PHYSIOLOGICALLY-BASED DETECTION OF COMPUTER GENERATED FACES IN VIDEO

V. Conotter, E. Bodnari, G. Boato *

Department of Information Engineering
and Computer Science
University of Trento, Trento (ITALY)

H. Farid †

Dartmouth College
Department of Computer Science
Hanover NH 03755 (USA)

ABSTRACT

We describe a new forensic technique for distinguishing between computer generated and human faces in video. This technique identifies tiny fluctuations in the appearance of a face that result from changes in blood flow. Because these changes result from the human pulse, they are unlikely to be found in computer generated imagery. We use the absence or presence of this physiological signal to distinguish computer generated from human faces.

Index Terms— Video Forensics, CGI, Photo Realism

1. INTRODUCTION

The photo realism of still and animated computer generated (CG) characters continues to improve thanks to the development of increasingly more powerful 3-D rendering software and hardware. When trying to distinguish the real from the fake, such CG characters pose significant challenges to the forensics community.

Over the past decade, some progress has been made in developing forensic techniques to discriminate CG from human characters. Most of these techniques exploit regularities in some low- to mid-level statistical features extracted from CG and natural images. The first such approaches to this problem used statistical features extracted from the wavelet domain [1, 2]. Related techniques leveraged sensor noise [3, 4], demosaicing artifacts [5], chromatic aberrations [6], geometric- and physics-based image features [7], and color compatibility [8]. More recently, non-statistically based techniques have been proposed which exploit facial asymmetries [9] and repetitive patterns of facial expressions [10].

These previous techniques were developed to operate on static images, and although they could be applied to video, they would not take advantage of the rich temporal data inherent to video. We describe a complementary and physiologically-inspired forensic technique for discriminating CG from human faces in a video. This technique directly measures differences in facial color that result from the human pulse. Such temporal variations are nearly invisible to the human eye, but can be revealed with "video magnification" [11]. This technique has the advantage that it does not require extensive data collection common to statistically-based techniques and it exploits a naturally and universally common signal to all humans – a pulse.

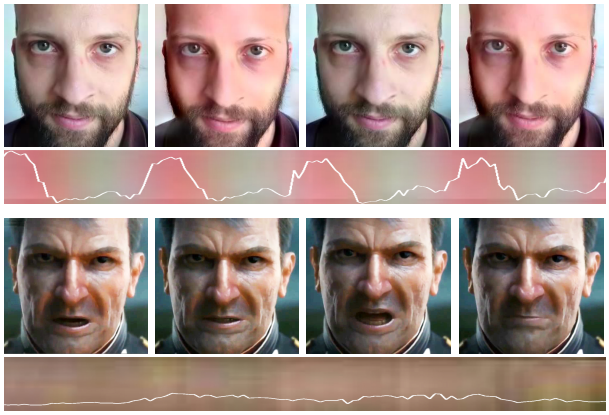

Fig. 1. Shown in the top row is a visualization of the change in blood flow due to the human pulse. Also shown is a space (vertical axis) and time (horizontal axis) plot of the color changes in the person's face, revealing the presence of a periodic pulse. No such physiological signal is present in the CG face shown in the bottom row. [CG video downloaded from: youtu.be/WO9W56KcCb8].

2. METHODS

In order to distinguish between computer generated and human faces, we propose to determine the absence or presence of a physiological signal that results from the human pulse. This signal manifests itself by changes in facial color which are magnified to be more visible, Fig. 1.

With a standard video as input, we manually identify a face on the first frame and then automatically track facial features over the length of the video. These tracked features are used to automatically fit a generic 3-D head model to the face on each frame. The head pose on each frame is then aligned to a canonical viewpoint from which the appearance of a patch of skin (the forehead or cheek) is extracted. This patch is sub-

*This work was partially supported by the ICT COST Action IC1105.
†This work was supported by a grant from the National Science Foundation (CNS-0708209).

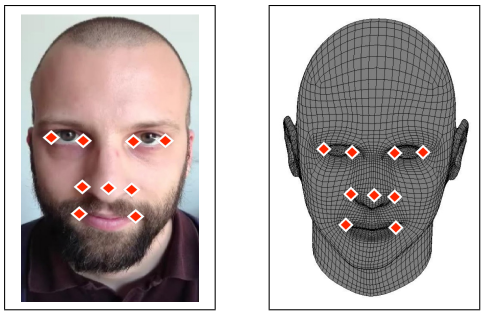

Fig. 2. Shown are nine features used for facial tracking, and an aligned 3-D model.

jected to "video magnification" [11] to enhance the desired physiological signal. The absence or presence of this physiological signal resulting from changes in blood flow is used to classify the person as computer generated or human. Each of these system components is described below in more detail.

2.1. Head tracking

We use a standard feature tracking approach to track a face over the length of the video. To begin, an analyst extracts a 5- to 15-second video and identifies nine facial features on the first frame, Fig. 2 (this step can be fully automated by employing a face and facial feature detector). Then, a standard SIFT-based descriptor is extracted at each feature point [12]. These features are localized on the next frame of the video using a simple tracking algorithm.

Assuming small motions between consecutive frames, we expect each facial feature to be located in a small neighborhood relative to its position in the previous frame. Denote the i^{th} SIFT descriptor at spatial location x, y and time t as $\phi(x_i, y_i, t)$. The location of this feature at time $t+1$ is given to be the location that minimizes the mean square error between descriptors:

$$E(u, v) = \|\phi(x_i, y_i, t) - \phi(x_i + u, y_i + v, t + 1)\|, \quad (1)$$

where $u, v \in [-6, 6]$ pixels. This process is repeated on each frame, where each feature descriptor is updated on each frame to contend with variations in lighting, scale, and pose.

2.2. Head alignment

Shown in Fig. 2 (right) is a 3-D head model aligned to the image shown on the left of this same figure. We employ a generic 3-D head model customized to each individual through only an anisotropic scaling in the horizontal and vertical directions. On the first frame of the video, this scaling and the overall head pose is automatically estimated. On each subsequent frame, the scaling is fixed, and only the head pose is re-estimated.

The head pose is estimated from the nine corresponding 2-D (video frame) and 3-D (model) features. Denote the location of the i^{th} facial feature in a video frame as $\mathbf{p}_i \in \mathcal{R}^2$ and its corresponding feature in the 3-D model as $\mathbf{P}_i \in \mathcal{R}^3$. Given these corresponding features, we estimate the extrinsic parameters that optimally map the 3-D model to the observed 2-D video frame [13]. The extrinsic parameters are composed of three rotation angles and three translation components. The intrinsic parameters are composed of only the focal length which we assume to be 35mm (we also assume that the principle point is the image center and assume negligible sensor skew and lens distortion). The six extrinsic parameters are estimated by minimizing the following error function:

$$E(\mathbf{R}, \mathbf{t}) = \sum_i \|\mathbf{p}_i - \mathbf{F}\begin{pmatrix}\mathbf{R} & \mathbf{t}\end{pmatrix}\mathbf{P}_i\|, \quad (2)$$

where $\mathbf{R}$ is a 3×3 rotation matrix, $\mathbf{t}$ is a 3×1 translation vector, $\mathbf{F}$ is the 3×3 intrinsic matrix, and the coordinates $\mathbf{p}_i$ and $\mathbf{P}_i$ are specified in homogeneous coordinates. On the first video frame only, an additional anisotropic scaling is added to the model estimation to yield the following error function to be minimized:

$$E(\mathbf{R}, \mathbf{t}, \mathbf{s}) = \sum_i \|\mathbf{p}_i - \mathbf{F}\begin{pmatrix}\mathbf{R} & \mathbf{t}\end{pmatrix}\mathbf{S}\mathbf{P}_i\|, \quad (3)$$

where $\mathbf{S}$ is an anisotropic scaling matrix in which the horizontal scaling is fixed at 1, and the vertical scaling $\mathbf{s}$ is a free parameter. This scaling adjusts the aspect ratio of the 3-D head model and allows for a minimum customization to the head being analyzed.

The error functions in Equations (2) and (3) are minimized using a standard unconstrained nonlinear optimization. An additional regularization term is added to these error functions, imposing a smoothness on the parameters over time. Shown in Fig. 4 and 5 are representative examples of the results of head tracking and alignment to a 3-D model.

2.3. Extracting a physiological signal

With a 3-D head model aligned to each video frame, the head on each frame is aligned to a canonical viewpoint. A common patch of skin (the forehead or cheek) is then extracted, from which the desired physiological signal is measured. However, the changes in color in a person's face due to their pulse is extremely small and nearly invisible to the human eye.

In order to make the measurement of this signal more reliable, we use a technique for magnifying tiny motions in a video. This technique, termed Eulerian video magnification [11], takes as input a standard video and outputs a video in which the color variations for a given temporal frequency are magnified. We magnify the color variations at a temporal frequency consistent with a typical pulse of 50-60 heart beats per minute (0.83Hz - 1.0Hz). For simplicity, we then compute the average luminance across a small extracted patch of skin.

Shown in Fig. 1, for example, are several frames of a human face (top) and CG face (bottom) after applying Eulerian

video magnification. Also shown are space-time plots extracted from these videos. In these plots, the horizontal axis corresponds to time and the vertical axis corresponds to the pixel color from a single vertical scanline in the center of the forehead. Shown superimposed on these space-time plots are the average luminance revealing the presence (top) and absence (bottom) of a pulse for the human and CG character, respectively.

3. RESULTS

We evaluated the efficacy of the proposed forensic technique on twelve videos, six containing human characters and six containing CG characters. The human characters consisted of videos of our creation and of videos downloaded from the Web. The CG characters were each downloaded from a variety of websites. For each video we manually extracted a 4.5 second section of the video. With a typical pulse of 50-60 beats per minute, this video length yielded approximately four beats in the extracted physiological signal.

Shown in Fig. 3 is the measured physiological signal extracted from human (top) and CG (bottom) characters.[1] Shown on the left is one frame of each 4.5 second long video and on the right is the physiological signal extracted following the procedure described in Section 2. A periodic human pulse is clearly visible for the three human characters, but not for the three CG characters. (Note that the scale of the vertical axis for the video of Lance Armstrong, third from the top, is different than the others).

Shown in Fig. 4 and 5 are two more detailed examples. Shown in the top row are five frames of a 4.5 second long video. Shown in the middle row is the corresponding aligned 3-D model. The nine facial features used for tracking and alignment are annotated in both the video frames and 3-D model. The rectangular region corresponds to the portion of the forehead from which the physiological signal is measured. As with the results in Fig. 3, we clearly see a pulse for the human face, Fig. 4, but not for the CG face, Fig. 5. Although not shown here, the remaining videos followed the same pattern of results.

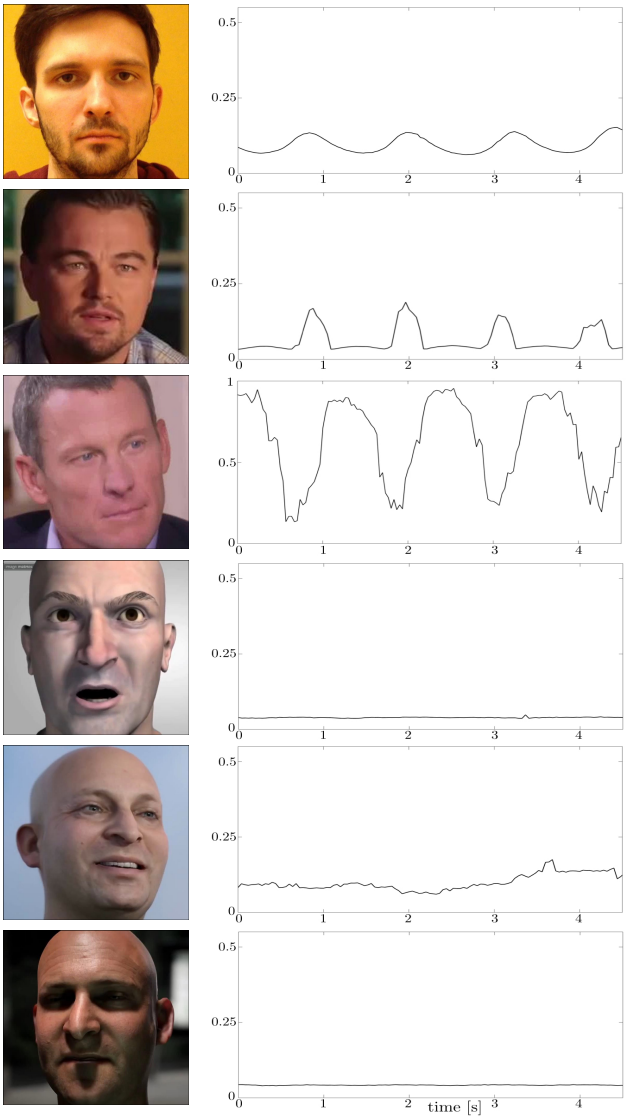

Fig. 3. Shown is one frame of a video and the measured physiological signal. The characters in the top three panels are human while the characters in the bottom three panels are CG. The periodic pulse is clearly visible in the human but not in the CG characters.

4. DISCUSSION

We have described a physiologically-based forensic technique for distinguishing between computer generated and human faces. This technique detects the absence or presence of a human pulse in a video. This physiological signal is, of course, naturally present in all living humans. Because this signal is nearly invisible to the human eye, there is no obvious reason why a modeler would add this signal to a computer generated character. While we have focused only on faces, the proposed approach could be applied to any part of visible skin.

The drawback of this technique is that it is only applicable to video of a person and might be vulnerable to countermeasures should a modeler choose to artificially introduce a pulse to a CG character, or should a forger remove the pulse from a human character. The benefits, however, are that it exploits a naturally occurring physiological signal, is nearly fully automatic, is effective when analyzing low quality and low resolution video, and does not require extensive data collection common to statistically-based techniques. We expect to fully automate this forensic technique by employing a face and facial feature detector. We also plan to estimate and remove luminance changes due to lighting which could, if they modulate at the same frequency as a pulse, confound our classification.

[1] Fig. 3, videos in rows 2-6 dowloaded from: youtu.be/B76UZCccdQE; youtu.be/Vq8NgepsFg8; youtu.be/WO9W56KcCb8; youtu.be/l6R6N4Vy0nE; youtu.be/CvaGd4KqlvQ.

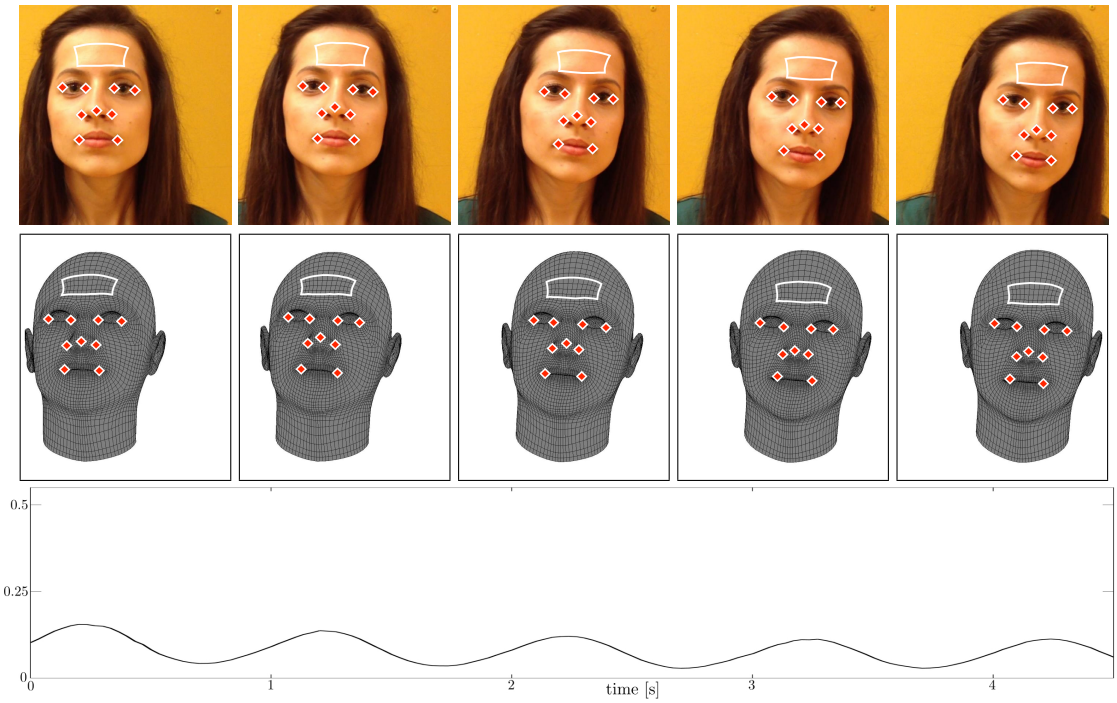

Fig. 4. Shown in the top row are five consecutive frames of a 4.5 second long video. Shown in the middle row is the corresponding aligned 3-D model. The nine facial features used for tracking and alignment are annotated in both the video frames and 3-D model. The rectangular region corresponds to portion of the forehead from which the physiological signal is measured. Shown in the bottom row is the measured physiological signal revealing the underlying human pulse.

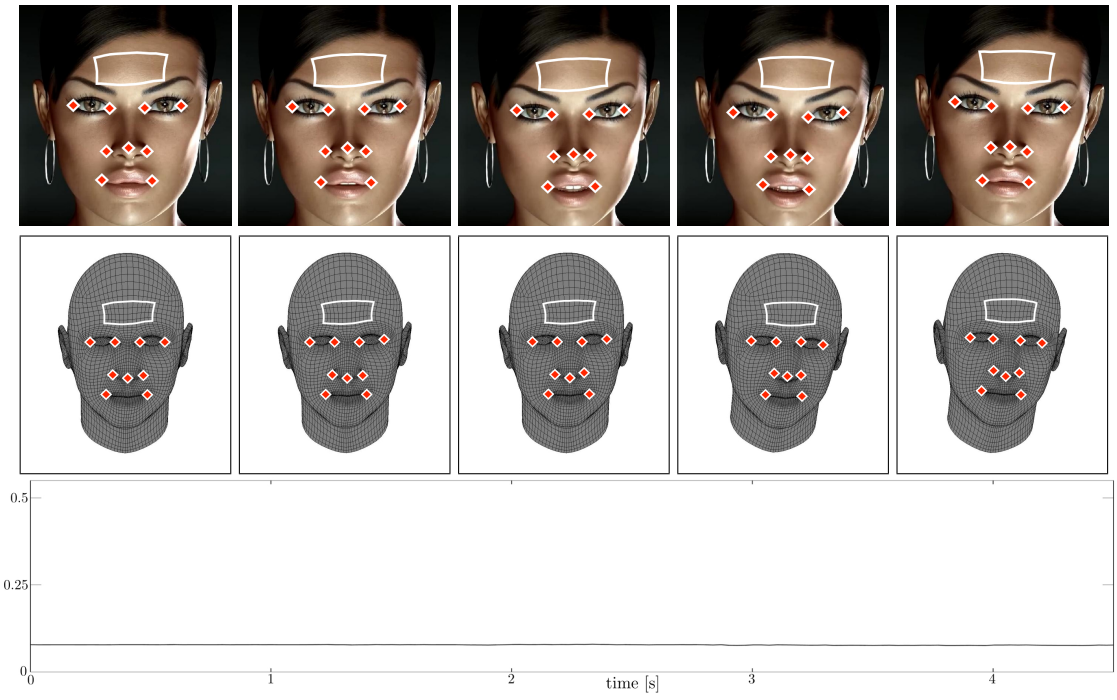

Fig. 5. Shown in the top row are five consecutive frames of a 4.5 second long video. Shown in the middle row is the corresponding aligned 3-D model. The nine facial features used for tracking and alignment are annotated in both the video frames and 3-D model. The rectangular region corresponds to portion of the forehead from which the physiological signal is measured. The lack of a periodic signal in the bottom row indicated the lack of a pulse for this CG model. [Video downloaded from: youtu.be/nX8KitVCcZM]

5. REFERENCES

[1] S. Lyu and H. Farid, "How realistic is photorealistic?," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 845–850, 2005.

[2] Y. Wang and P. Moulin, "On discrimination between photorealistic and photographic images," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. II.161 – II.164.

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

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

[5] A.E. Dirik, S. Bayram, H.T. Sencar, and N. Memon, "New features to identify computer generated images," in *IEEE International Conference on Image Processing*, 2007, vol. 4, pp. IV.433 – IV.436.

[6] A. Gallagher and T. Chen, "Image authentication by detecting traces of demosaicing," in *IEEE Computer Vision and Pattern Recognition Workshop*, 2008, pp. 1–8.

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

[8] J.-F. Lalonde and A. A. Efros, "Using color compatibility for assessing image realism," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[9] D.-T. Dang-Nguyen, G. Boato, and F. G. B. De Natale, "Discrimination between computer generated and natural human faces based on asymmetry information," in *European Signal Processing Conference*, 2012, pp. 1234 – 1238.

[10] D.-T. Dang-Nguyen, G. Boato, and F. G. B. De Natale, "Identify computer generated characters by analysing facial expressions variation," in *IEEE International Workshop on Information Forensics and Security*, 2012, pp. 252–257.

[11] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)*, vol. 31, no. 4, 2012.

[12] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 2, no. 60, pp. 91–110, 2004.

[13] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.