Deepfake Video Detection through Optical Flow based CNN

Irene Amerini[♢], Leonardo Galteri[♢], Roberto Caldelli[♢♣], Alberto Del Bimbo[♢]
[♢]Media Integration and Communication Center (MICC), University of Florence, Florence, Italy
[♣]National Inter-University Consortium for Telecommunications (CNIT), Parma, Italy

Abstract

Recent advances in visual media technology have led to new tools for processing and, above all, generating multimedia contents. In particular, modern AI-based technologies have provided easy-to-use tools to create extremely realistic manipulated videos. Such synthetic videos, named Deep Fakes*, may constitute a serious threat to attack the reputation of public subjects or to address the general opinion on a certain event. According to this, being able to individuate this kind of fake information becomes fundamental. In this work, a new forensic technique able to discern between fake and original video sequences is given; unlike other state-of-the-art methods which resorts at single video frames, we propose the adoption of optical flow fields to exploit possible inter-frame dissimilarities. Such a clue is then used as feature to be learned by CNN classifiers. Preliminary results obtained on FaceForensics++ dataset highlight very promising performances.*

1. Introduction

Deep learning techniques are escalating technology sophistication regarding creation and processing of multimedia contents. A new phenomenon, known as *Deep Fakes* (DF), has recently emerged: it permits to quite simply create realistic videos where people faces, or sometimes only lips and eyes movements, are modified in order to likely simulate the presence of another subject in a certain context or to make someone speak coherently with a different and, probably compromising, speech. The effects can be straightforwardly imagined when this fake information is deliberately used to harm a person such a public figure or a politician, or even an organization like a political party. The impact of Deep Fakes can also be amplified by the action of social networks that deliver information quickly and worldwide. According to this, machine learning community has dedicated a particular and twofold attention to this phenomenon. From one side, an effort has been spent to develop new kinds of effective synthesized video generation techniques such as Face2Face [14], Deep Video Portraits [7], StarGAN [5] and Deep Fake[1]. From another side, various studies have lastly focused on the problem to detect deepfake-like videos; most of them by analyzing possible inconsistencies within RGB frames of the video [9, 10, 1]. Usually, well established and pre-trained CNN techniques are directly applied to learn distinctive features from each single frame of the sequence. In [11], a recurrent convolutional strategy is used for face manipulation detection where a group of frames is evaluated as an ensemble. Other approaches consider physical characteristics like the work in [8] where the authors propose a detection of eye blinking to expose generated fake face videos and in [2] where facial expression is modeled in order to distinguish a fake speaking pattern from natural one.

In this extended abstract, a new technique able to detect deepfake-like videos from original ones is introduced. In particular, unlike state-of-the-art methods which usually act in a frame-based fashion, we present a sequence-based approach dedicated to investigate possible dissimilarities in the temporal structure of a video. Specifically, optical flow fields have been extracted to exploit inter-frame correlations to be used as input of CNN classifiers.

The paper layout is the following: Section 2 describes the proposed methodology by discussing the usage of motion vector fields while Section 3 discusses some preliminary experimental results; finally, Section 4 draws conclusions.

2. Proposed method

In this section the proposed method, whose basic architecture is depicted in Figure 1, is described. Such a structure has been built up to understand the actual effectiveness of *optical flow fields* to distinguish a deepfake from an original video. Optical flow [4, 3] is a vector field which is computed on two consecutive frame $f(t)$ and $f(t+1)$ to extract apparent motion between the observer and the scene itself. In particular, our hypothesis is that the optical flow is able to exploit discrepancies in motion across frames synthetically created with respect to those naturally generated by a video camera. It should be more appreciable in the

[1]Deepfakes: github.https://github.com/deepfakes/faceswap.

optical flow matrices, the introduction of fake and unusual movements of the lips, eyes and in general of the whole face. So, for this reason, for each frame $f(t)$, at a certain time t, a forward flow $\mathbf{OF}(f(t), f(t+1))$ is extracted using the CNN model for optical flow called PWC-Net [13]. This technique is based on pyramidal processing and warping and on the use of a cost volume processed by the CNN itself to estimate the optical flow. Successively (see Figure 1), the computed forward flow $\mathbf{OF}(f(t), f(t+1))$ is given as input to a semi-trainable CNN named *Flow-CNN*, based on some pre-trained network. In our experiments we have tested VGG16 [12] and ResNet50 [6] as backbones.

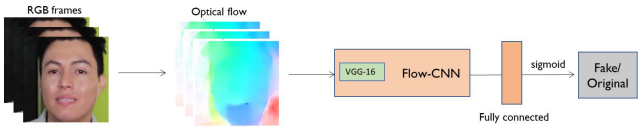

Figure 1. The proposed architecture.

Since the dimension, in terms of number of samples, of the common datasets for the deepfake identification task is not sufficient to train a net from scratch, we adopt the transfer learning technique on a portion of the net while only the rest of the net is fine-tuned on the deepfake dataset. So during fine-tuning, the first layers of the network have been fixed while the last convolutional layers together with the dense ones have been trained. After that a final fully connected layer with one output unit followed by a sigmoid activation is placed at the end of the net for the binary classification for each frame (fake or original). To exploit existing implementations and pre-trained networks trained on raw RGB images, the optical flow is transformed to a 3-channel image using a fixed color-coding approach. The color of pixels is determined by the angle between the flow vector and the horizontal axis, while the intensity of the motion is encoded by the saturation of the color.

3. Preliminary experimental results

In this section some preliminary experimental results are introduced to evaluate the goodness of the proposed concept. In Figure 2, an example of the optical flow fields calculated on two consecutive frames for an original video (left side) and for the corresponding deepfake one (right side) are pictured respectively to just provide a visual inspection. As expected, it can be noticed that the motion vectors around the chin in the real sequence are more noisy in comparison with those of the altered video that appear much smoother. On this basis, we have tried to verify if such a clue can be properly learned by a neural network. To do so, a generic net has been trained on randomly left-right flipped squared patches of size 224×224 pixels randomly chosen on a bigger patch of 300×300 containing the face; for the training, we used Adam optimizer with 10^{-4} learning rate, default momentum values and a batch size of 256. We run our experiments on the FaceForensics++ dataset proposed in [10]; the set consists of 1000 original video sequences that have been manipulated with three automated face manipulation methods: Deepfakes, Face2Face and FaceSwap. 720 of the videos are used for training, 120 for validation and another 120 for testing.

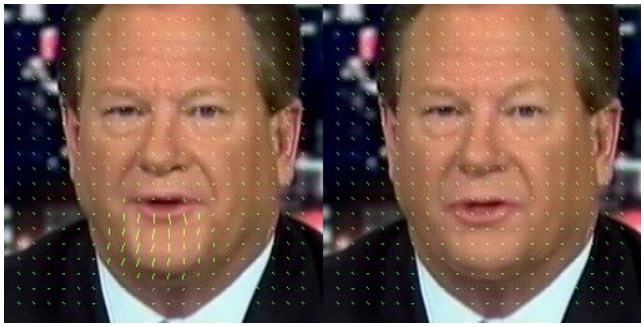

Figure 2. Optical flow for original (left) and deepfake (right) videos.

	VGG16	ResNet50
Face2Face	81.61%	75.46%

Table 1. Binary detection accuracy (%) of our architectures on Face2Face manipulation.

Preliminary results on two implemented networks (VGG16 and ResNet50) are presented in Table 1. They are obtained on the whole testset of FaceForensics++ for the manipulation Face2Face and witness that the method is able to distinguish the two kinds of videos.

4. Conclusions

In this work, the idea to exploit optical flow field dissimilarities as a clue to discriminate between deepfake videos and original ones has been introduced and investigated. This is a very innovative attempt to take into account possible anomalies in the temporal dimension of the sequence. In this initial experiments, to solve the problem to use pretrained network, motion vectors have been represented as 3-channels image and then considered as input for a neural network. Preliminary results, obtained on FaceForensics++ dataset with different types of networks, are very promising and show that this kind of feature seem to be able to point out some existing dishomogeneities between the two analyzed cases. This evidence paves the way for many possible future works: firstly, it is to evaluate the reliability of optical flow fields for deepfake video identification by testing against more datasets and with other neural networks, secondly, it would be interesting to study how this approach which exploits inconsistencies on the temporal axis can be combined with well-known state-of-the-art frame-based methodologies to improve their performances.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. pages 1–7, 12 2018. 1

[2] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. Protecting world leaders against deep fakes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1

[3] L. Alparone, M. Barni, F. Bartolini, and R. Caldelli. Regularization of optic flow estimates by means of weighted vector median filtering. *IEEE Transactions on Image Processing*, 8(10):1462–1467, Oct 1999. 1

[4] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Comput. Surv.*, 27(3):433–466, Sept. 1995. 1

[5] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, abs/1711.09020, 2017. 1

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4):163:1–163:14, July 2018. 1

[8] Y. Li, M. Chang, and S. Lyu. In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking. *CoRR*, abs/1806.02877, 2018. 1

[9] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179, 2018. 1

[10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. *CoRR*, abs/1901.08971, 2019. 1, 2

[11] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos, 05 2019. 1

[12] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014. 2

[13] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[14] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner. Demo of face2face: Real-time face capture and reenactment of RGB videos. In *ACM SIGGRAPH 2016 Emerging Technologies*, SIGGRAPH '16, pages 5:1–5:2, New York, NY, USA, 2016. ACM. 1