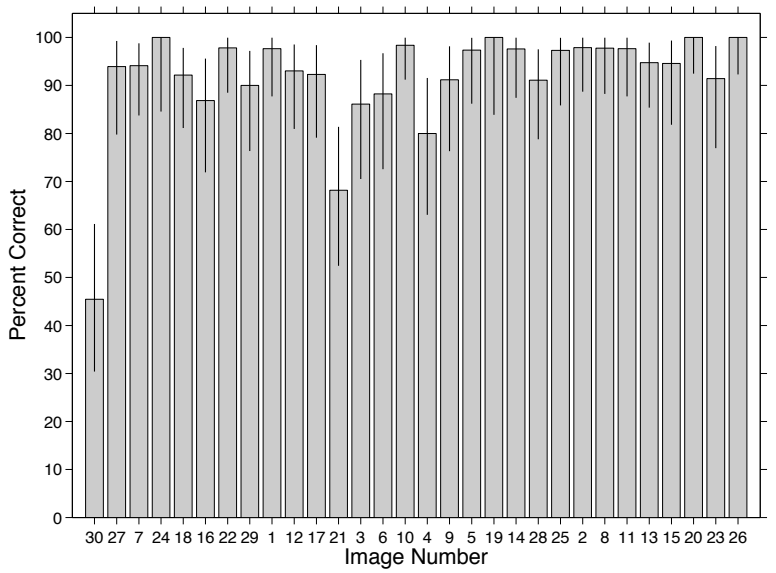

Figure 23: Photographic images sorted by accuracy at 100 pixel resolution. Note that the y-axis begins at 0. Image numbers correspond to the numbers in Figures 3 and 4.

updated collection of CG images (rendered between 2013 and 2014 instead of between 2007 and 2010).

In order to understand and quantify the progress that has been made in the field of computer graphics and how this has affected observers' judgement, it is important to compare the results of these two studies. In [29], observers were tested on images at 11, 22, 44, 109, 218, and 436 pixel resolutions. The results at 109, 218, and 436 pixel resolutions have reasonable homologues in our study (100, 200, and 400 pixel resolution, respectively). Thus, these results have been reproduced in Figure 8 so as to be easily comparable with our own results.

The first major difference to note between Figures 7 and 8 is the pronounced decrease in observer ability to identify CG images as CG from 2010 to the present day. This trend can be reasonably attributed to the fact that the images for [29] were far less photorealistic than the current CG images and therefore more easily identifiable to observers as CG.

The progress that has been made in computer graphics is truly astounding and can be seen in how easily human observers are fooled by current CG images. For example, in 2010, accuracy for CG images rendered at 109 pixels was 75.50%, yet by the time of this study, that number had dropped to near chance (52.31%) at virtually the same resolution. Also of note is that the highest accuracy for CG images in the 2010 study was over 80%, yet in 2014, observers were not able to perform above 65% accuracy for CG images.

While performance on photographic perception seems to have stayed fairly consistent over time, observer reliability in identifying images as CG has fallen considerably. This becomes most apparent in a comparison of d' and beta between the two studies. In 2010, d' stayed fairly consistent across resolutions at an average of 2.32, considerably higher than the best d' from our study (1.73). However, observer bias does not seem to have changed over time: average beta from this study and [29] are very similar (2.46 and 2.37, respectively).

3 Experiment 2: Contrast

The results from Experiment 1 demonstrated how observers' ability to distinguish CG from photographic images fall as the resolution is diminished. Another way to degrade image quality is by affecting the level of contrast. Although an image may be large in size, with the contrast affected, the contents of the image become more difficult to discern. Experiment 2 investigated the question: at what level of contrast does observers' ability to distinguish CG from photographic images become impaired?

3.1 Methods

Unless stated otherwise, the methods used in this second experiment were the same as those used in Experiment 1, Section 2.1.

3.1.1 Images

This experiment employed the same set of 60 color-adjusted CG and photographic images that were used in Experiment 1. However, instead of producing the 60 images at six different resolutions, in this experiment the 60 images were produced at 100%, 50%, 10%, 5%, 3%, and 1% of their original contrast. There was no variation in image resolution; all images were viewed at the highest resolution of 600 pixels. A contrast adjusted image, F_c, is generated from the original image, F, as follows:

$$F_c = \frac{c}{100}\left(F - \frac{1}{2}\right) + \frac{1}{2}, \tag{3}$$

where c is the percent contrast modulation and image pixel values are between $[0, 1]$. See Figure 24 for a sample CG image rendered at six contrast levels.

3.1.2 Psychophysical Setup

We recruited a total of 100 observers for this experiment. Each observer viewed a total of 60 images, one at a time, and never saw the same image twice. Each image that he/she viewed was rendered at a randomly selected contrast level (from our set of six predetermined levels) so that, over 60 trials, the observer saw both CG and photographic images rendered at a variety of

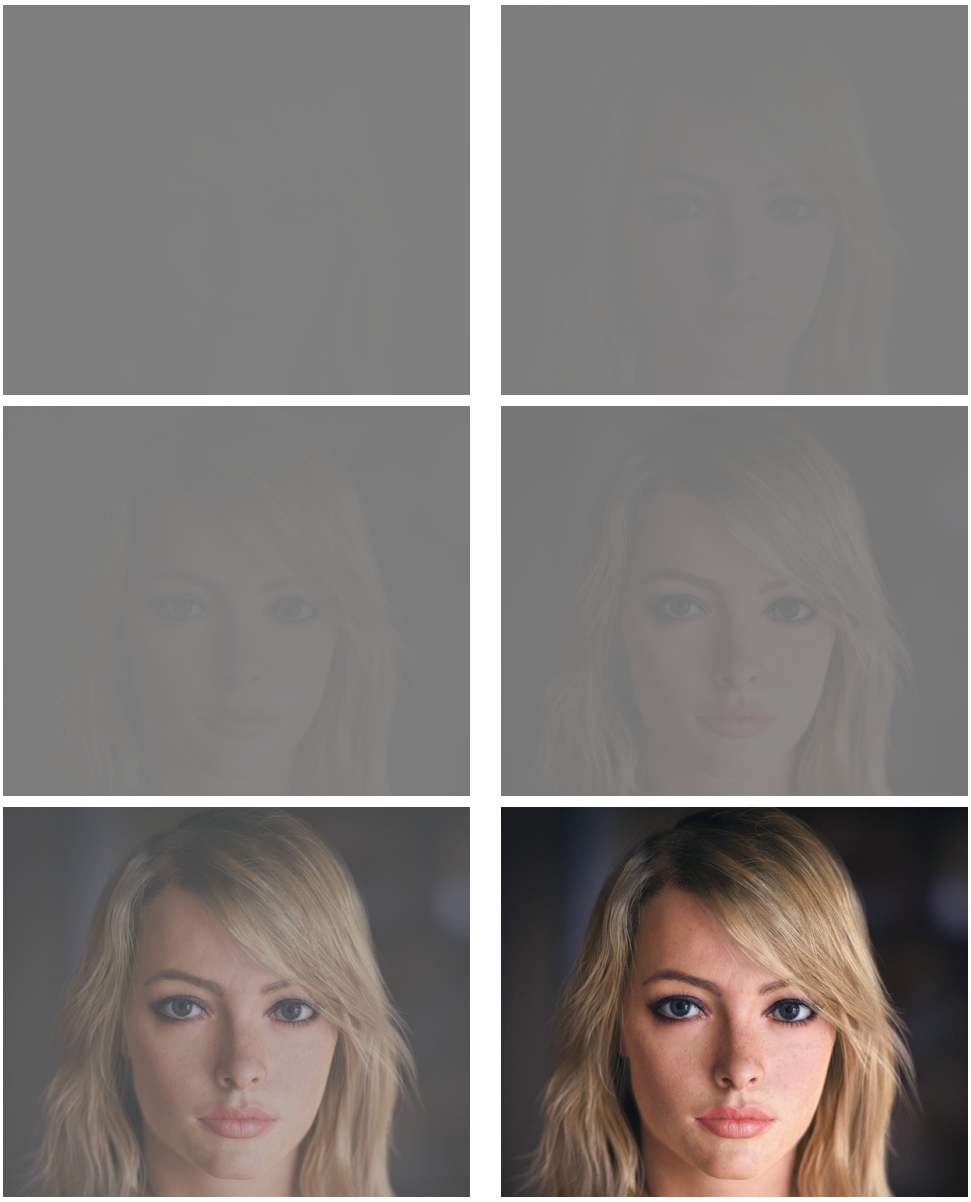

Figure 24: A sample CG image rendered at all six levels of contrast (1%, 3%, 5%, 10%, 50%, and 100%, from left to right, top to bottom).

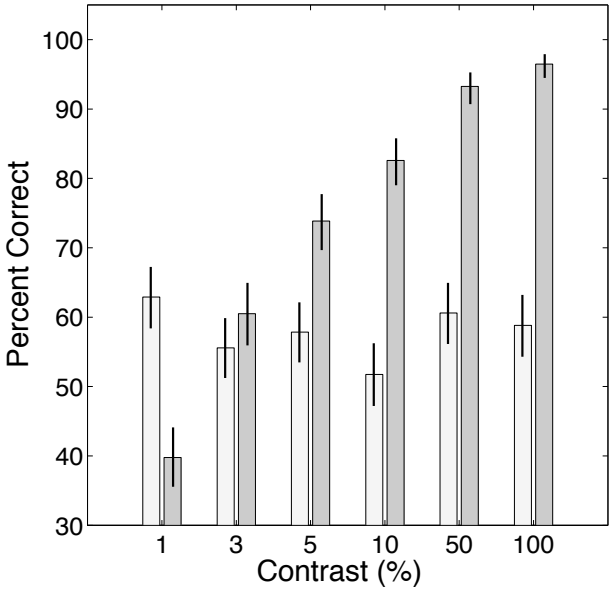

Figure 25: Observer accuracy as a function of contrast. Light grey bars indicate performance on CG images and grey bars, photographic images. Error bars are shown in black. Chance performance is 50%.

levels of contrasts. The order in which the 60 images were presented to the observer was randomized.

After each image was presented, the observer was asked to make a judgement as to whether the image was CG or photographic and whether the person was female or male. After a delay of three seconds, the observer could click a button to indicate their choice: "male/CG", "male/photographic", "female/CG", or "female/photographic".

3.2 Results

We were able to use all 100 responses collected through Mechanical Turk because observer accuracy in determining the sex of the faces in the images was above 95% for all observers.

To see observer accuracy as a function of contrast refer to Figure 25. While d', seen in Table 1(c), at 100% and 50% contrast demonstrates some observer reliability, as the level of contrast falls, d' indicates how steadily, and rapidly, user performance falls as well.

4 Experiment 3: Features

After discovering two ways (degradation in resolution and contrast) to hinder observer performance in identifying images as CG or photographic, we were interested if it was possible to *improve* performance. Our first attempt at achieving this goal was to isolate prominent features of the face in the images. If the observer concentrated only on select parts of the face, we postulated, they may better be able to discover details that indicate an image is either CG or photographic. This work was also motivated by a previous study [30] whose results indicate that eyes best communicate the animacy of a subject. Considering this, we hypothesized that isolating the eyes of an image would be the best aid to observers in forming their judgements.

4.1 Methods

Unless stated otherwise, the methods used in this experiment were the same as those used in Experiment 1, Section 2.1.

4.1.1 Images

This experiment employed the same set of 60 color-adjusted CG and photographic images that were used in Experiment 1. The purpose of this experiment was to see if it was possible to improve performance by showing observers only isolated facial features. To test this, the 60 images were rendered so that only one of five facial features (eyes, nose, mouth, cheek, or hair) was visible, as shown in Figure 26. There was no variation in image resolution; all images were viewed at the highest resolution, 600 pixels.

In order to ensure systematic selection of the location and size of a feature, we estimated a geometric warping between each face and a generic template face, Figure 27 (left). We modeled this warping using either a localized affine or 2^{nd} degree polynomial transformation, depending on the selected feature.

To create a mapping from the template face to a CG or photographic face, we specified 18 predetermined landmarks on each of the 60 images in our set as well as the template face. These 18 points are listed here and shown in Figure 27: left and right temple, top of the hairline, bottom of the chin, left and right corner of both eyes, the center of the iris in both eyes, outer left and right side of nose, top of the nose, center of the tip of the nose, bottom of the nose, left and right corner of mouth, and in between the eyebrows.

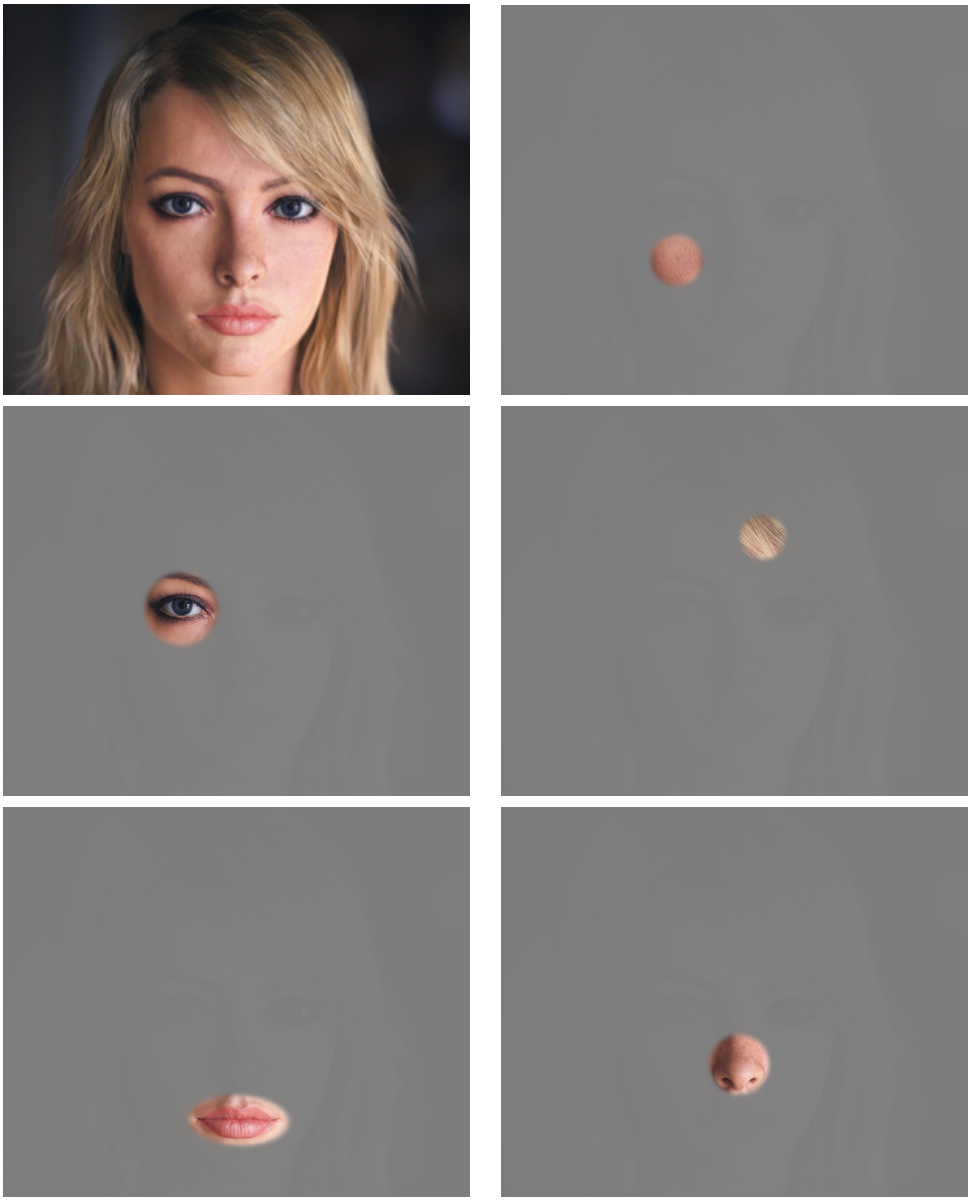

Figure 26: A sample CG image with each of the five features isolated. Left column from top to bottom: original, eye, mouth. Right column from top to bottom: cheek, hair, nose.

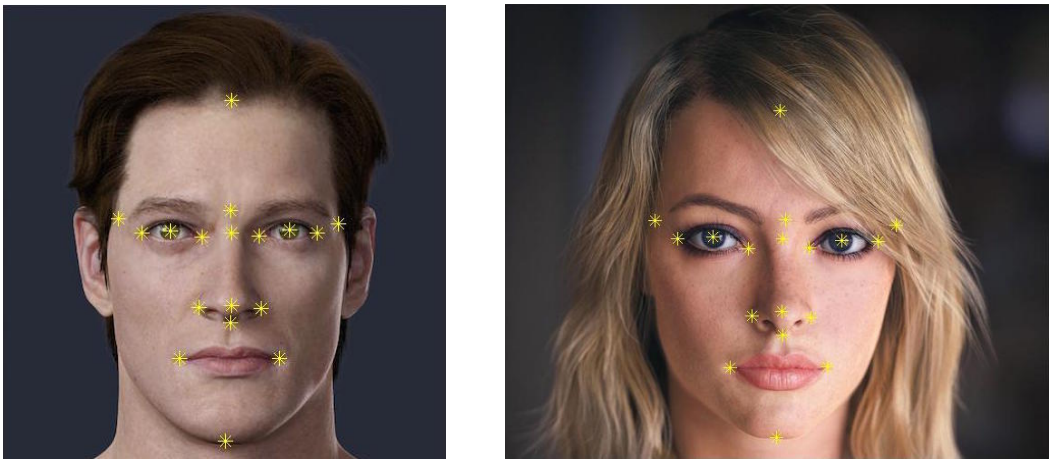

Figure 27: The template face and a sample CG image with the 18 predetermined points highlighted on each face.

The masking was done using an anti-aliased circular or elliptical region. The inside area of the circle/ellipse had 100% (unaltered) contrast. The surrounding area of the image was reduced to 2.5% of its original level of contrast. The circular/elliptical regions were rendered with an anti-aliased transition in order to avoid visual artifacts. The purpose of reducing the contrast outside of the circle/ellipse was to allow the observer to perceive the context of the entire image without allowing that information to help them determine image type. The diminished level of contrast was selected based on our findings from Experiment 2 (Figure 25), which showed that observer accuracy fell to chance at approximately 2.5%.

Note that some manual adjustments were made in the process of masking the images. However, this was only done when the programmatic mapping proved insufficient.

4.1.2 Psychophysical Setup

We recruited a total of 250 observers for this experiment. Each observer viewed a total of 60 images and never saw the same image twice. Each image that he/she viewed had one randomly chosen feature isolated (from our set of five predetermined features). The order in which the 60 images were presented to the observer was randomized as well.

After seeing each image, the observer was asked to make a judgement

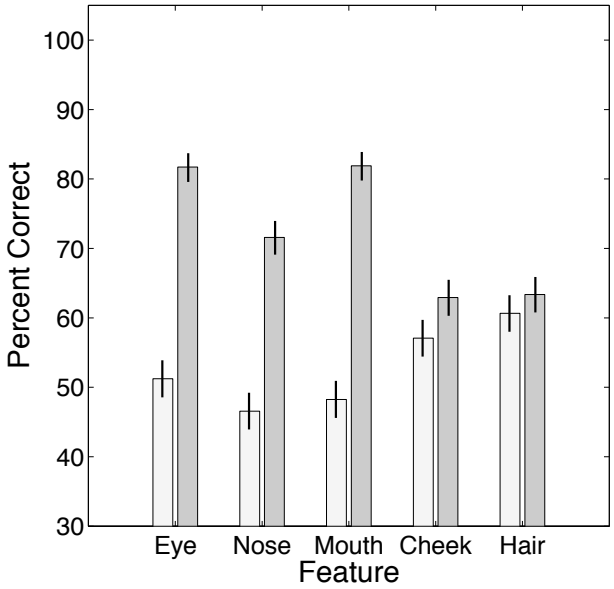

Figure 28: Observer accuracy as a function of facial feature. Light grey bars indicate performance on CG images and grey bars, photographic images. Error bars are shown in black. Chance is 50%.

as to whether the image was CG or photographic and whether the person was female or male. After a delay of three seconds, the observer could click a button to indicate their choice: "male/CG", "male/photographic", "female/CG", or "female/photographic".

4.2 Results

Although 250 observers participated in the experiment through Amazon's Mechanical Turk, twenty participants' results were discarded because their accuracy in determining the sex of the faces in the images was below 80%. This cut-off point, which is lower than the 95% cut-off used in the previous experiments, was chosen because identifying sex in this experiment was more difficult due to the low contrast of the image apart from the isolated feature.

It is clear from d' in Table 2 that observer performance for this task was very poor, regardless of which feature an observer was asked to focus on. However, observers did perform somewhat better on the mouth and eyes (with d' of 0.87 and 0.94, respectively) than they did on the other features

Feature	d'	beta
Cheek	0.51	1.04
Hair	0.61	1.02
Nose	0.49	1.17
Mouth	0.87	1.51
Eye	0.94	1.51

Table 2: D' and beta for Experiment 3.

(average d' of 0.53). The mouth and eyes also both had the highest beta of 1.51, indicating an observer bias to classify these images as photographic. Interestingly, this bias nearly disappears for the other three features, which have an average beta of 1.08.

To see observer accuracy as a function of feature refer to Figure 28.

4.3 Discussion

In Section 2.3 of Experiment 1, it was noted how unreliable observers are at classifying images as CG or photographic in comparison to the results from the 2010 study. In [29], observers were considered to be quite reliable, with an average d' across resolutions of 2.32. We concluded that, because the d' had dropped to an average of 1.64 in Experiment 1 and there was a strong observer bias, that reliability had decreased considerably. Considering these conclusions from Experiment 1, we can conclude from the even lower d' from this experiment that isolating features do not help observers in classifying images, but actually serve to impair their judgement.

5 Experiment 4: Training

Because the methods employed in Experiment 3 clearly did not help observers in their judgements as to whether an image was CG or photographic, we postulated that perhaps by taking the opposite approach (exposing observers to *more* material), we could succeed in improving their performance. By providing observers with examples of recently rendered, photorealistic images, we hoped to create a baseline familiarity with CG images across all observers. To test this idea, a short training exercise was added before the actual task (which was identical to the task in Experiment 1) wherein observers were shown images as well as the correct response for each image.

5.1 Methods

Unless stated otherwise, the methods used in this experiment were the same as those used in Experiment 1, Section 2.1.

5.1.1 Images

This experiment employed the same set of 60 color-adjusted CG and photographic images (rendered at the same six resolutions) that were used in Experiment 1.

An additional 10 CG and 10 photographic matches were collected to serve as training images for this experiment. All CG images were downloaded from the following popular CG websites: `www.cgsociety.org` and `www.3dtotal.com`. The content and context of these websites virtually guaranteed that these images were computer generated in nature.

The primary aim in choosing the CG training images was to select the most photorealistic images possible so that observers would be exposed to high-quality CG before the actual task began. While most of the training images that were chosen were produced after the completion of Experiment 1 (and were therefore exemplars of the most recently rendered CG images), our preference for quality over render date forced us to use some older images (although all training images were rendered between 2013 and 2014). The CG images that were chosen were meant to include the diversity (in terms of age, race, etc.) that observers would see in the CG images in the actual task. The final set of 10 CG images is composed of 5 male and 5 female

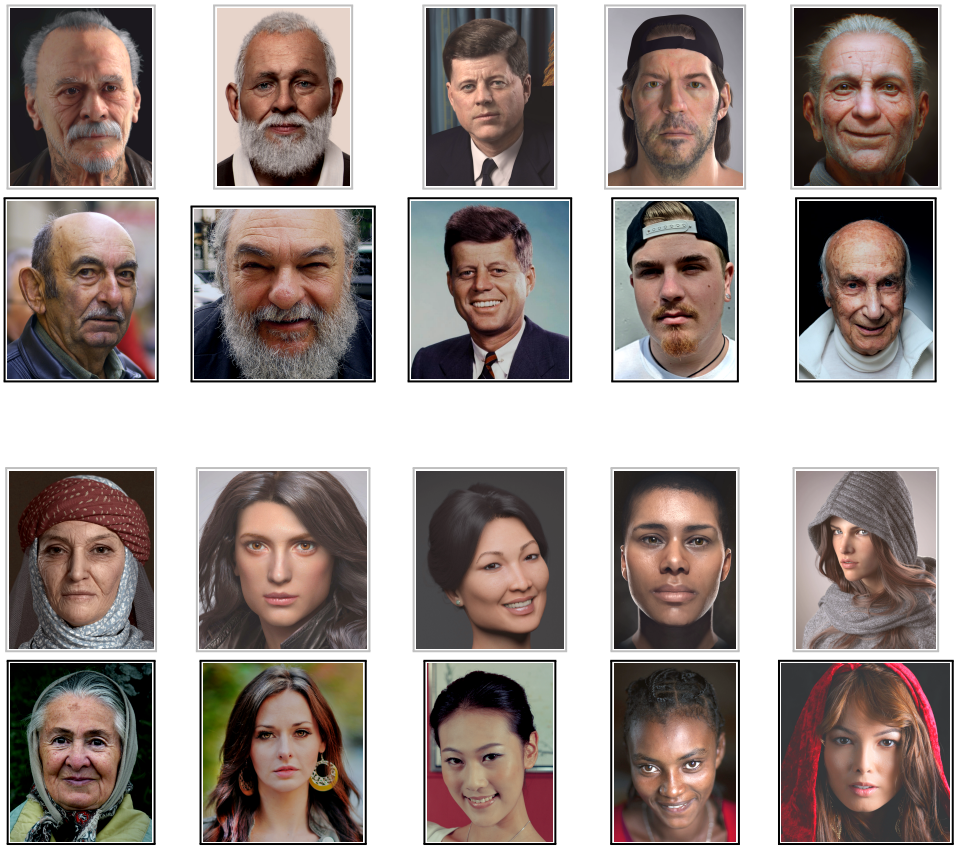

Figure 29: CG training images (top, with grey border) paired with their photographic matches (bottom, with black border).

faces. These images all followed the same set of qualifications laid out for the original 30 CG images in Section 2.1.1.

Ten photographic "matches" were also chosen. These photographic counterparts were selected in the same manner as detailed in Section 2.1.1. The photographic images that were chosen were downloaded from several websites (although the majority were found on `www.flickr.com`). The content and context of these websites virtually guaranteed that these images were photographic in nature.

As in Experiment 1, the backgrounds of all 20 training images were reviewed for systematic low-level cues. Additionally, the brightness and contrast of the 20 training images were adjusted in the same manner as in Experiment 1. Referring back to Equations (1) and (2), for this set of training

images, $\vec{\sigma} = [1876, 41, 76]$ and $\vec{\mu} = [115, 118, 143]$. The final set of color-adjusted training images are shown in Figure 29.

5.1.2 Psychophysical Setup

We recruited a total of 250 observers to participate in this experiment using Amazon's Mechanical Turk. The experiment involved two distinct parts: (1) the training session at the beginning of the experiment and (2) the actual task, identical to the one performed in Experiment 1.

After an observer consented to the terms of the experiment, they were informed that they would first observe a set of training images for which they would be told the correct response. In this portion of the experiment, observers saw a total of 20 images, one at a time, in the same format that they would encounter in the actual task. However all images were rendered at 600 pixel resolution and, beneath each image, the correct response (in terms of sex and image type) was presented to the user. An observer never saw the same training image twice and the order in which the images were presented to each observer was randomized.

Upon completing the training, observers were informed that they were about to begin the actual task, where they would not be told the correct response for each image and would have to make their choice based on their own judgements. Each observer then viewed a total of 60 images, never viewing the same image twice. Each image that he/she viewed was rendered at a random resolution so that, over 60 trials, the observer saw both CG and photographic images at a variety of sizes. The order in which the 60 images were presented to the observer was randomized.

5.2 Results

Although 250 observers participated in the experiment through Amazon's Mechanical Turk, three participants' results were discarded because their accuracy in determining the sex of the faces in the images was below 95%.

Figure 30 shows observer accuracy as a function of resolution after observers underwent training. It is clear that observers are fairly reliable in identifying photographic images regardless of resolution: accuracy over all six resolutions varies between a maximum of 87.21% and a minimum of 79.11%.

In contrast, there is more variation in observer accuracy in classifying CG images across resolutions with a maximum accuracy of 80.41% at a resolution

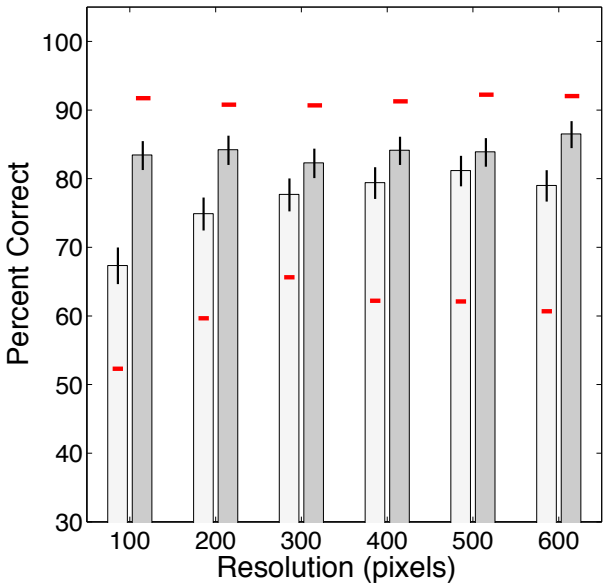

Figure 30: Observer accuracy as a function of resolution after training. Light grey bars indicate performance on CG images and grey bars, photographic images. Error bars are shown in black. Performance from Experiment 1 (without training) is shown by the horizontal red bars. Chance performance is 50%.

of 500 pixels and a minimum of 67.43% at a resolution of 100 pixels. With the exception of images rendered at 600 pixels, as resolution is reduced, observers become less reliable in correctly classifying images as CG.

The aforementioned trends cummulate to produce d' and beta that can be seen in Table 1(b). At 600 pixels, d' peaks at 1.91, after which it falls steadily (although slowly) as resolution decreases. Interestingly, beta stays relatively close to 1.00 across almost all resolutions (especially 300, 400, and 500 pixel resolutions), indicating that, although a bias exists towards photographic images, it is relatively small. This bias is most noticeable at the 100 pixel resolution, but this is also the resolution at which observers are the most inaccurate.

	Without Training				With Training			
	d'		**beta**		**d'**		**beta**	
Resolution	first	second	first	second	first	second	first	second
100	1.35	1.55	2.44	2.83	1.47	1.37	1.49	1.41
200	1.51	1.63	2.27	2.41	1.59	1.77	1.34	1.31
300	1.79	1.66	2.33	2.11	1.59	1.81	1.26	1.02
400	1.76	1.58	2.74	2.13	1.81	1.84	1.15	1.20
500	1.71	1.75	3.01	2.29	1.97	1.79	1.31	0.94
600	1.67	1.70	2.75	2.46	1.99	1.84	1.36	1.31

Table 3: D' and beta as a function of resolution for Experiment 1 (without training) and Experiment 4 (with training), comparing performance between the first and second halves of both experiments.

5.3 Discussion

In Figure 30, it is immediately clear that, with training, observers have improved in performance across all six resolutions. Maximum observer accuracy for CG images without training, 65.63%, is still less than the *minimum* accuracy for CG images after participants were trained (67.43%). To further emphasize this point, the difference between the maximum accuracy without training and with training is nearly 15 percentage points.

Note also that bias is reduced, accounting for the overall drop in percentage correct for photographic images. The average beta without training is 2.46, much higher than the average beta with training, which is 1.25. It seems that after observers have been exposed to training, their bias towards classifying images as photographic drops dramatically. Additionally, there seems to be a slight improvement in d' from an average of 1.64 without training to an average of 1.73 with training. The largest improvement in d' can be seen for the highest resolution images, suggesting that, when presented with a high-quality image in which they can presumably differentiate details more clearly, observers who participated in training perform consistently better than those who did not receive training.

Although it seems like training helped to improve observer performance, it is important to determine whether the same improvement can be seen without the feedback that was given to observers in training. To control against the effect of feedback, we compared d' and beta for observers in the

first half of Experiments 1 and 4 to d' and beta for the second half of the experiments.

In Table 3, there is no consistent improvement in d' from the first half of Experiment 1 to the second half of Experiment 1. In fact, the average d' for the first half of the experiment (1.63) is virtually the same as that of the second half of the experiment (1.64). Similarly, the beta for the first and second half of the experiment show no consistent trend.

The same statements can be made for Experiment 4: the average d' for the first half of the experiment (1.74) is identical to that of the second half of the experiment. In general, there is a slight decrease in beta from the first half to the second half of the experiment, but the overall average values do not differ greatly (1.32 and 1.20, respectively).

Why didn't observers improve on the second half of Experiment 1 after they had already been exposed to 30 CG and photographic images? Given the improvement we see in observers after a training in which they were informed of the correct responses, we can infer that feedback is the key to enhancing observers' learning and, in turn, their performance on the subsequent task.

Why did observers not improve on the second half of Experiment 4? The lack of change in d' from the first half to the second half of the experiment and the fact that these d' were consistently higher than those from Experiment 1 suggests that the training that observers underwent at the beginning affected observer accuracy *overall* and did not help it improve over time. An interesting corollary for future study would be investigating whether length and the resolution of the images in the training effects the amount of improvement made by observers.

6 Conclusion

From its inception in 1960, computer graphics technology has progressed quickly from simple 3D models to complex, photorealistic recreations of the human body. Concurrently, lawmakers and courts in the United States have struggled to define what is "obscene", what is illegal, and what is protected under the First Amendment with regards to child pornography. Nonetheless, *Ashcroft v. Free Speech Coalition* established that indeed there is a difference in how the law deals with photographic and CG images of child pornography and thus it is essential to have a reliable method of distinguishing the two image types.

While a reliable and consistent computational method for distinguishing image type may become a viable possibility in the future, its strength remains to be seen as of this writing. With these methods unavailable, we turned to testing the reliability of the human visual system in Experiment 1 by quantifying how good observers are in distinguishing CG from photographic images.

The seminal study [29] on this same topic, using CG images rendered between 2007 and 2010, concluded that observers are, in general, fairly reliable in performing this task. However, Experiment 1 showed that, when tested on CG images that were rendered between 2013 and 2014, observer performance fell considerably, from a maximum of over 80% accuracy in 2010 to 65% in 2015. Based on this poor observer performance, we conclude that observers are no longer reliable in their judgements regarding image type.

Experiments 3 and 4 showcased two very different results of attempting to improve observer performance. The results of Experiment 3 showed that isolating one feature of the face like the mouth or nose actually hurts observer performance in classifying images as CG or photographic.

However, the impact of a possible design flaw should be considered for Experiment 3 in that attempts to isolate a single feature did not always include *only* the desired feature (often a stray hair, piece of skin, etc. were included in the masking of a particular feature). However, removing these intrusions completely would require increased manual masking of the features and it is unclear whether this had a significant impact on observer performance.

Without the results from Experiment 4, one would have had to conclude that observers can no longer be trusted to reason about whether an image is CG or photographic. In a world where the line between photographic and CG

is becoming increasingly blurry, one might begin to wonder if our eyes have failed us completely in making a judgement at which we were once skilled. However, the results from Experiment 4 show that, with training, observer performance can be improved. In addition, training observers before the task helps to remove the strong bias towards classifying an image as photographic that was present without training. In further examining how exactly training helped observers in the task, the results from Experiments 1 and 4 were divided temporally. In neither experiment did observer performance improve from the first half to the second half of the task, indicating that training *with feedback* is the key to improving performance overall.

All of the experiments performed for this study utilized Amazon's Mechanical Turk. While this allowed us easy and quick access to a large group of observers, some sacrifices were made in choosing to use this utility. The largest concession was that the experiments were in many ways uncontrolled. For example, given that observers performed this task in the comfort of their own homes, it was impossible to dictate the type of computer the images were viewed on, the resolution of the observers' monitor, the physical environment in which observers viewed the images, etc.

While not controlling for these environmental inconsistencies can be seen as a limitation on our results, one can also view them as a feature. In a courtroom, for example, it is impossible to predict the quality of the image on which a jury may be asked to make a judgement and it will likely not be perfect (variations in size, contrast, parts of the face visible, etc. are inevitable). In this way, allowing observers in our experiment to view images in different environmental conditions in many ways mimics real-world situations in which they may be asked to make this judgement. If this is true, then our results are perhaps a more realistic predictor of an observer's accuracy than an experiment in which the environment was completely controlled.

In considering the observers who participated in these experiments, one can reasonably infer that they represent a lower-bound on overall observer performance. Observers were unmotivated to do well (their cash reward was given regardless of performance), they viewed each image on average between five and six seconds, and presumably had no expert knowledge of CGI before performing this task. In a courtroom setting, a jury would presumably be able to view an image for a much longer period of time and see more of a full-body image (rather than just the face) which is harder to render photorealistically in CG.

If it has been demonstrated that observers can improve their performance

in distinguishing CG and photographic images through a very small amount of training, the question then becomes can we improve their performance even further? Future research should investigate elongating the length of observer training and providing different forms of feedback to build upon these results.

The most useful application of this work would be in standardizing a training program for law enforcement officials and jury members before they are asked to make judgements as to image type in future child pornography cases. This task is going to become increasingly difficult as CG technology continues to improve. It is therefore imperative that we give observers (especially those involved in important legal cases) all of the tools that are possible to enable them to make an informed decision.

7 Acknowledgements

I would like to thank Professor Hany Farid, without whose guidance, this research would not have reached the interesting and thoughtful conclusions that it did. I was lucky that Professor Farid is such an expert in this field and was able to act as the best mentor, guiding my research when I was unsure how to proceed, and encouraging me to try out my own ideas as well.

References

[1] K. Sathyanarayana and G.V.V. Kumar. Evolution of computer graphics and its impact on engineering product development. In *Fifth International Conference on Computer Graphics, Imaging and Visualisation*, pages 32–37, 2008.

[2] The story of computer graphics. *Computer Graphics World*, 22(8):64, 1999.

[3] Ben F. Laposky. Oscillons: Electronic abstractions. *Leonardo*, pages 345–354, 1969.

[4] Charles Csuri and James Shaffer. Art, computers and mathematics. In *Proceedings of the Fall Joint Computer Conference, Part II*, pages 1293–1298. ACM, 1968.

[5] A Michael Noll. The beginnings of computer art in the United States: A memoir. *Leonardo*, pages 39–44, 1994.

[6] Frederick Ira Parke. A parametric model for human faces. Technical report, University of Utah, 1974.

[7] Cartoon character made by computer puts students at head of animation. *The Citizen*, 1985.

[8] *New York v. Ferber*, 458 U.S. 747, 1982.

[9] Child Pornography Prevention Act (CPPA), 1996.

[10] *Ashcroft v. Free Speech Coalition*, 535 U.S. 234, 2002.

[11] Prosecutorial Remedies and Other Tools to end the Exploitation of Children Today (PROTECT) Act, 2003.

[12] *Miller v. California*, 413 U.S. 15, 1973.

[13] *United States v. Dwight Whorley*, 550 F. 3d 326, 4th Cir. 2008.

[14] Sintayehu Dehnie, H.T. Sencar, and S Memon. Digital image forensics for identifying computer generated and digital camera images. In *IEEE International Conference on Image Processing*, pages 2313–2316, 2006.

[15] Ying Wang and Pierre Moulin. On discrimination between photorealistic and photographic images. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, 2006.

[16] Nitin Khanna, G. T. C. Chiu, Jan P Allebach, and Edward J Delp. Forensic techniques for classifying scanner, computer generated and digital camera images. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1653–1656, 2008.

[17] Ahmet Emir Dirik, Sevinc Bayram, Husrev T Sencar, and Nasir Memon. New features to identify computer generated images. In *IEEE International Conference on Image Processing*, volume 4, 2007.

[18] Husrev T Sencar and Nasir Memon. Overview of state-of-the-art in digital image forensics. *Algorithms, Architectures and Information Systems Security*, 3:325–348, 2008.

[19] Chen Wen, Q Shi Yun, and Xuan Guorong. Identifying computer graphics using hsv color model. In *The 22nd IEEE International Conference*, 2007.

[20] Siwei Lyu and Hany Farid. How realistic is photorealistic? *Signal Processing, IEEE Transactions on*, 53(2):845–850, 2005.

[21] Andrew C Gallagher and Tsuhan Chen. Image authentication by detecting traces of demosaicing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[22] Tian-Tsong Ng, Shih-Fu Chang, Jessie Hsu, Lexing Xie, and Mao-Pei Tsui. Physics-motivated features for distinguishing photographic images and computer graphics. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 239–248, 2005.

[23] J. F. Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[24] Duc-Tien Dang-Nguyen, Giulia Boato, and Francesco G. B. De Natale. Discrimination between computer generated and natural human faces based on asymmetry information. In *Proceedings of the IEEE 20th European Signal Processing Conference*, pages 1234–1238, 2012.

[25] Duc-Tien Dang-Nguyen, Giulia Boato, and Francesco G. B. De Natale. Identify computer generated characters by analysing facial expressions variation. In *WIFS*, pages 252–257, 2012.

[26] C. Neil Macrae and Susanne Quadflieg. *Perceiving People*. John Wiley and Sons, Inc., 2010.

[27] Randolph Blake and Maggie Shiffrar. Perception of human motion. *Annual Review of Psychology*, 58:47–73, 2007.

[28] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: 20 results all computer vision researchers should know about. *Department of Brain and Cognitive Sciences Massachusetts Institute of Technology*, 2005.

[29] Hany Farid and Mary J Bravo. Perceptual discrimination of computer generated and photographic faces. *Digital Investigation*, 8(3):226–235, 2012.

[30] Christine E Looser and Thalia Wheatley. The tipping point of animacy how, when, and where we perceive life in a face. *Psychological science*, 21(12):1854–1862, 2010.