Temporal Consistency Based Deep Face Forgery Detection Network

Chunlei Peng[1](✉), Wenbo Zhang[1], Decheng Liu[2], Nannan Wang[3], and Xinbo Gao[2,4]

[1] State Key Laboratory of Integrated Services Networks,
School of Cyber Engineering, Xidian University, Xi'an 710071, China
clpeng@xidian.edu.cn

[2] Video and Image Processing System Laboratory, School of Electronic Engineering,
Xidian University, Xi'an 710071, China

[3] State Key Laboratory of Integrated Services Networks,
School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

[4] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts
and Telecommunications, Chongqing 400065, China

Abstract. With the rapid development of deep learning techniques as well as increasingly more visual information being made publicly available on the Internet, image translation methods have achieved great progress and encouraging performance. The manipulation and fabrication of visual information has become accessible and difficult to distinguish by the naked eye, which will have adverse effects on cloud and communication security. Thus, face forgery detection techniques have recently attracted increasing attention. Most recent works regard the face forgery detection problem as the typical image classification task, ignoring the exploration of inherent properties of forgery visual information itself. In this paper, we first explore the inherent limitation of fake videos, and find that the temporal consistency could help distinguish fake faces from real faces. A temporal consistency based deep face forgery detection network is proposed to directly detect fake videos when given multiple consistent video frames. The proposed method effectively considers the frame consistency property and achieves promising detection performance. Experimental results on the face forgery detection dataset demonstrate the superior performance of the proposed method.

Keywords: Cloud security · Visual information analysis · Face forgery detection

1 Introduction

In cloud and communication security, facial visual information has drawn increasing attention because of its convenience and safety. Especially for face recognition tasks, many recent works [1,2] have achieved encouraging superior

X. Chen et al. (Eds.): ML4CS 2020, LNCS 12488, pp. 55–63, 2020.
https://doi.org/10.1007/978-3-030-62463-7_6

performance, even when facial visual information is captured with different sensors [3–7]. It is due to the rapid development of deep learning techniques and the availability of a large amount of cloud visual information on the Internet. In addition, image translation tasks are also becoming an important topic in computer vision and machine learning. Generative adversarial network methods [8–10] have yielded fine realistic textures and have further improved the quality of generated images. Because of the wide application of these image translation techniques, manipulating facial videos and images has become easier, and the results are becoming harder to distinguish by human eyes. These image forgery manipulation techniques can not only generate fake images or videos, but also create fake news and scams. There is no doubt that social media will make the propagation of fake visual information more convenient. Thus, it is indeed an important task to develop an effective forgery detection method for cloud and communication security.

In the early stage, the manipulation of visual content in the media requires complex sophisticated editing tools, and high image manipulation expertise. Additionally, the forgery of these videos is always time consuming, and the degree of realism is limited. For example, [11] utilized existing footage to create a new video of a person mouthing words, which tracked points on the mouth of the speaker. However, due to the increasing high computing power and rapid progress of machine learning applications, visual content manipulation is becoming easier. The end-to-end deep learning technique reduces the computational time. Existing face forgery detection methods can be roughly categorized into two categories: the traditional classification based methods [12–15] and deep learning based methods [16–19]. These methods mostly consider the face forgery detection task as the common binary classification task, and their aim is usually to train a strong and robust classifier to accurately distinguish fake images from real images.

With the development of deep learning, researchers have found that the convolutional neural networks can learn extremely powerful image features for classification, which indeed further promotes the face forgery detection field. Although recent deep learning based methods have achieved superior detection performance, these works do not consider the inherent properties of forgery media. Thus, the performance of these face forgery detection techniques is not good enough to be widely applied in the real world.

To address this challenge, we propose a temporal consistency based deep face forgery detection network. Our method first analyzes the frame consistency of forgery media, and finds the difference between fake videos and real videos. Then we design a temporal consistency based deep face forgery detection network to directly distinguish fake videos with an end-to-end network model. A novel objective function is designed to integrate the consistency information for better detection performance.

The main contributions of this paper are summarized as follows:

1. The proposed temporal consistency based framework considers the inherent characteristics of fake face videos, and experimental results prove its efficiency.

2. We design a temporal consistency based deep face forgery detection network that can effectively integrate frame temporal consistency to boost the detection performance. Furthermore, the backbone network of our method can be extended for better performance.
3. Experimental results on the UADFV dataset illustrate the superior performance of the proposed method.

We organize the rest of this paper as follows. Section 2 gives a review of face forgery detection works. Section 3 presents the temporal consistency based deep face forgery detection network. Section 4 shows the experimental results and provides an analysis of the algorithm, and the conclusion is drawn in Sect. 5.

2 Related Work

In this section, we review the face forgery detection methods in the aforementioned categories: traditional classification based methods and deep learning based methods.

In the early stage, researchers always focused on the biometric feature extraction to detect face forgeries [12,20]. [13] presented a holistic liveness detection paradigm that collaborated with face biometrics. Considering the lighting differences in optical flow fields generated by movements, [21] utilized the degree of differences between the two fields to distinguish a real face from a photograph. [14] proposed a method for masked fake face detection that utilized reflectance disparity. [15] captured facial physiological patterns with the bioheat information contained in the thermal images for face forgery analysis.

Recent works in the computer vision filed prove that convolutional neural networks can effectively extract strong and robust visual features for downstream tasks. Thus, deep learning based forgery detection methods have attracted increasing attention. [16] presented the incremental learning based method for the classification of GAN generated images, where multiple binary classifiers are utilized for the detection task. [22] found that fake videos are always created by splicing synthesized face regions into holistic regions, and proposed a novel method to estimate 3D head poses as features to distinguish fake videos. [17] directly targeted the artifacts in affine face warping to distinguish real and fake images, without using any generated images as training data. [18] presented a novel method to expose fake face videos generated and utilized the detection of eye blinking to detect generated videos. [19] combined a recurrent convolutional neural network model and face alignment approach to improve detection performance.

Inspired by previous works, we find that the key is to explore the inherent property of the generated fake face videos and train a strong binary classifier. In this work, we aim to train an end-to-end deep network to directly distinguish fake face videos, where temporal fame consistency information is considered as the clue.

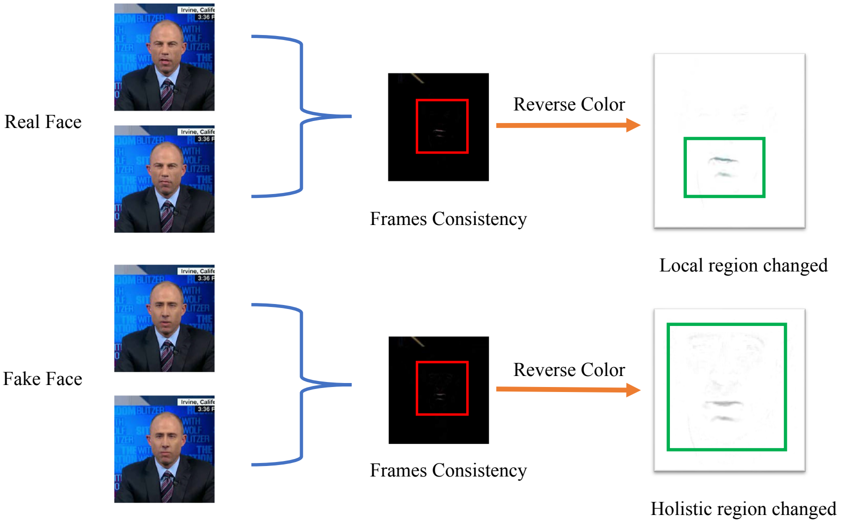

Fig. 1. The temporal consistency analysis results of face forgery videos.

3 Proposed Method

In this section, we present a novel temporal consistency based deep face forgery network to distinguish fake face videos. In the subsection, we will first describe the motivation, and then give a detailed explanation of our proposed method.

3.1 Motivation

In real-world scenarios, fake face videos are always generated from a deep network based image translation model. Naturally, our assumption is that the difference in temporal consistency will help to detect fake videos because these fake face videos are usually created with consistently generated with consistent generated frames from the image translation model, where the temporal relationship is ignored. In this work, we regard the temporal consistency of generated videos as the inherent characteristic for face forgery detection.

For convenience, we directly calculate the difference between the previous frame and the subsequent frame to find the distinct feature. As shown in Fig. 1, when we analyze the frames consistency of both real and fake videos, we find that there exists only a local changed region in the real videos, but a holistic appearance changed region is found in fake videos because the fake videos are generated by combining individual fake frames, which makes it hard to control the temporal consistency of face poses, expressions and countenances. Thus, we regard the temporal consistency of face videos as the distinguishing feature for forgery detection. In addition, the inherent features of the original videos themselves also provide discriminative information. Next, we provide more details of our proposed temporal consistency based deep face forgery detection network.

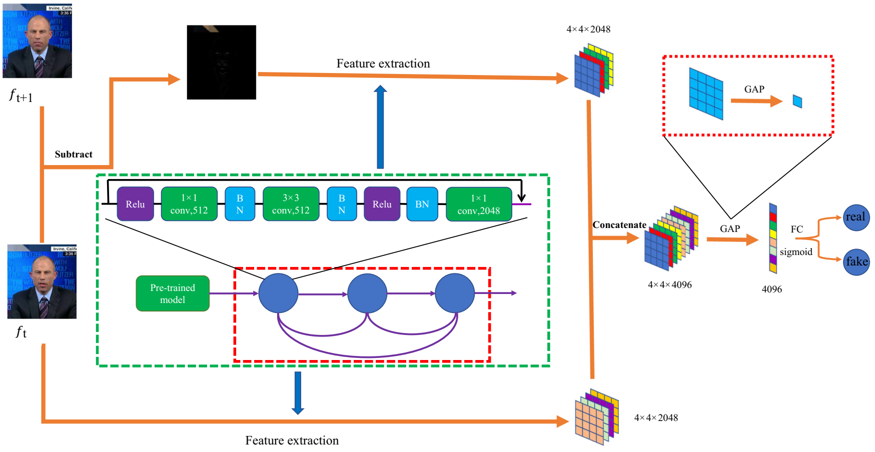

Fig. 2. Overview of the proposed temporal consistency detection based deep face forgery detection algorithm.

3.2 Temporal Consistency Based Deep Face Forgery Detection Network

In this subsection, we provide a detailed description of the proposed deep face forgery network. Here, we choose ResNet50 as the backbone network to introduce our method. Figure 2 shows the framework of the proposed algorithm. We consider that the input face video contains consistent frames $\{f_t\}_{t=1}^{T}$, where T is the number of all frames in one video. To avoid overfitting problem, we utilize images from ImageNet dataset to pre-train the backbone network to improve the generalization ability of the extracted features. Inspired by related works, we set the parameters of former layers to effectively extract low-level common discriminative information.

We separately input the original frame f_t and the difference in consistent frames $f_t - f_{t-1}$ into two network branches at time t. Next, the pre-trained backbone network extracts the discriminative image features as shown in Fig. 2. We directly concatenate frame feature fea_t with the consistent frame feature fea_t^{cons} as the final fused distinguishing feature $[fea_t, fea_t^{cons}]$ for face forgery detection. A global average pool layer is utilized to fuse different features, and then a fully connected layer is designed for binary classification.

The objective function of our network is designed as follows:

$$L = E[-log(p(label|fea_t, fea_t^{cons}))], \tag{1}$$

where $label \in \{0, 1\}$. Here $p(label|fea_t, fea_t^{cons})$ is the predicted probability of the consistent frames fused feature at time t.

3.3 Implementation Details

The backbone network is pre-trained by images from the ImageNet dataset. Different backbone architectures can be adopted in the future for better performance. The proposed network is designed with the Keras platform and run on a GTX 1660Ti GPU. Here we use the root mean square prop algorithm to train the parameters in our network. The batch size is set as 10 in our experiments. The learning rate is set to 1e–3, and the number of epochs is set to 10.

4 Experiments

In this section, we evaluate the performance of our proposed temporal consistency based deep face forgery detection network on the UADFV dataset [22]. We will present details of experimental results and illustrate the effectiveness of the proposed method.

4.1 Dataset and Evaluation Metric

The UADFV dataset [22] contains 98 videos in total, with 49 real face videos and 49 fake face videos. The average length of the videos is 11.14 s, and the size of most frames is 294×500. With the same protocol as that proposed in [22], we select 35 real face videos and 35 fake videos as training data which contains 23,981 frames in total, and we use the remaining 14 real videos and 14 fake videos as testing data, which contain 10,241 frames in total.

To further evaluate the face forgery detection performance to mimic real-world scenarios, we utilize the evaluation metrics mentioned in [22]: individual frames based metric. Here, we consider the face forgery detection task a binary classification task. Naturally, the individual frames of videos could be evaluated as units with the area under the ROC curves (AUC) as the evaluation metric. In the following experiment, we all choose the AUC to analyze individual frames as the performance metric.

4.2 Experimental Results

As mentioned before, we evaluate four different backbones of our proposed method on the UADFV dataset. Here we choose four common pre-trained models VGG16, VGG19, ResNet50 and ResNet101 as the backbone networks to prove the generalization ability and efficiency of our method. As shown in Fig. 3, the VGG16, VGG19, ResNet101 and ResNet152 models achieveed AUCs of 87.9%, 83.6%, 97.9% and 98.1% respectively. It is noted that the recent state-of-the-art method [22] achieves an AUC of only 89.0%. Compared with the state-of-the-art algorithm [22], our proposed method increases the AUC by 9.1%. It is because our method considers the inherent temporal consistency property of face forgery videos, and the deep networks could indeed extract strong discriminative features for forgery detection.

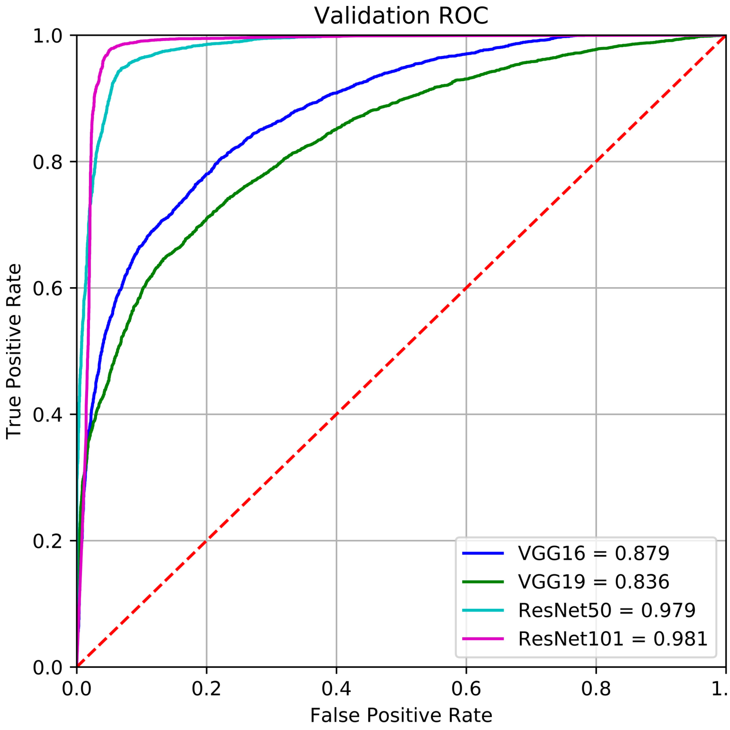

Fig. 3. ROC curves of the different backbone networks of our proposed method on the UADFV dataset.

4.3 Ablation Study

In this subsection, we further analyze the proposed temporal consistency based deep face forgery detection network. In the ablation study, we compare our network with the networks after removing the temporal consistency branch. As shown in Table 1, AUCs of all the CNN models apparently decrease after removing the temporal consistency network branch. Using a VGG16 backbone

Table 1. The experimental results (AUC) of the proposed temporal consistency based deep face forgery detection network with different backbones on the UADFV database.

Models	W/O temporal consistency	Final performance
Method [22]	/	89.0%
Proposed with VGG16	85.3%	87.9%
Proposed with VGG19	81.5%	83.6%
Proposed with ResNet50	96.6%	97.9%
Proposed with ResNet101	97.0%	98.1%

in our proposed algorithm even decreases the AUC by 26%. This finding demonstrates that our proposed temporal consistency based deep face forgery detection network can help integrate the inherent temporal consistency information, and boost the face forgery detection performance.

5 Conclusion

A novel temporal consistency based deep face forgery detection network is proposed in this paper. The proposed method explores the inherent temporal consistency property of face forgery videos, and design an effective two-branches deep forgery detection network to fuse different discriminative features for better performance. Benefiting from the proposed temporal consistency discriminative features, our algorithm outperforms the state-of-the-art method by achieving a superior AUC. Additionally, we explore different backbone networks of our framework to show the generalization ability of our method. In the future, we will try to integrate more inherent characteristics of face forgery videos, and evaluate our proposed method on more datasets.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0801001, in part by the National Natural Science Foundation of China under Grant 61806152, Grant 61876142, Grant 61432014, Grant U1605252, Grant 61772402, and Grant 61671339, in part by the Fundamental Research Funds for the Central Universities under Grant JB190117 and JB191502, in part by Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2019JM-289, in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, in part by the Key Industrial Innovation Chain in Industrial Domain under Grant 2016KTZDGY04-02, in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001, in part by the Young Elite Scientists Sponsorship Program by CAST under Grant 2016QNRC001, in part by the Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by the CCF-Tencent Open Fund, in part by the China 111 Project under Grant B16037, in part by the Key Research and Development Program of Shaanxi under Grant 2020ZDLGY08-08, in part by the China Post-Doctoral Science Foundation under Grant 2018M631124 and 2019T120880, and in part by the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence.

References

1. Wang, H., et al.: Large margin cosine loss for deep face recognition. In: IEEE Conference Computer Vision Pattern Recognition, pp. 5265–5274 (2018)
2. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: IEEE Conference Computer Vision Pattern Recognition, pp. 4690–4699 (2019)
3. Liu, D., Li, J., Wang, N., Peng, C., Gao, X.: Composite components-based face sketch recognition. Neurocomputing **302**, 46–54 (2018)

4. Liu, D., Gao, X., Wang, N., Li, J., Peng, C.: Coupled attribute learning for heterogeneous face recognition. IEEE Trans. Neural Netw. Learn. Syst. (2020)
5. Peng, C., Wang, N., Li, J., Gao, X.: Dlface: deep local descriptor for cross-modality face recognition. Pattern Recogn. **90**, 161–171 (2019)
6. Peng, C., Gao, X., Wang, N., Li, J.: Graphical representation for heterogeneous face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(2), 301–312 (2017)
7. Peng, C., Gao, X., Wang, N., Li, J.: Sparse graphical representation based discriminant analysis for heterogeneous face recognition. Signal Process. **156**, 46–61 (2019)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference Computer Vision Pattern Recognition, pp. 1125–1134 (2017)
9. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint (2017)
10. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: unified generative adversarial networks for multi-domain image-to-image translation. arXiv preprint, 1711 (2017)
11. Bregler, C., Covell, M., Slaney, M.: Video rewrite: driving visual speech with audio. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, pp. 353–360 (1997)
12. Galbally, J., Marcel, S., Fierrez, J.: Biometric antispoofing methods: a survey in face recognition. IEEE Access **2**, 1530–1552 (2014)
13. Kollreider, K., Fronthaler, H., Bigun, J.: Verifying liveness by multiple experts in face biometrics. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–6. IEEE (2008)
14. Kim, Y., Na, J., Yoon, S., Yi, J.: Masked fake face detection using radiance measurements. JOSA A **26**(4), 760–766 (2009)
15. Buddharaju, P., Pavlidis, I.T., Tsiamyrtzis, P., Bazakos, M.: Physiology-based face recognition in the thermal infrared spectrum. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 613–626 (2017)
16. Marra, F., Saltori, C., Boato, G., Verdoliva, L.: Incremental learning for the detection and classification of gan-generated images. In: International Workshop on Information Forensics and Security (2019)
17. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2019)
18. Li, Y., Chang, M.C., Lyu, S.: In ICTU oculi: exposing AI created fake videos by detecting eye blinking. In: International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE (2018)
19. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, E., Masi, I., Natarajan, P. Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) **3**, 1 (2019)
20. Hadid, A., Evans, N., Marcel, S., Fierrez, J.: Biometrics systems under spoofing attack: an evaluation methodology and lessons learned. IEEE Signal Process. Mag. **32**(5), 20–30 (2015)
21. Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: International Conference on Image Analysis and Signal Processing, pp. 233–236. IEEE (2009)
22. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265. IEEE (2019)