Perceptual Judgments to Detect Computer Generated Forged Faces in Social Media

Suzan Anwar[1,2](✉), Mariofanna Milanova[1], Mardin Anwer[2,3], and Anderson Banihirwe[1]

[1] University of Arkansas at Little Rock, Little Rock, USA
sxanwar@ualr.edu
[2] Salahaddin University, Erbil, Iraq
[3] Lebanese-French University, Erbil, Iraq

Abstract. There has been an increasing interest in developing methods for image representation learning, focused in particular on training deep neural networks to synthesize images. Generative adversarial networks (GANs) are used to apply face aging, to generate new viewpoints, or to alter face attributes like skin color. For forensics specifically on faces, some methods have been proposed to distinguish computer generated faces from natural ones and to detect face retouching. We propose to investigate techniques based on perceptual judgments to detect image/video manipulation produced by deep learning architectures. The main objectives of this study are: (1) To develop technique to make a distinction between Computer Generated and photographic faces based on Facial Expressions Analysis; (2) To develop entropy-based technique for forgery detection in Computer Generated (CG) human faces. The results show differences between emotions in both original and altered videos. These computed results were large and statistically significant. The results show that the entropy value for the altered videos is reduced comparing with the value of the original videos. Histograms of original frames have heavy tailed distribution, while in case of altered frames; the histograms are sharper due to the tiny values of images vertical and horizontal edges.

Keywords: Video manipulation · ASM · Face expression · Entropy based histogram

1 Introduction

With advances in computer vision and graphics, it has become possible to generate image/videos with realistic synthetic faces. Companies like Google, Baidu, Nvidia, Adobe and startups such as Voicecey have recently funded efforts to fabricate audio or video. These companies have released do-it yourself software and open source tools available on GitHub such as DeepFake. Currently in-use methods can generate manipulated videos in real time Face2Face, can synthesize video based on audio input or can artificially animate static images. New technologies allow users to edit facial expressions. This gained incredible attention in the context of fake-news discussions.

F. Schwenker and S. Scherer (Eds.): MPRSS 2018, LNAI 11377, pp. 38–48, 2019.
https://doi.org/10.1007/978-3-030-20984-1_4

The results are raising concerns that face swaps technology can be used to spread misleading information.

According recent publication [1] “Right now, there is no tool that works all the time” says Mikel Rodriguez, a researcher in Mitre DCorp. The overview of face image synthesis approaches using deep learning techniques is presented in [2]. Most of the techniques used to swap faces generate an output as a face image or 3D facemask. For instance, it was virtually impossible to distinguish between the real Paul Walker and the computer-generated one in the film “The Fast and the Furious 7”. The death of the actor during filming led the director to use previously recorded digital 3D scan data to reconstruct Mr. Walker’s face for the unfinished scenes. Another example is Pro Evolution Soccer2, a video game developed and published by Konami. Since the 2012 version, the images of the soccer players are rendered so realistically that they look almost like real people [3].

In June, 2017 NVIDIA created a GAN that used CelebA-HQ’s database of photos of famous people to generate images of people who don’t actually exist. In 2018 NVIDIA proposed a new GAN that increases the variation in generated images [4]. GANs are used to apply face aging, to generate new viewpoints, or to alter face attributes like skin color [5].

Virtual Worlds have been used constructively for the benefit of the society. However, there are safety and security concerns as well e.g. cyberterrorism activities, child pornography detection and economic crimes such as money laundering.

Unreal images and videos can be used to harm people or to gain political and/or economic advantage. For example, fake images or videos about aliens, disasters, statesmen, or businessmen can create confusion or change people’s’ opinions. Social media platforms such as Facebook, Twitter, Flickr, or YouTube are ideal environments to widely disseminate these fake images and videos.

To combat this threat, CG manipulation-detection software will need to become more sophisticated and useful in the future. This technology, along with robust training and clear guidelines about what is acceptable, will enable media organizations to hold the line against willful image manipulation, thus maintaining their credibility and reputation as purveyors of the truth. The challenges to create new technologies are:

1. The algorithms used to fabricate images/video are based on convolutional neural network, widely used in object recognition. In Deep learning approach, features are automatically learned from training samples rather than being manually designed. However deep learning – based approaches are using mostly supervised learning. Although deep-learning-based approaches are promising, they are not yet mature in digital image forensics; a considerable amount of work remains to be done in this area.
2. Lack of sharing datasets, maintenance, and availability. Coming to a world where everything is connected (IoT) there is a need to collect data from streaming devices, such as Roku or AppleTV and Unmanned Aerial Vehicle (UAV) and variations of computer-generated images using new deep learning architectures.

We propose to investigate techniques based on perceptual judgments to detect image/video manipulation produced by deep learning architectures. The main objectives of this study are:

- To develop techniques to make a distinction between computer generated and photographic faces based on facial expressions analysis. The hypothesis is that facial emotions expressed by humans and facial expressions generated from fake faces are different. Humans can produce a large variety of facial expressions with a high range of intensities.
- To develop entropy based technique for forgery detection in CG human faces. The hypotheses is that natural images have some special properties different from the other types of images.

2 Related Work

Given the need for automated real-time verification of the digital image/video content, several techniques have been presented by researchers. There are two major categories of digital image treatment detection approaches: active approaches and passive approaches. Active approaches involve various kinds of watermarks or fingerprints of the image content and embedding them into the digital image [6]. With rising number of images used in social networks, it is impossible to require all the digital images on the internet to be watermarked before distribution. Therefore, passive forensics approaches have become a more popular choice.

Passive approaches detect changes in digital image by analyzing specific inherent clues or patterns that occur during the modification stage of digital images. Passive approaches do not rely on any prior or preset information and they have a broader application in image forensics. These techniques are successfully applied for tracking true and false news. In [7] the traces are classified in three groups: traces left in image acquisition, traces left in image storage, and traces left in image editing. Recently new category becomes popular images generated by computer graphics software.

For forensics specifically on faces, some methods have been proposed to distinguish computer generated faces from natural ones [8] and to detect face retouching [9]. In biometry, two pre-trained deep CNNs, VGG19 and AlexNet are proposed to detect morphed faces [10]. In [11] the authors proposed detection of two different face swapping manipulations using a two-stream network: one stream detects low-level inconsistencies between image patches while the other stream explicitly detects tampered faces.

Researchers from the Technical University of Munich have developed a deep learning algorithm that potentially identifies forged videos of face swaps on the internet. They trained the algorithm using a large set of face swaps that they made themselves, creating the largest database of these kinds of images available. They then trained the algorithm, called XceptionNet, to detect the face swaps [12].

In [13] different algorithms to detect and classify original and manipulated video are presented. In fact, this is difficult task for humans and computers alike, especially when the videos are compressed and have low resolution, as it often happens on social media. The authors also present a large-scale video dataset called “Face Forensics”.

Some forgery detection methods also use statistical features to detect forgery. This technology is based on methods using natural image statistics. Natural images have some special properties different from the other types of images [14]. In [15] CG faces and real faces are discriminated by analyzing the variation of facial expressions in a video by analyzing sets of feature points.

3 Facial Emotion

In the experiments, the software FaceXpress which proposed in [16] is used to recognize the emotion for each frame within the FaceForansic dataset. The software starts with deting the face using Viola-Jones detector, followed by indicating 116 face landmarks using a multi resolution tracker Active Shape Model (ASM) tracker [17]. FaceXpress detects facial triangulation points using Active Shape Model tracker (see Fig. 1). Attributes are obtained by measuring the length of among the detected facial triangulation points. For some attributes such as, mouth width, mouth height, and the distance of the midpoint of eye gap to eyebrow midpoints, are obtained using Mahalanobis distance. The other attributes such as, the domain of vertical edge in the forehead and domain of horizontal edge in the mid forehead are obtained by filtering with Gauss core. The tracking point’s location is used to compute the changes in facial regions such as eye brow wrinkles, forehead wrinkles, wrinkles in cheeks, distance eye to eyebrows, and vertical and horizontal measures of mouth [18] (see Fig. 2). Finally, a support vector machine (SVM) is used identified the detected facial emotion among the seven universal emotions; surprise, anger, happiness, sadness, fear, disgust, and neutral.

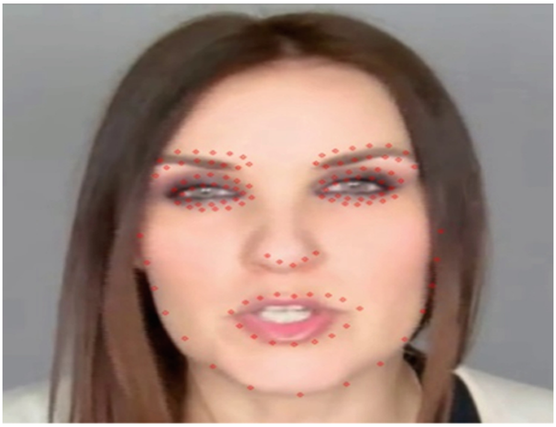

Fig. 1. Facial landmarks detection using ASM tracker

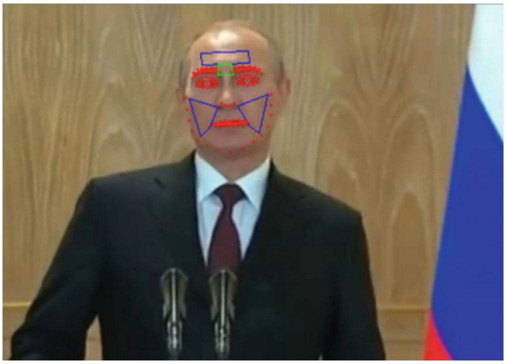

Fig. 2. Facial attributes that are used to detect the regions of interest

4 Entropy Based Histogram

Histogram processing includes image altering by modifying its histogram. To make the histogram of an image flat, normalization process is performed on both original and altered images from FaceForensics dataset (see Sect. 5.1). This process is called contrast enhancement where the function of intensity transformation based on information such as compression, description, and segmentation, are extracted. To compute the histogram of an image, the following discrete function is applied for intensity levels within [0, L−1] range.

$$h(r_k) = n_k \tag{1}$$

r_k is the intensity value,
n_k is the number of pixels in the image with intensity r_k,
$h(n_k)$ is the histogram of the digital image with Gray level r_k.

The total number of pixels is used for normalizing the image histogram by assuming an M × N image. This normalization computation is related to r_k probability of occurrence in the image. The equation to normalize the histogram is given below:

$$p(r_k) = \frac{n_k}{MN}, K = 0, 1, 2, \ldots..L - 1 \tag{2}$$

$p(n_k)$ computes the probability of occurrence estimation of image level r_k.

The summation of all normalized histogram components should be equal to 1 [19]. The histograms for same frames in both original and altered videos are different (see Fig. 3).

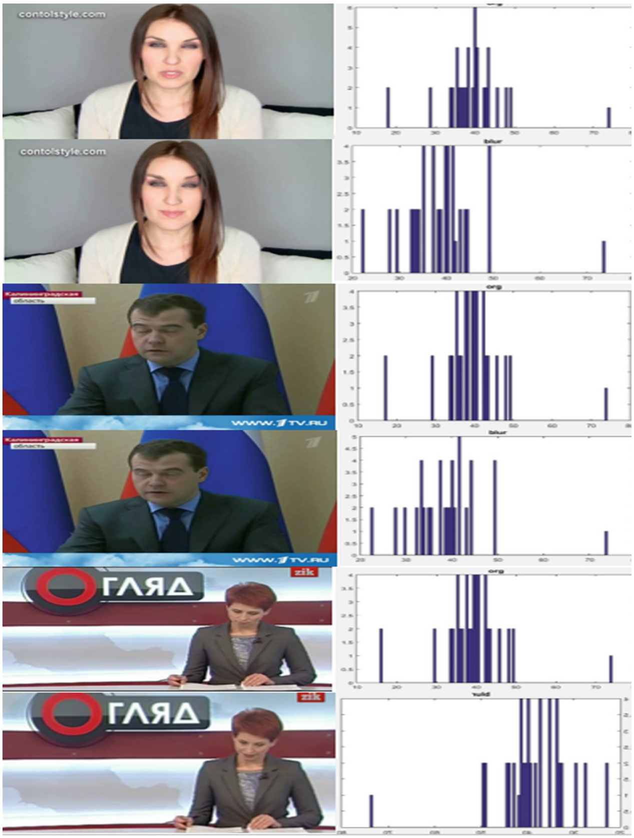

Fig. 3. Results of applying image histogram on both original and altered frames

Histograms of original frames have heavy tailed distribution. In case of altered frames, the histograms are sharper due to the tiny values of images vertical and horizontal edges. Also, an image mean information or entropy is determined from the images histogram. The purpose of computing the images entropy is to find its automatic focusing. For any random variable X, with probability density function f(x), the entropy definition is:

$$H(X) = -E[log\, f(X)] = -\int f(x) log\, f(x) dx \tag{3}$$

The range of the variable is divided into n intervals (l_k, u_k), k = 1, 2,…., n. The relation between the above entropy definition and the density that is represented as a histogram is shown in the following equation:

$$H(X) = -\sum_{k=1}^{n} \int_{l_k}^{u_k} f(x) log\, f(x) dx \tag{4}$$

The relation between kth bin of a histogram to the kth term of the above summation with width is represented in the following equation:

$$w_k = u_k - l_k \tag{5}$$

The bin probabilities p_k, $k = 1, 2, \ldots.\ n$ is defined as:

$$p_k = \int_{l_k}^{u_k} f(x)dx \tag{6}$$

Which can be approximated as $w_k f\,(x_\mathrm{k})$, where:
$f(x_k)$ is the area of a rectangle,
x_k is the interval $(l_k,\ u_k)$ value,
To the kth integral for Eq. (4) can be approximated as $w_k f(x_\mathrm{k}) log(x_k)$, this expression is used in term of bin probabilities to rewrite the entropy as:

$$H(X) = -\sum_{k=1}^{n} p_k log(p_k/w_k) \tag{7}$$

The above expression is given for a discrete distribution by Harris [20] and for a histogram by Rich and Tracy [21], if $w_k = 1$. When w_k is constant and not equal to 1, we used:

$$H(X) = -\sum_{k=1}^{n} p_k\ log\ p_k + log\ w \tag{8}$$

5 Results

5.1 FaceForensic Video Dataset

In this paper, we used the faceForansics video dataset [13] which consists of about 500.000 faces frames from around 1004 videos was collected from YouTube. The dataset has been manipulated using state-of-the-art face editing approach including classification and segmentation. The original face2face reenactment approach is used where the mouth interiors is selected from a mouth database depending on the target expression.

5.2 Facial Emotion

The FaceXpress software produces a csv file contains the recognized emotion for each frame for original and altered videos. To evaluate the differences in emotion between the original and altered videos, the mean square error (MSE) is the mean square error between the original and the altered video [22] from FaceForensic dataset. The metric

MSE between the produced emotions stored in the csv files for both original and altered videos is computed. For each frame in the original and altered video, the difference in emotion was squared and averaged. Table 1 shows the result of computing MSE for some videos' frames in the FaceForensic dataset, the results are dreadful and noticeable. By applying the FaceXpress software on some FaceForensic's videos, the results show a clear difference between emotions express in original and altered videos (see Fig. 4).

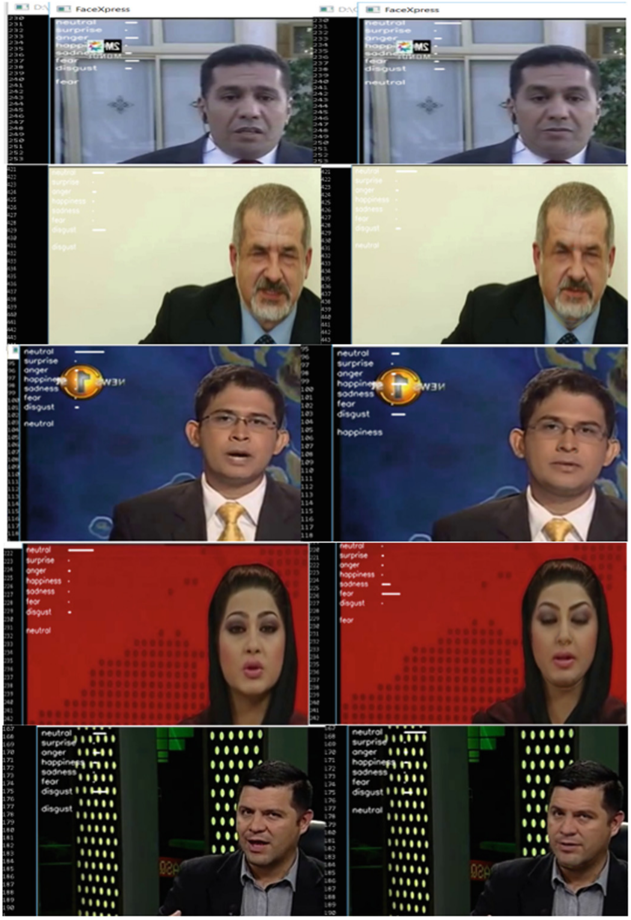

Fig. 4. Result of applying FaceXpress on some FaceForensic videos

Table 1. Results of MSE calculating

Video	MSE mean square error
v1	43.12965
v2	494.748
v3	420.4217
v4	224.6175
v5	327.7263
v6	96.96217
v7	322.4336
v8	172.4224
v9	367.1219
v10	366.6556
v11	306.3773
v12	123.6393
v13	22.05341
v14	162.6497
v15	38.28153
v101	168.4492
v102	226.1895
v148	113.8129
v149	29.48472
v150	364.4547

5.3 Entropy Based Histogram

Another measurement for image quality evaluation is computing the value of Entropy for both original and altered videos. We applied the entropy formula represented in Eq. 8, on both original and altered videos for the same frames. Table 2 shows that the entropy values for the altered frames are reduced comparing with their values for original videos.

Table 2. Results of computing Entropy value for three selected frames

Frame	Entropy value	
	Original	Altered
1	3.9122	3.8919
2	3.9114	3.8922
3	3.9106	3.8924

6 Conclusion

In this paper, we applied two different methods to test the quality and differences in emotions in FaceForensic original and altered videos dataset. In the first method, we used FaceXpress software to recognize the emotions in the videos for comparison. The differences in emotions between both original and altered videos are calculated using MSE measurement. The results of MSE values concluded that the differences in emotion are clear and noticeable between both original and altered videos. In the second method, we compute the Entropy values that are generated from the frame's histogram to test the quality of the videos. The result for the second method showed that the Entropy values for the altered videos are reduced comparing with their value for the original videos. Histograms of original frames have heavy tailed distribution, while in case of altered frames; the histograms are sharper due to the tiny values of images vertical and horizontal edges.

References

1. Nordrum, A.: Forging voices and faces. Spectrum IEEE 14–15, May 2018. https://doi.org/10.1109/mspec.2018.8352562
2. Lu, Z., Li, Z., Cao, J., He, R., Sun, Z.: Recent progress of face image synthesis (2017). https://arxiv.org/abs/1706.047173
3. Face-Swapping Porn: How a Creepy Internet Trend Could Threaten Democracy, Rolling Stone, 4.18.18. https://www.rollingstone.com/culture/features/face-swapping-porn-how-creepy-trend-could-threaten-democracy-w518929
4. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation, ICLR 2018. http://research.nvidia.com/publication/2017-10_Progressive-Growing-of
5. Antipov, G., Baccouche, M., Dugelay, J.-L.: Face Aging With Conditional Generative Adversarial Networks, May 2017. https://arxiv.org/abs/1702.01983
6. Milanova, M., Kountchev, R., Ford, C., Kountcheva, R.: Watermarking with inverse difference pyramid decomposition. In: International Signal Processing Conference, Dallas, USA, pp. 346–362 (2003)
7. Lin, X., et al.: Recent advances in passive digital image security forensics: a brief review. Engineering **4**, 29–39 (2018)
8. Rahmouni, N., Nozick, V., Yamagishi, J., Echizeny, I.: Distinguishing computer graphics from natural images using convolution neural networks. In: IEEE Workshop on Information Forensics and Security, pp. 1–6 (2017)
9. Bharati, A., Singh, R., Vatsa, M., Bowyer, K.: Detecting facial retouching using supervised deep learning. IEEE Trans. Inf. Forensics Secur. **11**(9), 1903–1913 (2016)
10. Raghavendra, R., Raja, K., Venkatesh, S., Busch, C.: Transferable Deep-CNN features for detecting digital and print-scanned morphed face images. In: IEEE Computer Vision and Pattern Recognition Workshops, pp. 10–18 (2017)
11. Zhou, P., Han, X., Morariu, V., Davis, L.: Two-stream neural networks for tampered face detection. In: IEEE Computer Vision and Pattern Recognition Workshops, pp. 1831–1839 (2017)
12. https://www.engadget.com/2018/04/11/machine-learning-face-swaps-xceptionnet/

13. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics: a large scale video dataset for forgery detection in human faces. CV Cornell University Library, March 2018
14. Souza, D., Yampolskiy, R.: Natural vs artificial face classification using uniform local directional patterns and wavelet uniform local directional patterns. In: IEEE CVPRW, pp. 27–33 (2014)
15. Dang-Nguyen, D.-T.: Discrimination of Computer Generated versus Natural Human Faces, February 2014. http://eprints-phd.biblio.unitn.it/1168/
16. Anwar, S., Milanova, M.: Real time face expression recognition of children with autism. IAEMR **1**(1) (2016)
17. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., et al.: Active shape models-their training and application. Comput. Vis. Image Underst. **61**, 38–59 (1995)
18. Ekman, P., Friesen, W.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto (1978)
19. Vij, K., Singh, Y.: Enhancement of images using histogram processing techniques. Int. J. Comput. Tech. Appl. **2**(2), 309–313. ISSN: 2229-6093
20. Harris, B.: Entropy. In: Balakrishnan, N., Read, C.B., Vidakovic, B. (eds.) Encyclopedia of Statistical Sciences, vol. 3, 2nd edn., pp. 1992–1996. Wiley, New York (2006)
21. Rich, R., Tracy, J.: The relationship between expected inflation, disagreement, and uncertainty: evidence from matched point and density forecasts. Staff Report No. 253, Federal Reserve Bank of New York. (Revised version published in Review of Economics and Statistics **92**(2010), 200–207 (2006)
22. http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/VELDHUIZEN/node18.html