

Deepfake Video Detection through Optical Flow based CNN

Irene Amerini¹, Leonardo Galteri², Roberto Caldelli¹, Alberto Del Bimbo¹
 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 Fake¹. 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

¹Deepfakes: [gi thub. https://gi thub.com/deepfakes/faceswap](https://github.com/deepfakes/faceswap).

