Pattern Recognition Letters 146 (2021) 31–37

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Optical Flow based CNN for detection of unlearnt deepfake manipulations

Roberto Caldelli[a,d,*], Leonardo Galteri[b], Irene Amerini[c], Alberto Del Bimbo[b]

[a] Universitas Mercatorum, Rome, Italy
[b] Media Integration and Communication Center, University of Florence, Florence, Italy
[c] Sapienza University of Rome, Rome, Italy
[d] National Inter-University Consortium for Telecommunications (CNIT), Parma, Italy

ARTICLE INFO

Article history:
Received 5 July 2020
Revised 22 February 2021
Accepted 8 March 2021
Available online 17 March 2021

MSC:
41A05
41A10
65D05
65D17

Keywords:
Deepfake manipulations
Optical Flow
Video forensics
CNN

ABSTRACT

A new phenomenon named Deepfakes constitutes a serious threat in video manipulation. AI-based technologies have provided easy-to-use methods to create extremely realistic videos. On the side of multimedia forensics, being able to individuate this kind of fake contents becomes ever more crucial. In this work, a new forensic technique able to detect fake and original video sequences is proposed; it is based on the use of CNNs trained to distinguish possible motion dissimilarities in the temporal structure of a video sequence by exploiting optical flow fields. The results obtained highlight comparable performances with the state-of-the-art methods which, in general, only resort to single video frames. Furthermore, the proposed optical flow based detection scheme also provides a superior robustness in the more realistic *cross-forgery* operative scenario and can even be combined with frame-based approaches to improve their global effectiveness.

1. Introduction

The growth of social networks together with the increasing number of their users have determined a new and very usual way to get, produce and share information. The web is full of malevolent image and video productions like posts, photoshopped pictures and counterfeited videos which are more or less credible. The *Deepfakes* (DF) phenomenon, has recently emerged among others and it obtained a remarkable attention by the media. In fact, through the use of *Deepfakes* techniques it is relatively easy to create synthetic but realistic videos; people faces, or sometimes only lips and eyes movements, can be simply modified in order to likely simulate the presence of another subject in a certain context or to alter someone's speech. The effects of this serious threat can be evident especially when this fake information is deliberately used to blackmail a person such as a public figure or a politician, or even a private citizen. The most widely adopted application of *Deepfakes* has probably been in the production of fake human faces. The ever-impressive level of realism reached by *Deepfakes* exposes public figures and society in general to a new menace. With the spread of social media, fake images or videos can easily go viral, giving a tremendous impact to fake news. As such, being able to argue if a human face is real or fake becomes a need to which image forensics community is asked for responding very quickly. Even though, computer-generated content has been used for a long time now, the recent development of deep learning techniques has attracted a lot of attention to this phenomenon from two different points of view. From one side, developing new kinds of effective synthesized video and image generation techniques from Face2Face [25] to Neural Textures [24] and DeepFakes,[1] StarGAN [6], ProGAN [13] and StyleGAN [14] among others. From another side, a lot of efforts have been made to distinguish Deepfakes-like videos (or face images) from the pristine ones [8,10]. Interesting overview information can be found in the paper by Matern et al. [19], where current facial editing methods and several characteristic artifacts determined from their processing pipelines are reviewed and, similarly, in Verdoliva [26] that presents a survey of algorithms used to detect deepfakes. Most of the existing meth-

[*] Corresponding author at: National Inter-University Consortium for Telecommunications (CNIT), Parma, Italy.
E-mail address: roberto.caldelli@unifi.it (R. Caldelli).

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

https://doi.org/10.1016/j.patrec.2021.03.005
0167-8655/© 2021 Elsevier B.V. All rights reserved.

ods exploit possible inconsistencies within RGB frames of a video [1,20,21] and well established pre-trained CNN techniques are directly applied to learn distinctive features from each single frame of the sequence. In Güera and Delp [11] and Sabir et al. [22], recurrent convolutional strategies are used for face manipulation detection where a group of frames is evaluated as an ensemble. Other approaches consider physical characteristics like the works of Li et al. [16] and Yang et al. [28] where the authors propose a detection of eye blinking or inconsistencies in head poses to individuate fake-generated face videos respectively. In Agarwal et al. [2], facial expression is modeled in order to distinguish a fake speaking pattern from natural one. Finally, the recent work by Li et al. [17] proposes a new dataset called *Celeb-DF* and gives a comprehensive overview of the most promising deepfake detection methods including those of the authors based on face warping artifacts. These approaches show promising results, however robustness issues need to be properly taken into consideration; in fact, performances drop in relation to operations like compression and, above all, when those methodologies are evaluated in a *cross-forgery* scenario. The term *cross-forgery* indicates when a model trained on a specific forgery is required to work against another unknown one. The generalization ability of forensics methods towards other unseen types of generated fake content is taken into consideration in the works by Cozzolino et al. [7], Marra et al. [18] and Wang et al. [27] especially focusing on GAN generated images. In general, state-of-the-art *Deepfakes* video detection methods are based on static frames features that though well-performing when trained on a specific kind of attack (*same-forgery* scenario), they show bad performances in a *cross-forgery* scenario.

For this reason in this paper, a new technique able to detect deepfakes-like videos generated by means of different kinds of manipulations that extends our preliminary work Amerini et al. [4], is introduced; in particular the *cross-forgery* scenario is in-depth investigated and the significant performances achieved by the proposed method are reported. We present a sequence-based approach oriented to investigate possible motion dissimilarities in the temporal structure of a video. Specifically, *optical flow* (OF) fields have been extracted to exploit inter-frame correlations to be used as input of a CNN classifier. In detail, the use of the optical flow based detection scheme seems to provide a gain in the robustness especially in the context of *cross-forgery* scenario. The idea which is behind the use of OF fields is that they constitute a structural part of the video sequence itself. OF takes into account of the movements of the objects belonging to the sequence, so it does not directly represent the visual part of the video frames which is altered by the deepfake techniques. Consequently, this should determine that diverse deepfake techniques, though differently acting on the visual part of the sequence, tend to similarly modify its structural part represented by the motion vector field. Furthermore, also static frames-based detection schemes seem to benefit from the merge with optical flow-based features in such a case. Experimental tests have been carried out, in particular, in order to investigate the actual behavior of the different methodologies to get rid of diverse kinds of deepfake attacks possibly unknown during the training phase. In fact, it could happen, in a real-world scenario, that a certain deep network, trained to detect a specific deepfake-like manipulation (e.g. *FaceSwap - FS*), faces a video sequence that is unreal but that has artificially been generated by resorting to another technique (e.g. *NeuralTextures - NT*). In this case, the classifier has never seen this type of artifacts and, above all, has not learnt on these during the training process. Notwithstanding with this, the system should be able to recognize such an example as fake though presenting slightly different visual features from those ones usually known. Moreover, additional experiments have been considered in a similar circumstance when the neural network is trained at the same time on more than one kind of deepfake-like attacks. In this situation, the classifier could already have seen this kind of manipulation, if it belongs to those learnt during training, but it might fall again in the previous operative case when the fake video to be evaluated has been crafted by means of a brand-new procedure even unpublished until then. On this basis, it is evident the importance to investigate the *cross-forgery* scenario, in order to better comprehend the effective resilience of deepfake detection methods and how the one here proposed based on OF fields could help in this quite unexplored direction. The main contributions provided in this work are the following:

- to introduce a novel method that exploits optical flow fields to detect deepfake videos;
- to present a solution to the need to integrate the bi-dimensional optical flow fields with pre-trained network usually receiving three channels inputs;
- to in-depth analyse the *cross-forgery* scenario and demonstrate that the proposed solution can provide superior performances in such a context with respect to the state-of-the-art methods.

The paper layout is the following: Section 2 discusses the usage of optical flow fields in relation to our objective, while in Section 3 the proposed method is outlined. Section 4 presents the experimental results in different operative scenarios and, finally, Section 5 draws conclusions.

2. Optical Flow fields for deepfake detection

In this section we briefly review the *Optical Flow* fields in order to motivate our choice to use this kind of feature to detect deepfake manipulation in videos. *Optical Flow* [3,5] is basically a vector field which is computed on two frames taken at two distinct temporal instants, that is $I(t)$ and $I(t+h)$ (usually $h=1$); it describes the apparent movements of the objects in a scene determined by the relative motion between the observer (e.g. the camera) and the scene itself. Motion is estimated by assuming two constraints: the first one concerns the brightness constancy of the scene, so that a point that has moved of Δx and Δy in a time interval Δt will conserve its intensity (see Eq. (1)),

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t) \tag{1}$$

while the second one assumes that the movements are small so that $I(x,y,t)$ can be developed in Taylor series (see Eq. (2)).

$$\begin{aligned} &I(x+\Delta x, y+\Delta y, t+\Delta t) \\ &\quad = I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + H.O.T. \end{aligned} \tag{2}$$

By neglecting the H.O.T. (higher order terms) and by using the first constraint of Eqs. (1) and (3) can be obtained

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0; \tag{3}$$

and then, by dividing by Δt, the well-known expression of the *Optical Flow* can be derived as in Eq. (4); where $V_x = \frac{\Delta x}{\Delta t}$ and $V_y = \frac{\Delta y}{\Delta t}$ represent the velocity components respectively, while I_x, I_y and I_t are the partial derivatives of $I(x,y,t)$ along the directions x, y and t.

$$I_x V_x + I_y V_y + I_t = 0 \tag{4}$$

Due to the fact that Eq. (4) contains two unknowns (V_x and V_y), it cannot be solved unless additional conditions are introduced. In the proposed method the technique based on TV-L1 optical flow estimation [29] has been adopted.

This method is based on the minimization of a functional containing a data term using the L^1 norm and a regularization term employing the total variation of the flow. The hypothesis behind

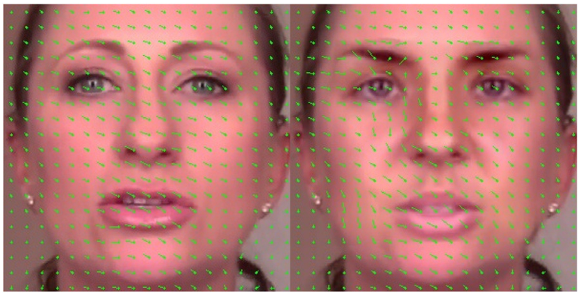


Fig. 1. Optical flow for original (left) and Deepfake (right) frames estimated with TV-L1 [29].

the usage of the optical flow in the proposed method is to take into account the information contained along the temporal axis of a frame sequence, as it usually happens when dealing with video in common applications like object detection and tracking or video coding. In fact, it is desirable that motion vectors will show a distinctiveness between synthetically created frames and naturally generated ones. It should emerge that the insertion of fake and unnatural movements of the lips, of the eyes and moreover of the whole face induce a distinctive pattern in the video with respect to the original case. A preparatory study has been carried out on various videos to investigate this issue and an example case is reported in Fig. 1. The OF field extracted from two consecutive frames of an original video (left side) and of the corresponding fake one (right side) are superimposed on each reference picture. It can be seen that the motion field on the original frame (left) is globally rather smooth and coherent. On the contrary, on the fake one (right), motion vectors belonging to the possibly modified areas (e.g. eyes, eyebrows and nose) present an incoherent pattern with respect to the other vectors of the frame located onto the peripheral parts of the face. These external parts of the face are generally unmodified and, in fact, they appear similar in both the situations (original and fake).

This lead to the fact that artificial facial elements should contain some intrinsic dissimilarities with respect to natural expressions and real faces that motion fields are able to properly grab.

3. The proposed Optical Flow based CNN method

In this section the proposed method, whose basic architecture is sketched in Fig. 2, is described. In the first phase of the pipeline, video frames are processed to estimate the optical flow fields that are then cropped according to a squared box of 300 × 300 pixels containing the speaker face. Such a bounding-box is computed on each frame by using *dlib* face detector [15]. Cropped OF fields are passed as input to a CNN whose final fully connected layer is represented by one output unit followed by a sigmoid activation used for the binary classification of each frame stating if such a frame is tampered or original. In our experiments, the well-known *ResNet50* [12] has been adopted as reference CNN. Different kinds of networks, such as VGG16 [23], had also been considered indeed in our preliminary study in Amerini et al. [4]; all of them have substantially provided similar results. However, it is out of the scope of this work to investigate on the CNN specific performances.

Going into details, the proposed net has been trained on randomly left-right flipped squared patches of size 224 × 224 pixels randomly chosen on the bigger patch of 300 × 300 containing the face, for data augmentation. Specifically for the training phase, we used Adam optimizer with 10^{-4} learning rate, default momentum values and a batch size of 256.

Using pre-trained networks is a reliable technique if not enough data are available for training. We benefit from such initialization as it helps to heavily mitigate the overfitting phenomenon and it determines even a faster convergence. When used with spatial (RGB) frames, we can employ directly *ResNet50* models trained on a large scale dataset such as *ImageNet* [9] but in our case we cannot feed optical flow frames to the available pre-trained networks as their distribution of values is very different from RGB data. To overcome this issue, we decide to bound the range of optical flow values to be the same of RGB frames. Therefore, we firstly clip OF values between -3 and 3 to eliminate outliers (this range is chosen to minimize the information loss) then we scale and discretize OF values into the range [0, 255] with a simple linear transformation.

As the input dimension of OF frames have only two channels (224 × 224 × 2), this does not match yet the pre-trained network requirements (three color channels), so we proceed to modify the weights of the first convolutional layer of the pre-trained network. Hence, we take the average (along the channels) of the weights of the first convolutional layer and the obtained weights are then replicated to become the new two-channels first layer of the network.

In addition to the proposed optical flow net described above in which the input to the net are OF cropped matrices (as evidenced in Fig. 2), we have also investigated in this paper, if typical training paradigm for Deepfakes detection can benefit from additional motion information provided by the optical flow estimation. For this reason (see Sections 4.3 and 4.4), we have independently trained two different networks having the same architecture, one just with optical flow (OF) frames and the other with spatial (RGB) frames following the idea of most of the state of the art methods. The two contributions derived from OF and from RGB nets are then combined together taken the average of the corresponding classifier outputs (such an approach has been named *MIX* in experimental Section 4).

4. Experimental results

In this section some experimental results are introduced to evaluate the effectiveness of the proposed methodology in different operative contexts. In particular, two distinct scenarios are basically considered: *same-forgery* and *cross-forgery*. The *FaceForensics++ (FF++)* dataset proposed in Rossler et al. [21] has been used for the experiments; it consists of 1000 original video sequences that have been manipulated with four face manipulation methods: two graphics rendering approaches *Face2Face* (F2F) and *FaceSwap* (FS) an the other two *DeepFakes* (DF) and *NeuralTextures* (NT), resorting to deep learning methods. DeepFakes and FaceSwap are two different methods for face replacement, while Face2Face and Neural Textures are two facial reenactment systems able to transfer the expressions of a source video to a target video while maintaining the identity of the target person. An amount of 740 videos is used for training, 120 for validation and another 120 for testing. The dataset is composed by three level-of-quality: uncompressed (C0) and compressed using the H.264 codec with a high visual quality (C23) level and a low visual quality (C40). It is worthy saying that the *FF+* dataset has been chosen because it has permitted to properly evaluate the actual performances of the optical flow in a cross-forgery scenario (see Sections 4.2 and 4.3).

4.1. OF-based approach in a same-forgery scenario

First of all, we have tested the proposed approach based on optical flow to understand if this new feature is able to highlight a distinctiveness between original and fake videos.

In Fig. 3, performances, in terms of accuracy, are pictured for the three kinds of video quality available in *FaceForensics++* dataset: C0 (yellow bars), C23 (blue bars) and C40 (red bars). In this case, the network has been trained and tested on the same

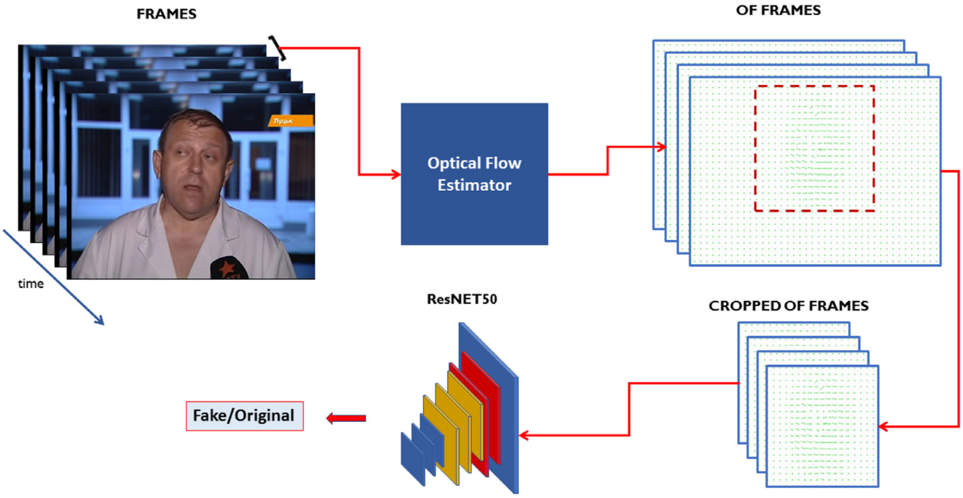


Fig. 2. The proposed pipeline. The TV-L1 [29] has been implemented as OF estimator.

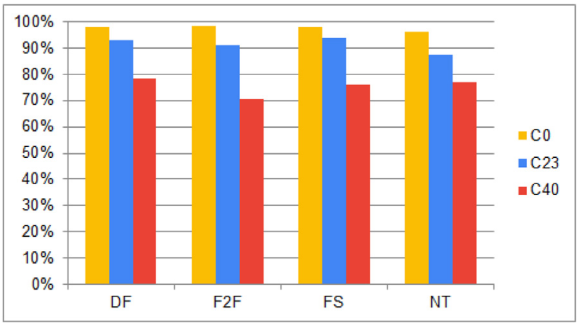


Fig. 3. Accuracy (%) of the proposed OF-based approach with respect to the four kinds of forgeries for the three diverse types of video quality: C0 (*yellow*), C23 (*blue*) and C40 (*red*).

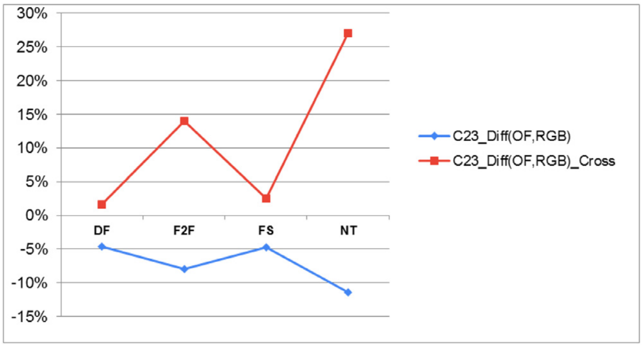


Fig. 4. Increment of accuracy (%) in the *cross-forgery* (red) and *same-forgery* scenario (blue) respectively (C23 case).

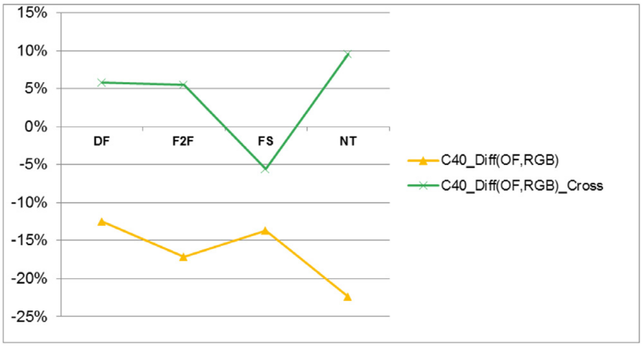


Fig. 5. Increment of accuracy (%) in the *cross-forgery* (green) and *same-forgery* scenario (yellow) respectively (C40 case).

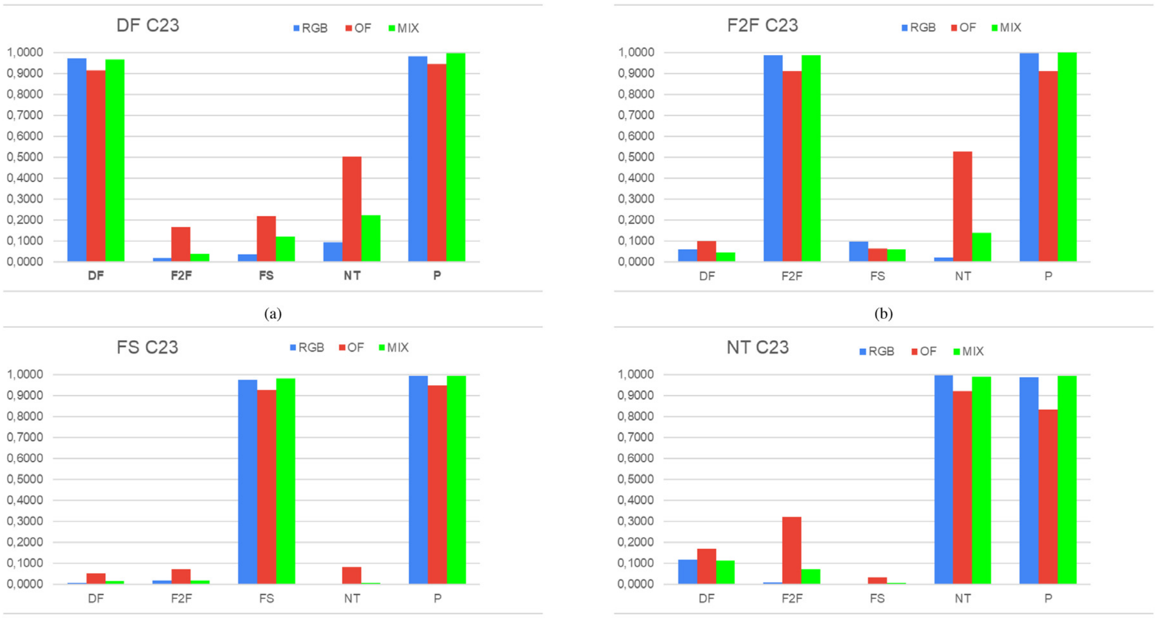
type of face manipulation (*same-forgery scenario*); it can be appreciated that achieved results are quite satisfactory for all the four forgeries: 97% for C0 and 91% for C23 is averagely obtained respectively, though accuracy is inferior for the circumstance of low video quality (C40) as expected (76% averagely). This witnesses that optical flow fields are a consistent feature to be learnt in order to detect deepfakes-like videos.

4.2. OF-based approach in a cross-forgery scenario

Going ahead, we have tried to understand if using OF-based features could help in a *cross-forgery* scenario, that is, when a model trained on a certain manipulation is asked to evaluate a video created by resorting to a diverse kind of forgery (e.g. F2F vs DF, F2F vs NT and so on) that has never seen before, as it often happens in real world. This issue is well-known as very challenging and also methods, such as the one based on RGB frames [21], usually presenting significant accuracy, drop their performances in this circumstance. In Figs. 4 and 5, a comparison, in terms of the achieved accuracy, is reported for the case C23 and C40 respectively.

In particular, in Fig. 4, the red line represents the average accuracy increment obtained by the OF-based method with respect to that one based on RGB spatial frames in the cross-forgery case (the manipulation used to train the model is reported on the x-axis and then it is tested on all the other manipulations). It can be pointed out that such an increment is always positive, sometimes small (as in DF and FS cases) but sometimes higher as in NT and F2F cases. On the contrary, a decrement is registered for the same-forgery cases (blue line). A similar behavior can be appreciated for the C40 case (see Fig. 5) where the increment is above 5% for three out of four forgeries, though with a reduced global effectiveness as expected.

However the performance increment in the cross-forgery cases (red and green lines) represents an interesting outcome that encouraged us to combine the two methods, RGB-based [21] and the proposed OF-based, as explained at the end of Section 3 and named as *MIX* in the next two subsections. The two techniques are equally balanced (i.e. with a weight of 0.5 each one) by weighing



(a)

(b)

(c)

(d)

Fig. 6. ResNet50: cross-forgery experiments on C23 dataset with neural networks trained on DF (a), F2F (b), FS (c) and NT (d). Accuracy achieved is pictured for frame-based method (RGB, blue bars), optical flow-based method (OF, red bars) and the mixed method, (MIX, green bars). *P* stands for pristine.

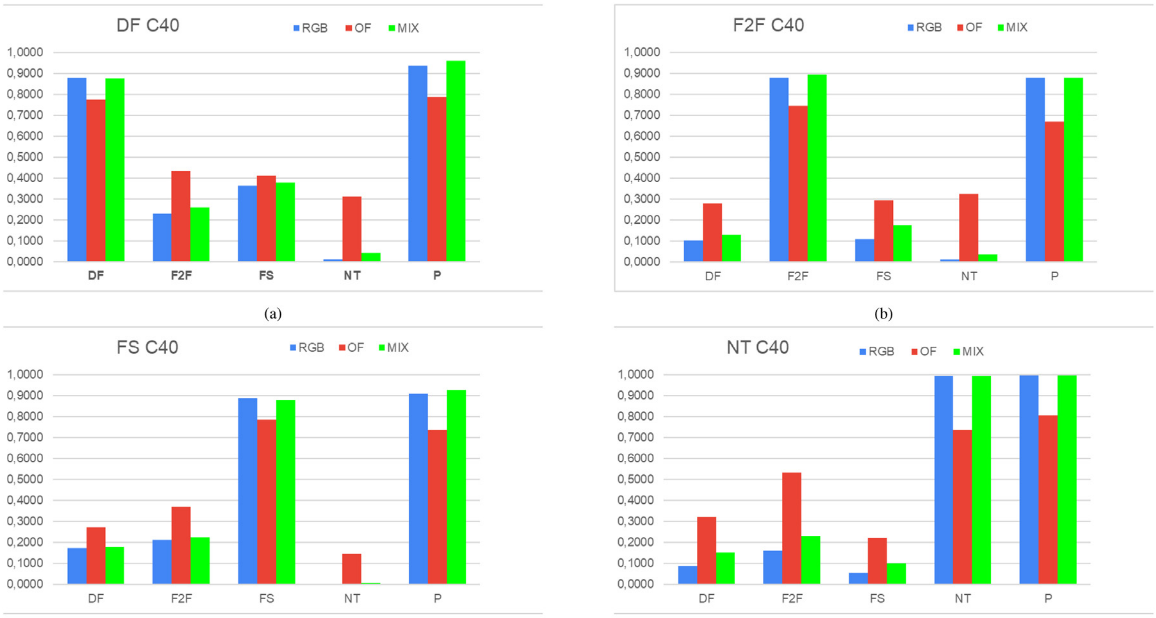
their outputs simply at the level of the final sigmoid function. The results obtained are discussed in the following subsections.

4.3. Results training on one manipulation and testing on all of them

In this subsection, we have investigated more in depth if the proposed idea to resort to OF fields can provide an advantage in a *cross-forgery* scenario. To do this, we have taken into account the following type of experiment: the classifier is binary and has been trained only on one kind of manipulation, for instance *FaceSwap - FS*, and on pristine examples of course, while, during the test phase, it will face, as in a real-world scenario, pristine videos and fake ones, but now these last ones have been generated both through the learnt method and also through other unknown techniques.

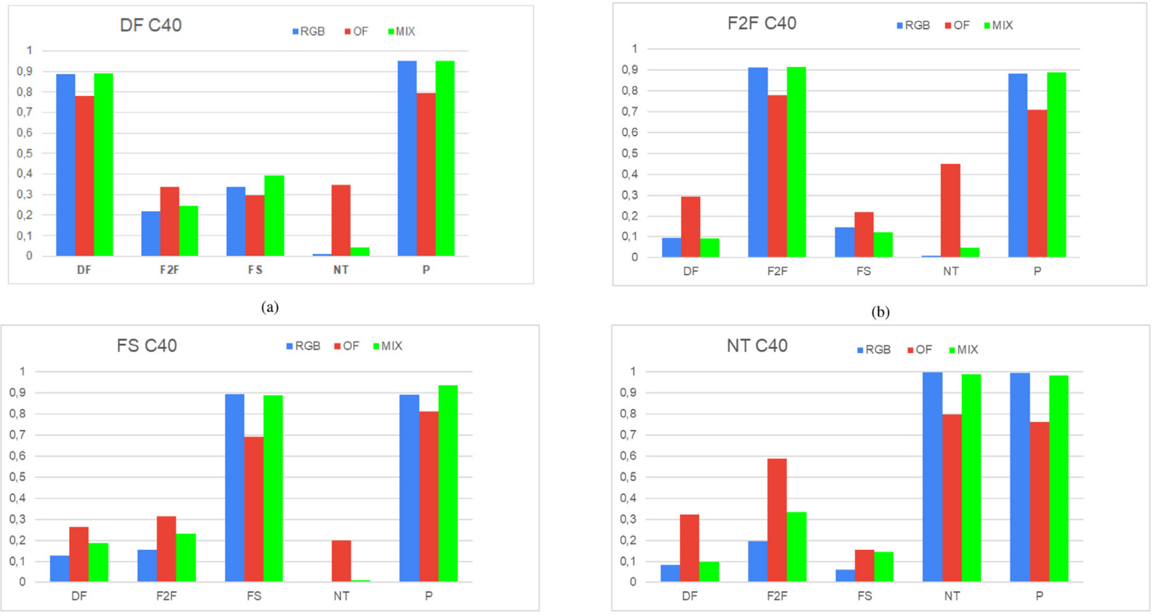
In Fig. 6, the values of accuracy obtained in the case of C23 dataset are pictured. The four graphs, going from Fig. 6(a) to (d), refer to the cases where the classifier has only been trained on *DF, F2F, FS* or *NT* manipulation respectively, as indicated by the title of each graph. Every colored bar represents the accuracy achieved by resorting to the frame-based method (RGB, blue bars) and at the OF-based one (OF, red bars) with respect to the diverse kinds of manipulations given at test time (on the x-axis, there are the four deepfake techniques, while *P* stands for pristine). The green bars values, labelled with MIX, are obtained by means of the combined approach. First of all, by looking at the blue and red bars, it can be observed that the frame-based method (blue) always outperforms the OF-based one (red) when the images to be classified are pristine or fake but generated by the same manipulation learnt during training (i.e. in a *same forgery* scenario). On the contrary, if we check the case when images, crafted with a not-learnt deepfake technique, are to be evaluated (i.e. in a *cross forgery* scenario), it can be appreciated that the situation is completely inverted: blue bars are always lower than red ones (except for a single case in Fig. 6(b) when the model trained on *F2F* is tested on *FS*). This seems to highlight that the OF-based method provide a superior robustness towards fake images created with methodologies unknown at training time though it appears slightly underperforming with respect to the frame-based approach in the classical *same forgery* scenario. The same trend is basically confirmed and also a bit more evidenced in Fig. 7 when the dataset C40 which contains images with lower quality is considered. Results achieved, for instance, in the case of *Neural Textures*-generated and unknown images is quite impressive (see Figs. 6 pictures (a) and (b)). A possible explanation for this interesting behaviour could be that OF represent the structural part of the video sequence by taking into account the movements of the objects belonging to the sequence; diverse deepfake techniques, though differently acting on the visual part of the sequence, tend to similarly alter its structural part represented by the motion vector field: this is detected by the models trained on OF features.

On the basis of such a finding, we have tried to understand if this apparent complementary behavior could be composed in order to get a general improvement. So we have mixed, as explained before, the two approaches and if we now look at the green bars in both the Figs. 6 and 7, we can effectively observe this phenomenon: the performances in the cases of the *same forgery* scenario (pristine and learnt manipulation) remain as very high as for the frame-based (RGB) method, sometimes even with a slight improvement (e.g. the average accuracy increase of 0.23% for the C23 dataset and of 0.59% for C40 one respectively). What is interesting is the increment in the cases of the *cross forgery* scenario where the accuracy is constantly augmented with respect to the values achieved by the frame-based method alone represented by the blue bars (e.g. the average accuracy increases of 3.14% for the C23 dataset and of 3.27% for C40 one respectively). As expected, the accuracy of the mixed approach (MIX), in these circumstances, is not as high as the OF-based technique by itself but it is generally intermediate between the two (OF and RGB). In order to better verify the reliability of such interesting behaviour, we have changed the underlying neural network from *ResNet50* to *XceptionNet* for all the three approaches RGB, OF and MIX and carried out again the experimental tests. In Fig. 8, the obtained results are reported for the C40 case (the case C23 is similar and is not pictured to avoid redundancy). It can be seen that performances are substantially the same as for the corresponding *ResNet50* case and good achievements for the cross-forgery situations are maintained. This



(a) (b) (c) (d)

Fig. 7. ResNet50: cross-forgery experiments on C40 dataset with neural networks trained on DF (a), F2F (b), FS (c) and NT (d). Accuracy achieved is pictured for frame-based method (RGB, blue bars), optical flow-based method (OF, red bars) and the mixed method, (MIX, green bars). *P* stands for pristine.



(a) (b) (c) (d)

Fig. 8. XceptionNet: cross-forgery experiments on C40 dataset with neural networks trained on DF (a), F2F (b), FS (c) and NT (d). Accuracy achieved is pictured for frame-based method (RGB, blue bars), optical flow-based method (OF, red bars) and the mixed method, (MIX, green bars). *P* stands for pristine.

seems to witness again the intrinsic good properties of optical flow also independently from the used neural network.

4.4. Results training and testing on all the manipulations

In this sub-section, in order to understand if what it has been evidenced in the previous experiments still holds in general, we have verified the behavior of the proposed CNN classifier (based on *ResNet50)* when trained on all the manipulations. In Table 1, results are reported for the C23 and C40 cases; the accuracy obtained by the frame-based (RGB) method and by the optical-flow based one respectively are presented. It can be seen that performances are lower with respect to the previous situations reported in Section 4.3, when the models were specialized on a specific kind of forgery, but are still satisfactory, though the OF-based approach points out a reduced accuracy. What is very interesting is that when merging the two techniques (MIX), an improvement is achieved for both C23 and the more difficult C40 case. It can be clearly seen that such improvement is consistent and it seems that the frame-based method benefits, in some way, of the decision support provided by the optical flow-based one in a constructive manner. This appears to confirm what evidenced in the previous experimental tests: information contained in the OF features basically helps in the cross-forgery scenario and seems to complement the action of the frame-based method so yielding to an overall improvement in accuracy.

Table 1
Accuracy (%) when the net is trained on all the four manipulations (C23 and C40 cases); *P* stands for pristine.

C23	P	DF	F2F	FS	NT	TOT
RGB	93.07	95.71	97.72	97.26	97.04	95.00
OF	81.01	88.92	85.19	84.34	81.38	82.99
MIX	**93.93**	**97.35**	**98.41**	**97.40**	**97.14**	**95.75**
C40	P	DF	F2F	FS	NT	TOT
RGB	69.01	86.64	86.64	86.58	95.38	78.91
OF	63.74	70.76	65.76	60.47	70.73	63.57
MIX	**71.65**	**87.20**	**88.40**	**86.59**	**95.70**	**80.56**

5. Conclusions

In this work, optical flow field dissimilarities are used to discriminate between Deepfakes videos and original ones through the use of CNN. Experimental results, obtained on FaceForensics++ dataset, are very interesting and show that this kind of feature is suited to extract peculiar features between the fake and real cases, especially when working in the challenging *cross-forgery* scenario. Furthermore, it is also evidenced how this approach exploits inconsistencies on the temporal axis that combined with well-known state-of-the-art frame-based methodologies improve their performances. Such findings pave the way for many possible future works in order to evaluate the reliability of the proposed method by testing it against more reference datasets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work has been partially supported by the EC under European H2020 Programme, grant number no. 951911 - AI4Media. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, 2018, pp. 1–7. doi:10.1109/WIFS.2018.8630761.
[2] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, Protecting world leaders against deep fakes, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.
[3] L. Alparone, M. Barni, F. Bartolini, R. Caldelli, Regularization of optic flow estimates by means of weighted vector median filtering, IEEE Trans. Image Process. 8 (10) (1999) 1462–1467, doi:10.1109/83.791974.
[4] I. Amerini, L. Galteri, R. Caldelli, A. Del Bimbo, Deepfake video detection through optical flow based CNN, in: The IEEE International Conference on Computer Vision (ICCV) Workshops, 2019.
[5] S.S. Beauchemin, J.L. Barron, The computation of optical flow, ACM Comput. Surv. 27 (3) (1995) 433–466.
[6] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, J. Choo, StarGAN: Unified Generative Adversarial Networks for multi-domain image-to-image translation, CoRR abs/1711.09020 (2017).
[7] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, L. Verdoliva, ForensicTransfer: weakly-supervised domain adaptation for forgery detection, arXiv:1812.02510, 2018.
[8] H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5780–5789, doi:10.1109/CVPR42600.2020.00582.
[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[10] D. Feng, X. Lu, X. Lin, Deep detection for face manipulation, in: H. Yang, K. Pasupa, A.C.-S. Leung, J.T. Kwok, J.H. Chan, I. King (Eds.), Neural Information Processing, Springer International Publishing, Cham, 2020, pp. 316–323.
[11] D. Güera, E.J. Delp, Deepfake video detection using recurrent neural networks, in: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6, doi:10.1109/AVSS.2018.8639163.
[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[13] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of gans for improved quality, stability, and variation, CoRR abs/1710.10196 (2017).
[14] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396–4405.
[15] D.E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[16] Y. Li, M. Chang, S. Lyu, In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking, CoRR abs/1806.02877 (2018).
[17] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A large-scale challenging dataset for deepfake forensics, in: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020.
[18] F. Marra, C. Saltori, G. Boato, L. Verdoliva, Incremental learning for the detection and classification of GAN-generated images, in: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), 2019, pp. 1–6.
[19] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019, pp. 83–92, doi:10.1109/WACVW.2019.00020.
[20] H.H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311.
[21] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, FaceForensics++: Learning to Detect Manipulated Facial Images, in: The IEEE International Conference on Computer Vision (ICCV), IEEE, 2019.
[22] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, 2019.
[23] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 (2014).
[24] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: Image synthesis using neural textures, ACM Trans. Graph. 38 (4) (2019), doi:10.1145/3306346.3323035.
[25] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Niessner, Demo of Face2Face: real-time face capture and reenactment of RGB videos, in: ACM SIGGRAPH 2016 Emerging Technologies, SIGGRAPH '16, 2016, pp. 5:1–5:2.
[26] L. Verdoliva, Media forensics and deepfakes: an overview, 2020. iew. arXiv:2001.06564.
[27] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A.A. Efros, CNN-generated images are surprisingly easy to spot... for now, 2019. arXiv:1912.11035.
[28] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8261–8265, doi:10.1109/ICASSP.2019.8683164.
[29] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime TV-L1 optical flow, in: Joint Pattern Recognition Symposium, Springer, 2007, pp. 214–223.