# Literature review

The research on image / video forensics has been rapidly developing over the last 15 years [Verdoliva, p.7]. In early years, traditional fake detection methods looked for in-camera (optical lens, color filter array, compression) and/or out-camera traces (traces of copy-move operations by editing software, etc) to decide if an image / video is forged. Verdoliva and Tolosana give comprehensive surveys of such and other methods [Tolosana survey], [Verdoliva].. But traditional methods are losing their relevance in times when social networks automatically modify original media content by compression and/or resize operation distorting possible in-camera and out-camera manipulation fingerprints. Machine learning methods with handcrafted features (double compression, noise residuals) proved quite good for fakes detection: 94% accuracy in 2013 challenge, but the accuracy drops dramatically if tested on unseen dataset [Verdoliva, p.7]. Moreover, low-resolution and high-compression significantly reduce performance of such methods.

In the recent years a spread of deepfakes (media content synthetically altered or generated by deep learning methods) posed new challenges and triggered a new wave of research on fake detection. The rise of deepfakes is associated with fast progress of Generative Adversarial Networks, a deep-learning technique introduced in 2014 by Ian Goodfellow, and their ability to generate high-quality face images. Face manipulations can include entire face synthesis, identity swap, expression swap or attribute manipulations. Tolosana et. al. in their survey [Tolosana survey] give up-to-date review of best-performing detectors for all of the mentioned face manipulations. Table 1 shows the best detectors from Tolosana survey for different kinds of facial manipulations:

**Table 1: The best performing face manipulation detectors from the Tolosana et.al. survey**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study** | **Methods** | **Classifiers** | **Best performance** | **Databases** |
| **Entire Face Synthesis** | | | | |
| Dang et al. (2019) | Deep Learning Features | CNN + Attention Mechanism | AUC = 100%  EER = 0.01% | DFFD (ProGAN, StyleGAN) |
| Neves et al. (2019) | Deep Learning Features | CNN | EER = 0.3%  EER = 4.5% | 100K-Faces (StyleGAN)  iFakeFaceDB |
| **Identity Swap** | | | | |
| Li et al. (2019) | Face Warping Features | CNN | AUC = 99.9%  AUC = 99.7% | DeepFakeTIMIT(LQ)  DeepFakeTIMIT(FQ) |
| R¨ossler et al. (2019) | Mesoscopic Features  Steganalysis Features  Deep Learning Features | CNN | Acc. = 98.0%  Acc. = 100.0%  Acc. = 97.0%  Acc. = 99.0% | FF++ (DeepFake, HQ)  FF++ (DeepFake, Raw)  FF++ (FaceSwap, HQ)  FF++ (FaceSwap, Raw) |
| Dolhansky et al. (2019) | Deep Learning Features | CNN | Precision = 93.0%  Recall = 8.4% | DFDC Preview |
| Sabir et al. (2019) | Image + Temporal Features | CNN + RNN | AUC = 96.9%  AUC = 96.3% | FF++ (DeepFake, LQ)  FF++ (FaceSwap, LQ) |
| Tolosana et al. (2020) | Facial Regions Features | CNN | AUC = 100%  AUC = 99.4%  AUC = 91.0%  AUC = 83.6% | UADFV  FF++ (FaceSwap, HQ)  DFDC Preview  Celeb-DF |
| **Attribute Manipulation** | | | | |
| Dang et al. (2019) | Deep Learning Features | CNN + Attention Mechanism | AUC = 99.9%  EER = 1.0% | DFFD (FaceApp/StarGAN) |
| **Identity Swap** | | | | |
| R¨ossler et al. (2019) | Mesoscopic Features  Steganalysis Features  Deep Learning Features | CNN | Acc. = 98.0%  Acc. = 100.0%  Acc. = 97.0%  Acc. = 99.0% | FF++ (Face2Face, HQ)  FF++ (Face2Face, Raw)  FF++ (NeuralTextures, HQ)  FF++ Neural Textures, Raw) |
| Dang et al. (2019) | Deep Learning Features | CNN + Attention Mechanism | AUC = 99.4%  EER = 3.4% | FF++ (Face2Face, -) |
| Sabir et al. (2019) | Image + Temporal Features | CNN + RNN | Acc. = 94.3% | FF++ (Face2Face, LQ) |

Source: Tolosana et.al. survey

It is clearly seen from the table that deep learning methods, most often CNN, show the best performance in detection of image/video forgery. As stated in Verdoliva’s survey, in the ideal conditions simple algorithms work well, but “in the presence of strong compression, there is a gap of about 15% between machine learning and deep learning, and 15% more using a very deep network” [Verdoliva, p.12]. Moreover, deepfakes are getting better very rapidly, there are little traces and fingerprints left to recognize a fake by a simple classifier.

However, classical forensics and ML methods can still be important. As stated in [Verdoliva], [Guarnera et.al.], deep learning networks is a black box, and it is difficult to explain why an image/video is a fake to a judge in a courtroom. Moreover, if some classical ML methods are computationally much more efficient compared to neural networks, so that less memory and time needed to do training. This is a reason to pay attention to classical ML classifiers and use these methods. Indeed, there are studies that show quite good results, when applying classical ML methods. Below we describe the most recent studies they focus on simple ML classifiers to detect a forged media content (images and video).

Matern et.al. use simple methods logistic regression, KNN and a simple neural network (NN) to find visual artifacts in fake images (no reflections in the eyes, different eyes color, imprecise face geometry) in generated faces (ProGAN and GLOW generated data), DeepFakes and Face2Face manipulations. The authors input face crops and face geometry statistics to detect generated faces and extract eyes, teeth, eyes&teeth, face borders, noses, and face crops as visual features for DeepFakes and Face2Face data. The best AUC of 0.852 is achieved by KNN classifier for the generated data, while logistic regression and NN show the highest AUC of 0.784 and 0.851 respectively for DeepFakes using eyes&teeth features and the AUC of 0.832 and 0.838 for Face2Face fakes using eyes features [Matern]. The results of Matern et. al are quite good, as they use simple methods that consume little time and efforts and can be used for prototyping. However, their method works only for images with open eyes and visible teeth, these prerequisites can be a limit for the efficiency and applicability of the method.

Dural et.al. in their study [Dural etl.al ] also use classical ML classifiers - SVM, KNN and logistic regression - to find fakes in the dataset based on different databases, including FaceForensics++. They apply Fourier Transform to a sample image and Azimuthal Average to receive 1D vector of 722 features. For high quality images the results are quite impressive - up to 100% for SVM and logistic regression and 96% for KNN. The experiments with low resolution images show that the accuracy drops to 90% and the results are sensitive to the sample size, especially for logistic regression classifier.

Bonettini et. al. offer a GAN image detector that is very accurate and computationally efficient. They use Benford’s law (First Digit law or Significant Digit law) to describe the distribution of the most significant digit for quantized Discrete Cosine Transform (DCT) coefficients of an image and use divergence coefficients as features for a random forest classifier. Their classifier shows excellent results (up to 100% accuracy on different GAN-generated datasets (not face images)) and compared to Xception does not need much data for training.

**Supplementary information (you do not have to read)**

Traditional forensics methods (Try to find fake manipulation fingerprints???)

Classical ML methods SVM

Carvalho et.al. 2013. They use physics and statistical illumination estimators, then extract edge and texture features and give them to SVM for classification. 86% accuracy. It’s not deepfakes, because deepfakes were introduced later

Matern 2020. The idea is very close to our article. They do simple logistic regression and small NN to detect fakes for Deepfakes and Face2Face manipulations (but not our database, as far as I understood, I should check). Very concise literature review, mentioning key traditional methods and deep learning. Good elaboration on artifacts - eyes reflection and geometry/edge distortions (i.e. teeth) - in fake videos/images. High AUC for both logistic and NN.

Dural 2020. They use Fourier Transform for feature extraction. They use FaceForensics++ dataset. They use Logistic Regression, SVM and KNN. Results are too good to be true.

Tolosana\_survey 2020. He mentions in-camera and out-camera fingerprints. Main drawback of traditional methods: they depend on specific training scenarios and not robust to the unseen conditions. (I am interested if these traditional methods use classical ML to classify the fingerprints???). Gives thorough review of different manipulation techniques. Lists key deepfakes images and video databases and key detection attempts with methods and results. The article mentions Maccloskey work 2018 that uses color differences between camera and deepfakes photos and SVM to detect fakes. Wang is another example of SVM here (in images): the input is artifacts at each layer of neuron modeling. The article mentions 5 other works, 4 of them are deep learning. Video detection: most interesting Matern (Logistic Regression, Visual Features), Yang (SVM, Head Pose Features), Agarwal and Farid (SVM, Head Pose and Facial Features). Others - CNN, RNN, Capsule Networks plus some other. The article contains separate table with detectors results for FF++, and Matern with his logistic regression is the only non-NN method.

Tolosana\_2020 They use FF++ and the same idea on train-test-split. This can be important. NN is not a black box (they visualize the result and eyes matter a lot). They show that 2nd generation fakes are better and more difficult to detect. As features they use facial regions (eyes, nose, mouth and rest of the face. All face is bette (except Capsule Network, where eyes produced slightly better results).

Guarnera. Describes in detail the detection methods that look at traces of image alteration/manipulation. In more detail they discuss Fourier transforms. . They state that its very important to study these traces rather than believe in the CNN black box.He uses classical forensics to analyse deepfakes and give the github link. I can look their for Fourier transform or other methods.

Marra. Use steganalysis and SVM. Good accuracy. But on compressed photos (like those in different social networks) Xception wins.

Bonettini. Their aim is to offer an excellent GAN photos fake detector with modest computational efforts. They use Banford’s law to extract features and a simple random forest classifier. They compare to Xception and SVM steganalyisis, but their dataset is specific and their feature-extraction and random forest wins.

Cazzolino. Steganalysis plus SVM. Quite an old article (not about GAN)

Yang. SVM and facial landmarks. The work of 2019.

Frank. Uses frequency spectrum and it improves all detector’s performance, especially the simplest ones (like KNN and eigenfaces).

Literature review: ideas and structure

Battle between deepfakes and detectors. 1st and 2nd generation deepfakes. NN are key players in the battle (NN proed better at traditional forensics? (Cazzolino?, Bonettini?, Mara?)) The results of NN of FF++ and 2nd generation deepfakes. SVM and steganalysis compete. But classical ML classifiers show good results (above all on 1st generation) and have their advantages: higher explanation power and computational simplicity. The ML classical methods have quite good results on FF++. Then describe Matern and Dural. Then Bonettini.

Battle between deepfakes and detectors. Traditional forensics and hand-crafted features. Then statistical features ?. Neural Nets. Neural nets beat the other. Why ML methods can be still important? (citation about judge from Verdoliva, smth about traces from Guarnera).

ML methods can work well on first generation deepfakes (Matern and Dural). What about second generation? Some examples. Bonettini?