DETECTION OF COMPUTER GENERATED FACES IN VIDEOS BASED ON PULSE SIGNAL

Bo Peng, Wei Wang, Jing Dong and Tieniu Tan*

Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
E-mail: { bo.peng, wwang, jdong, tnt }@nlpr.ia.ac.cn

ABSTRACT

The rapid development in the field of computer graphics (CG) makes it quite easy to create photo-realistic images and videos. This brings forward an emergent requirement for techniques that can distinguish CG from real contents. In this paper, we propose a method that leverages human pulse signal to distinguish between CG and real videos that include human faces. We use a robust tracking method to locate a patch of skin on the face. Then, a chrominance-based algorithm is employed to robustly extract pulse signal. By checking the frequency waveform of the extracted pulse signal, we can tell CG and real videos apart. The experiment shows encouraging results, which demonstrate the efficiency of our method.

***Index Terms*—** Video forensics, CG characters, pulse detection

1. INTRODUCTION

With the fast evolution of Computer Graphic (CG) technologies, Computer Generated Imagery (CGI) is increasingly photo-realistic. Many easy-to-use 3D modeling and rendering softwares such as Maya, 3ds Max, Blender etc. make it convenient for ordinary people to create Hollywood-like special effects. This posts a new challenge for the forensic society, as it is becoming more and more difficult to distinguish CG from real contents.

CG technologies can be used by vicious groups to produce deceiving and malicious videos. Imagine the vicious ones manipulated a CG movie star to acknowledge some of his rumors, this will pollute his reputation. Even worse, it can cause social instability or international intension if a CG politician is made to make a fake statement. Motivated by the above reason, in this work, we focus on judging the authenticity of videos that include human faces.

*Corresponding author

*This work is funded by the National Basic Research Program of China (Grant No. 2012CB316300), the National Nature Science Foundation of China (Grant No.61303262), and the National Key Technology R&D Program (Grant No.2012BAH04F02).

In the literature, there has been some work to distinguish between CG and real contents, most of which focus on images and are based on statistical features. The techniques proposed in [1, 2] extract statistical features from wavelet domain and train a classifier for CG and real images. Some other methods leverage clues from digital cameras which include chromatic aberration [3], color filter array demosaicing [4] and sensor pattern noise [5, 6]. Geometric- and physics-based features are also explored in [7]. The methods mentioned above are general purpose and can apply to any image. Recently, some methods specifically designed for human faces are proposed either for images [8] or for videos [9, 10, 11]. The method in [8] is based on face asymmetry information and [9, 10] explore the difference of facial expression variation patterns between CG and natural characters in videos.

Different from [9, 10], [11] proposes to use a physiological signal – human pulse to identify CG characters. The advantage of this method is that it exploits a naturally and universally common signal to humans. Besides, it does not require extensive data collection as in statistically-based methods. The method in [11] uses an Eulerian video magnification technique [12] to amplify the pulse signal of human facial skin. The presence or absence of this pulse signal is a sign of real human or CG character. Although this idea is very innovative, the technique suffers from several practical problems. A critical one is the coarse frequency resolution caused by a very short duration of the video clip, which we will elaborate in the next section. The Eulerian video magnification technique used in this work also has some limitations. This technique is designed for videos of visually still objects[1]. Although it is applied to a tracked patch of skin in [11], this technique cannot suppress the noise introduced by head motion. Another drawback is that the Eulerian technique is sensitive to parameters and it needs a try and error procedure to get a satisfying result.

In this paper, we propose a new method to more reliably distinguish between CG and natural characters by explicitly extracting pulse signals. This method improves the detected

[1]See `http://people.csail.mit.edu/mrub/vidmag/` and `https://videoscope.qrilab.com/` for example videos

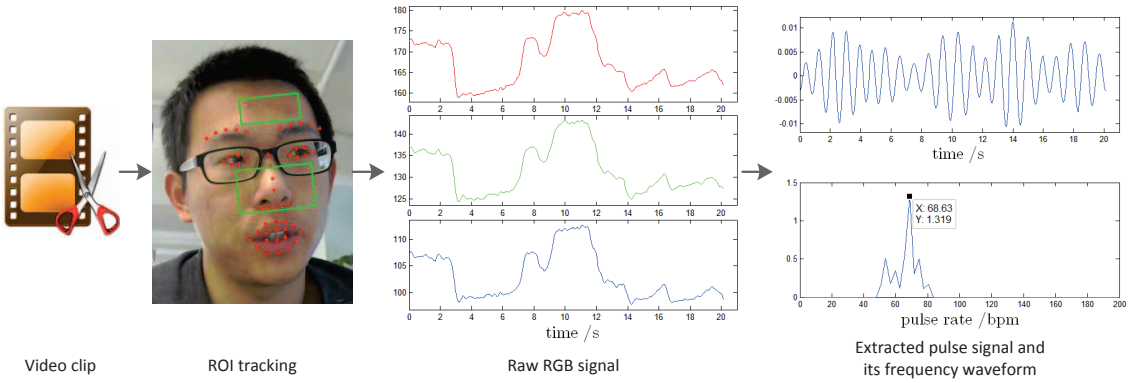

Fig. 1. Overview of the proposed method. (Better viewed in color)

pulse signal's frequency resolution by choosing a longer clip duration compared to [11]. It also uses a more sophisticated facial landmark detection method to track the Region of Interest (ROI) on faces. The chrominance-based pulse extraction technique in our method can resist motion artifact to some extent which makes this method more robust to use.

2. METHODS

The overview of the proposed method is illustrated in Fig. 1. To extract the pulse signal to distinguish between CG character and real human, a clip of video consistently showing a human face is first selected manually. A ROI on the face is tracked based on a facial landmark detection result. The raw RGB signal is then obtained by averaging the intensity value in the ROI for each of the RGB channel and for each frame. The pulse signal can be extracted by analyzing the chrominance components of the raw RGB signal. A final decision can be made by observing the frequency waveform of the extracted pulse signal. In the following subsections, we will describe the proposed method in more details.

2.1. Video clip selection

Analyzing the whole video is not necessary because it is too time-consuming. We just select a clip of video which consistently includes a character's face in its full duration. It is necessary to first study the relation between the selected clip's duration and signals frequency resolution, since we will eventually analyze the extracted pulse signal in its frequency domain. In the following, a video clip's duration is denoted as T, its number of frames as N and video framerate as f_s. In digital signal processing field, T, N, f_s are also known as sampling time, number of sampling points and sampling rate. We treat the skin color value in each video frame as a sampling point of the pulse-related time-variant signal that is shown in Fig. 1 as the raw RGB signal. By applying Discrete Fourier Transform (DFT), the extracted pulse signal's frequency waveform will have the highest frequency sampling point at $\frac{f_s}{2}$ in Hz, or $30f_s$ in beat per minute (bpm). Its frequency resolution, i.e. the lowest frequency sampling point is equal to the inverse of sampling time, which is

$$d_f = \frac{1}{T} = \frac{f_s}{N} \tag{1}$$

where d_f denotes frequency resolution. As we can see, the frequency resolution is determined solely by sampling time. With a longer clip duration, we can get higher frequency resolution and more accurate pulse rate measurement. But in practical applications, most videos do not consistently contain a human face for more than one minute. Besides, in long duration clips, faces are more likely to have large motions and this will introduce too much motion noise into the extracted pulse signal. To find a proper balance between high frequency resolution and practical use, we tried different clip duration lengths. We found that using clips around 20 seconds can obtain satisfying results. With this choice, the extracted pulse signal's frequency resolution is 0.05Hz or 3bpm which is high enough for a reliable analysis.

Different from our choice, the method in [11] selects a clip of 4.5s, which is too short to get a meaningful result. We can also apply the above relation between clip duration and frequency resolution to analyze [11]. With a duration of 4.5s, the frequency resolution is 13.3bpm. This number is intolerably coarse. In [11], the authors only choose the frequency band of 50∼60bpm, which corresponds to common pulse rates. With a resolution of 13.3bpm, the first few frequency sampling points are at 0, 13.3, 26.6, 39.9, 53.2, 66.5, ... in bpm. We can see that there is actually only one point inside this passband which is at 53.2bpm (the 5th point). Obviously, not all people has the same pulse rate around 53.2bpm. A person who has a pulse rate not exactly as 53.2bpm will be decided as a CG. Thus, the coarse frequency resolution of short clip in [11] hinders this method from making meaningful distinguishments.

2.2. ROI tracking

Robust ROI tracking is necessary for extracting pulse signal if the character in the video clip has head motion. In this subsection we describe how to track a ROI on a character's face based on a facial landmark detection method [13].

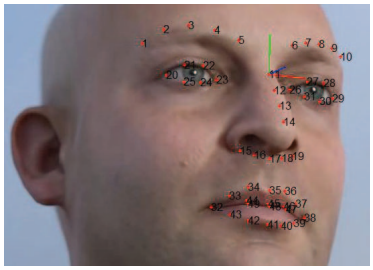

Fig. 2. 1∼49 are the facial landmarks. The red, green and blue lines are the projected world coordinate axes attached to the head indicating the head pose. (Better viewed in color)

With the help of "IntraFace"[2] in [13], we can get 49 facial landmarks and the head pose as shown in Fig. 2. These facial landmarks serve as a reference for locating a ROI. We define a reference oblique coordinate system. The origin is located between the eyes at point 11 in Fig. 2. The x and y axes of the oblique coordinate are along the red and green lines in Fig. 2. The unit length of x axis is defined as the distance between two outer corners of the eyes, i.e. point 20 and 29. The unit length of y axis is defined as the the distance between upper and lower points of the nose, i.e. point 11 and 17. The ROI shape used in our method is a parallelogram. This ROI parallelogram is defined by its top-left and bottom-right corners with respect to the reference coordinate. The ROI's edges are in parallel with the coordinate axes. The relative position of the ROI with respect to the reference coordinate is fixed as specified in the first frame. This allows for accurate extraction of raw RGB signal from basically the same area of face skin.

We also want to get some insights into the strategy of choosing an appropriate ROI placement for signal extraction. To this end, we experiment on some steady face videos, for which we try different placements of the ROIs and compare the quality of the raw RGB signals in the ROIs. Obviously, the ROI should not contain the eyes or mouth, where the signal is likely to be corrupted by eye blinking and talking movement. We choose two candidate ROIs, which are the forehead area, and the area around the nose. As can be seen in Fig. 3, we found that the area around the nose has stronger pulse-related signal compared to the forehead. However, the nose area suffers from more deformation when the character has severe expression change. This deformation can introduce color variation noise that interferes with pulse signal. Other

[2]The software is publicly available at `http://www.humansensing.cs.cmu.edu/intraface/`

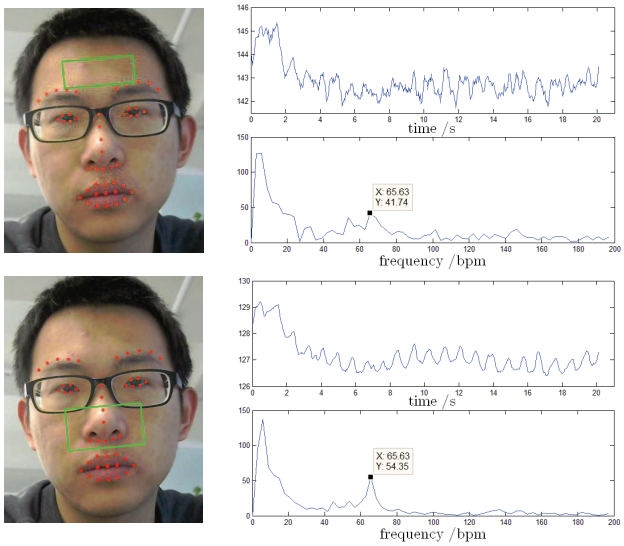

Fig. 3. The left panels show different ROI placements. The right panels show raw signal in G channel and its frequency waveform (only 0∼200bmp is shown).

factors can also influence the placement decision of ROI. For example, if the character's forehead is covered by hair, just nose area can be selected as ROI. When the face is not uniformly lit and the nose cast an obvious shadow, or the face has severe expression change, selecting the nose ROI can be error-prone.

As a final decision, we leave the ROI placement step to the user for flexibility. In our implementation, a user can click on the first frame to decide the placements of an arbitrary number of parallelograms as one combined ROI. Usually, the combination of forehead area and nose area can get satisfying result. The algorithm then robustly track the specified ROI in the following frames.

2.3. Pulse signal extraction

Based on the tracked ROI in each frame, we average the intensity value of all pixels in the ROI for each of the RGB channel to get the raw RGB signal. We use a chrominance-based method proposed in [14] to extract the pulse signal. The pulse extraction method is based on a sophisticated light reflection model of human skin. The influence of head motion is seen as an equal intensity modulation for all channels. Hence, pulse signal can be got from chrominance components to suppress motion noise. The extracted pulse signal is in the form of a linear combination of the bandpass filtered and normalized raw RGB signal. Two chrominance signals are defined as

$$X = 3R_n - 2G_n \tag{2}$$

$$Y = 1.5R_n + G_n - 1.5B_n \tag{3}$$

where C_n is the normalized version of $C \in \{R, G, B\}$ by dividing its samples by their mean over a temporal interval.

$$C_{ni} = \frac{\sum_{j=-n}^{n} C_{i+j}}{2n+1} \tag{4}$$

Here, C_{ni} denotes the ith sampling point of C_n and $2n+1$ is the temporal interval length. To eliminate the disturbance from irrelevant frequencies, X, Y is bandpass filtered to get X_f, Y_f. Then the pulse signal is obtained as

$$S = X_f - \alpha Y_f \tag{5}$$

or equally

$$S = 3(1 - \frac{\alpha}{2})R_f - 2(1 + \frac{\alpha}{2})G_f + \frac{3\alpha}{2}B_f \tag{6}$$

where α is the ratio of the standard deviations of X_f and Y_f

$$\alpha = \frac{\sigma(X_f)}{\sigma(Y_f)} \tag{7}$$

In our experiment, we set the passband to $50 \sim 80bpm$ which corresponds to the range of normal human pulse rate. The readers are referred to the original paper [14] for more details about the pulse extraction algorithm. We apply DFT to the extracted pulse signal S and analyze its frequency waveform to distinguish between CG and natural characters. If the character in a video is real human, we will find a clear peak in the pulse frequency band. On the contrary, the frequency waveform of the "pulse" extracted from a CG character will be just random artifact noise and has multiple separated peaks.

3. RESULTS

Since the resources of CG videos are quite limited, we conducted an experiment on only 6 videos which include 3 real and 3 CG. These videos are either downloaded from websites or captured by our own web camera. The resolution of these videos is 1280×720, because higher quality videos contain stronger pulse signal. Video clips of around 20 seconds are selected manually. The ROI is placed on the forehead and nose areas. Fig. 4 shows the results of this experiment. As can be seen, the frequency waveforms of the extracted pulse signals from real and CG characters are quite different. The real human videos show one dominant peak which indicates the frequency of human pulse rate. While the CG character videos have no clear peak or have multiple equal peaks. The distinguishment is made by human observation at the present implementation. If enough CG videos are available, a machine learning method can be used to automatically distinguish them.

To compare with the method proposed in [11], we implemented their algorithm to our best. However, we can not reproduce any meaningful results as reported in [11] due to coarse frequency resolution and uncertainty of the parameters for Eulerian video magnification technique.

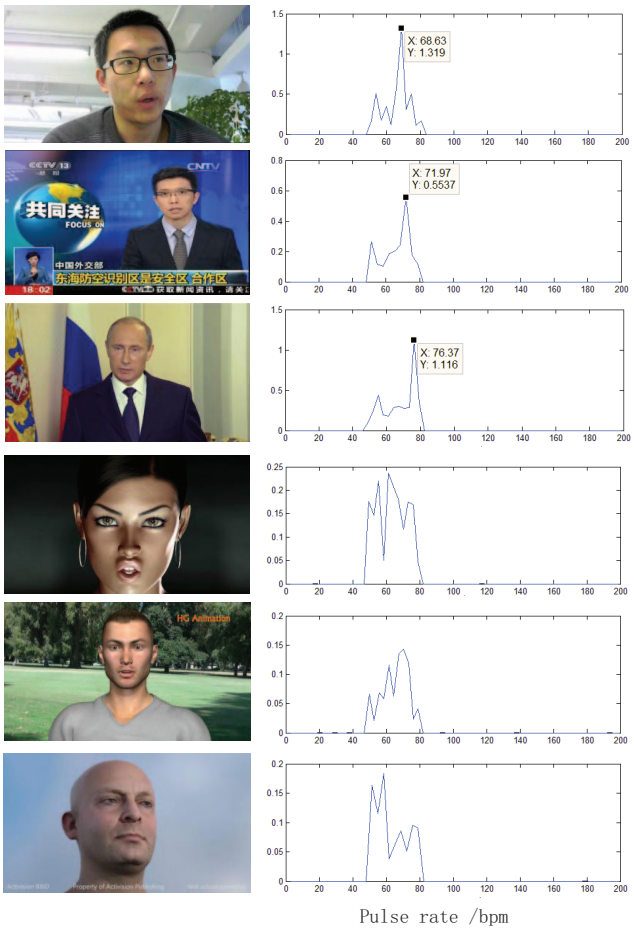

Fig. 4. Results of the proposed method. Shown in the left panels are one frame of each of the videos. Shown in the right panels are the frequency waveform of the extracted pulse signal from these videos. The top three panels show real humans, and the bottom three show CG characters. A dominant peak can be observed for real human videos indicating the detected pulse rate, while the waveform of CG show no clear peaks.

4. CONCLUSIONS

In this paper, we proposed an efficient method to distinguish between CG and real videos showing human faces. We extract pulse signal from face skin and examine its frequency waveform to make a decision. Given a video clip, the ROI tracking and pulse signal extraction procedures are fully automatic. Our chrominance-based pulse signal extraction algorithm has a better resistance to motion related noise and gives reliable results. We also improved the pulse signal's frequency resolution by analyzing its relation to clip duration. Experiment on several videos that include naturally talking characters shows encouraging results. The resulting frequency waveforms of real and CG videos are quite different and easy to distinguish.

5. REFERENCES

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

[2] Ying Wang and P. Moulin, "On discrimination between photorealistic and photographic images," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, May 2006, vol. 2, pp. II–II.

[3] A.C. Gallagher and Tsuhan Chen, "Image authentication by detecting traces of demosaicing," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, June 2008, pp. 1–8.

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

[5] S. Dehnie, "Digital image forensics for identifying computer generated and digital camera images," in *Image Processing, 2006 IEEE International Conference on*, Oct 2006, pp. 2313–2316.

[6] 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 *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 1653–1656.

[7] 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*, New York, NY, USA, 2005, MULTIMEDIA '05, pp. 239–248, ACM.

[8] Duc-Tien Dang-Nguyen, G. Boato, and F.G.B. De Natale, "Discrimination between computer generated and natural human faces based on asymmetry information," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug 2012, pp. 1234–1238.

[9] Duc-Tien Dang-Nguyen, G. Boato, and F.G.B. De Natale, "Identify computer generated characters by analysing facial expressions variation," in *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, Dec 2012, pp. 252–257.

[10] Duc-Tien Dang-Nguyen, Giulia Boato, and Francesco G.B. De Natale, "Revealing synthetic facial animations of realistic characters," in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 5327–5331.

[11] V. Conotter, E. Bodnari, G. Boato, and H. Farid, "Physiologically-based detection of computer generated faces in video," in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 248–252.

[12] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *Acm Transactions on Graphics*, vol. 31, no. 4, pp. 1–8, 2012, 998OP Times Cited:12 Cited References Count:11.

[13] Xiong Xuehan and F. De La Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 532–539.

[14] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 10, pp. 2878–2886, 2013.