Forensic Analysis of Tampered Digital Photos

Sara Ferreira[1(✉)], Mário Antunes[2,3], and Manuel E. Correia[1,3]

[1] Faculty of Science, University of Porto, Porto, Portugal
up201606726@up.pt, mcc@dcc.fc.up.pt
[2] CIIC, School of Technology and Management, Polytechnic of Leiria, Porto, Portugal
mario.antunes@ipleiria.pt
[3] INESC-TEC, CRACS, University of Porto, Porto, Portugal

Abstract. Deepfake in multimedia content is being increasingly used in a plethora of cybercrimes, namely those related to digital kidnap, and ransomware. Criminal investigation has been challenged in detecting manipulated multimedia material, by applying machine learning techniques to distinguish between fake and genuine photos and videos. This paper aims to present a Support Vector Machines (SVM) based method to detect tampered photos. The method was implemented in Python and integrated as a new module in the widely used digital forensics application Autopsy. The method processes a set of features resulting from the application of a Discrete Fourier Transform (DFT) in each photo. The experiments were made in a new and large dataset of classified photos containing both legitimate and manipulated photos, and composed of objects and faces. The results obtained were promising and reveal the appropriateness of using this method embedded in Autopsy, to help in criminal investigation activities and digital forensics.

Keywords: Digital forensics · Deepfake · Photo tampering · Support Vector Machines · Discrete Fourier Transform

1 Introduction

Cybercrime assumes different shapes. By having a computer connected to the Internet, cybercriminals are able to carry on a widespread of illegitimate and malicious activities against companies and individuals. In the last five years, there has been an increase of 67% in the incidence of security breaches worldwide [7], being malicious activities like phishing, ransomware, and crypto-jacking, some of the most popular threats to cybersecurity [14].

Intrinsically related, and in some way more silent, defacing and deepfake take advantage of multimedia contents manipulation techniques to modify digital photos and videos. In this type of crime, attackers are interested in defacing individuals' digital identity, by spreading malicious multimedia content and exposing

J. M. R. S. Tavares et al. (Eds.): CIARP 2021, LNCS 12702, pp. 461–470, 2021.
https://doi.org/10.1007/978-3-030-93420-0_43

individuals in an odd context. Broadly speaking, the motivations for deepfake are digital kidnap, usually involving under-aged and vulnerable victims [4].

Digital forensics analysis integrates techniques and procedures for the collection, preservation, and analysis of evidence in electronic equipment. It is an imperative tool for criminal investigation teams, namely in the analysis and identification of artifacts and digital evidence. Criminal investigation has recently encountered several challenges in detecting manipulated multimedia content, being even more affordable as cybercriminals massively use digital equipment connected to the Internet. Once seized, these equipment have to be analyzed to identify digital evidence and artifacts of suspicious activity.

Machine Learning (ML) has boosted the automated detection and classification of artifacts in a digital forensics investigation. Existing techniques to detect manipulated photos [1] are not yet properly integrated into forensic applications and therefore a module to automate this type of detection is relevant. The enhancements observed in the reported ML methods were not yet been translated into substantial improvements for cybercrime investigation, as those are not yet massively incorporated in state-of-the-art digital forensics tools.

Autopsy (https://www.autopsy.com/) is an open-source digital forensics application, widely used by criminal investigation police, dedicated to analyse and identify digital evidence and artifacts of suspicious and anomalous activities. It incorporates a wide set of native modules to process digital images (e.g. disks) and also allows the community to develop others more specific.

This paper describes the deployment and development of a module for Autopsy, that incorporates an SVM based method [6] to detect manipulated photos. The method was developed as a Python based module for Autopsy and is able to detect distinct anomalies in photos, like splicing and copy-move. The features were calculated by the Discrete Fourier Transform (DFT) and extracted for further processing by a SVM-based method. The module was tested with a classified dataset of about 40,000 photos, composed of both faces and objects, where it is possible to find examples of splicing and copy-move manipulations. The results proved the precision of the SVM-based method that achieved an averaged precision and recall of 99,4%, when detecting manipulated photos.

The remaining of the paper is organized as follows. Section 2 describes the fundamentals behind the topics covered in this paper, namely digital forensics, detection techniques, and deepfake. Section 3 depicts the overall architecture and pipeline delineated to process the input photos. Section 4 presents the experimental setup, the datasets, and the performance metrics used. Section 5 describes the results obtained. Finally, Sect. 6 describes the main conclusions and delineates the future work.

2 State of the Art

This Section describes the fundamentals of digital forensics, video manipulation techniques, and the most relevant ML techniques to deal with the detection of fake multimedia content.

2.1 Digital Forensics

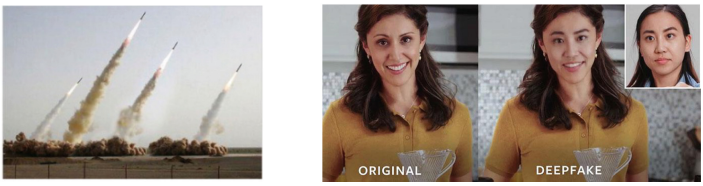
Digital forensics embodies techniques and procedures to collect, preserve and analyze digital evidence in electronic equipment, namely disks, smartphones, and other devices with storage capacity. The underpinning main goal is to produce a sustained reconstruction of events, that may help digital forensics investigators to build a list of evidence that may dictate about suspect's innocence or guilt.

Cybersecurity professionals understand the value of digital forensics information and the importance of maintaining it protected. For this reason, it is essential to establish strict guidelines and procedures, namely detailed instructions about authorized rights to retrieve digital evidence, how to properly prepare systems for evidence retrieval, where to store any recovered evidence, and how to document these activities to guarantee data authenticity and integrity. Among the digital forensics tools to extract, collect, and analyze digital artifacts, Autopsy, EnCase, FTK, XRY are the most relevant and widely used ones.

2.2 Multimedia Manipulation Techniques

Photos manipulation is appealing, mostly in the context of spreading fake news, defacing, deepfake, and digital kidnap. There are three main types of photo manipulation, that are described below.

Copy-move (Fig. 1(a)) consists of copying or moving part of a photo to another place in the same photo. The goal is to give the illusion of having more elements in the photo than those that are there.



(a) Copy-move (b) Splicing

Fig. 1. Photos manipulation types.

Splicing (Fig. 1(b)) consists of superimposing different regions of two photos, being deepfake the most relevant consequence. It is an artificial and automated manipulation of media, usually by means of artificial intelligence techniques, in which a person's face in an existing photo or video is swiped by someone else's face [2].

While deepfake of photos and videos is not new and can be seen in numerous digital contents, it has leveraged powerful ML and artificial intelligence techniques to improve contents manipulation. The most common ML methods used to improve deepfake are based on deep learning and involve training generative neural network architectures, such as auto encoders or Generative adversarial

networks (GANs) [3]. Deepfake has garnered widespread attention, as it has been used in digital campaigns of spreading fake news. This manipulation technique is also responsible for digital kidnap, revenge porn, and financial fraud [16].

Finally, Resampling consists of changing the scale or even the position of an element in a photo.

2.3 Techniques Used to Detect Photos Manipulation

Bearing in mind that the use of deepfake and copy-move in digital crimes is a growing problem and has a great impact on today's society, there are already documented algorithms to tackle with this type of manipulation.

The difference between Gaussian (DoG) and Oriented Rotated Brief (ORB) are techniques used to detect copy-move in manipulated photos. The method suggested by Niyishaka et al. [8] comprises three steps: corners detection with Sobel algorithm; features extraction with DoG and ORB; and finally, features correspondence.

Unmasking deepfake with DFT and ML is a method described in [6]. It is based on a classical frequency domain analysis with DFT, followed by a classification based on ML techniques. The frequency characteristics of a photo can be analyzed in a space defined by a Fourier transform, namely a spectral decomposition of the input data indicating how the signal's energy is distributed over a range of frequencies. In this method it is used a DFT, which is a mathematical technique to decompose a discrete signal into sinusoidal components of various frequencies ranging from 0 (constant frequency, corresponding to the image mean value) up to the maximum of the admissible frequency, given by the spatial resolution. The frequency-domain representation of a signal carries information about the signal's amplitude and phase at each frequency, and can be computed as (1):

$$X_{k,l} = \sum_{n=0}^{N-1} \sum_{0}^{M-1} x_{n,m} \cdot e^{(-\frac{i2\pi}{N} k_n)} \cdot e^{(-\frac{-i2\pi}{M} l_m)} \tag{1}$$

After applying a Fourier Transform to a photo, the returned values are represented in a new domain but within the same dimensionality. Therefore, given that we work with photos, the output still contains 2D information. We then apply azimuthal average to compute a robust 1D representation of the DFT power spectrum. At this point, each frequency component is the radial average from the 2D spectrum. Support Vector Machines (SVM) is then used to create a model based on a training dataset with manipulated and genuine photos. The model will then applied to a test dataset, to identify an optimal separating hyperplane that maximizes the margin between both classes.

Image splicing detection with artificial blurred boundary is a method described in [13]. It is based on image edge analysis and blur detection and has two steps: image edges analysis along with feature extraction and detection algorithm. For the first step, it is performed a Non-Subsampled Contourlet Transform (NSCT) in order to get more details of the high-frequency components with the size of all the directions the same as that of the original image. The

last step is to classify all features with SVM, similarly to the method described above. By using NSCT the authors claim to achieve a 95.12% true positive rate on detecting image splicing. The architecture developed in this paper, which is described in Sect. 3, applies the DFT method, having in mind the promising results previously obtained [6].

3 Architecture

This Section describes the architecture that was deployed to process input photos and to classify them as being genuine or manipulated. It also describes the Autopsy module developed to classify photos in a digital forensics context.

3.1 General Architecture

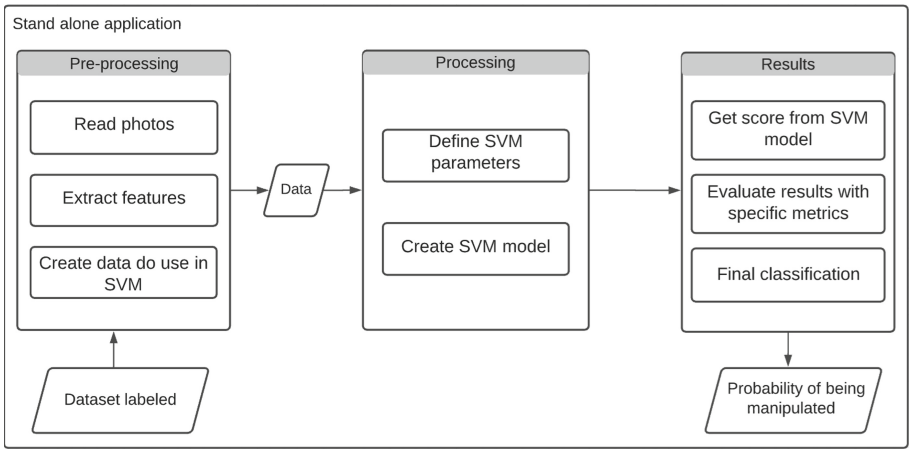


Fig. 2. Overall architecture.

The overall architecture of the standalone application developed to classify photos is depicted in Fig. 2. It has three main building blocks: pre-processing, processing, and results.

Pre-processing consists of extracting the features from the photos, by applying DFT method described in Sect. 2.3 [6] to produce the labeled input datasets for both training and testing. The pre-processing phase reads the photos through the `OpenCV` library and further extracts their features [6]. Using this method, exactly fifty features were obtained for each photo, that were then loaded into a new file with the corresponding label (0 for fake photos and 1 for the genuine ones). A training file with the features previously extracted, was created being then used to feed the SVM model.

The processing phase corresponds to the SVM processing. SVM is included in a set of kernel-based learning methods, in which the problem is addressed by mapping the data to a larger dimension space. This mapping may not be linear, as the function that allows this mapping is called a kernel [18]. The RBF (Radial basis function) kernel and a regularization parameter of 3 were chosen based on the experiments. The implementation of SVM processing was made through `scikit-learn` library for Python 3.9. The tests were carried on with a split of the whole dataset in two parts: 67% for training and 33% for testing. Both datasets (training and testing) are balanced regarding the amount of fake and genuine photos.

The results obtained in each processing are the following: SVM score; confusion matrix, precision, and recall; and the calculated prediction that allow us to deduce the probability of an image being manipulated.

3.2 Autopsy Module Architecture

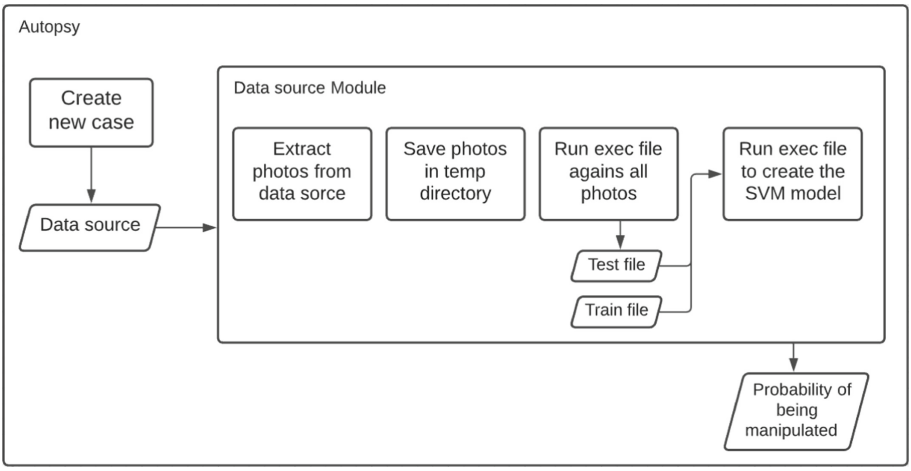


Fig. 3. Autopsy module architecture.

As stated before, Autopsy is among the most used digital forensics applications and is open to the integration of third-parties modules. Autopsy processes the input data and shows the user the results by means of report modules.

Autopsy uses Jython in new modules development, to enable Python scripting. Jython is converted into Java byte code and runs on the JVM. However, it is limited to Python 2.7. To overcome this limitation and the fact that some libraries used by the SVM classification method do not work in python 2.7, two Python executables were created to: one to process the photos; and another to create the SVM model and to classify the photos.

The data source ingest module, that runs against a data source added to the Autopsy, was developed and its architecture is depicted in Fig. 3. To start this analysis, it is needed to create a new "case" inside Autopsy and add one data source to it. An example of a data source is a disk image. Then, the module starts by extracting each photo within the data source added to the Autopsy case and saves them in a temporary directory. Next, the first Python executable extracts the features of the photos stored in the temporary directory, being all of them labeled with 0 (we assume that all photos are fake until proven otherwise). With the output obtained and the training file already created and distributed with the module, the second Python executable comes to action and the SVM classification model is created. Finally, the artifacts with the classification results are calculated and post in the Autopsy blackboard, which are further displayed to the user.

It is possible to note that the standalone application architecture corresponds also to the method used in Autopsy data source ingest module (Fig. 3). The standalone application was developed before the Autopsy module, which gave the possibility to develop and test the method while disregarding the needed compatibility with the Python libraries and with the strict format that is required by Autopsy to develop new modules.

4 Experimental Setup

This Section describes the dataset used for the experiments and the evaluation metrics that were applied to evaluate the SVM classification.

4.1 Datasets

A dataset containing both people's faces and objects was created to train the classification model. The dataset described in [6], which was used in the tests, is a compilation of photos available in `CelebA-HQ` dataset [9], `Flickr-Faces-HQ` dataset [10], `"100K Facesproject"` ("https://generated.photos/") and `"this person does not exist"` project (https://thispersondoesnotexist.com/).

Some complexity was added to the dataset, by including objects and other people's faces being possible to detect other types of manipulations aside deepfake. `COVERAGE dataset` [11] was included to add photos with objects and other people's faces, as well as a copy-move forgery database with similar but genuine objects that contains 97 legitimate photos and 97 manipulated ones. `Columbia Uncompressed Image Splicing Detection Evaluation Dataset` [12] was also added, which consists of high-resolution images, 183 authentic (taken using just one camera and not manipulated), and 180 spliced photos. Finally, 14 legitimate and 14 fake ad-hoc photos were also added, containing splicing and copy-move manipulations created by us. Putting it all together, the new dataset used in this paper is balanced and has 40,629 photos divided into two classes: 20,335 genuine (or real) photos and 20,294 that were manipulated.

4.2 Evaluation Metrics

To have a correct evaluation of the classification method, it is necessary to discuss the evaluation metrics. The metrics used to evaluate the results were Precision (P), Recall (R), and F1-score, which can be calculated through the well-known and documented confusion matrix [17].

In the confusion matrix, each row represents the instances in a predicted class, while each column represents the instances in an actual class. The positive class refers to the manipulated photos, while the negative class refers to the genuine ones. TP represents the events where the model has correctly predicted the positive class, that is a manipulated photo. TN calculates the events that were correctly predicted as negative, that is genuine photos. FP and FN evaluate the events that were incorrectly predicted by the model, namely those that legitimate photos classified as manipulated and those manipulated that were classified as genuine, respectively.

Precision and Recall correlate the metrics described above. Precision measures the percentage of examples identified as true that are really true. That is, those photos that are manipulated, from those that were classified as manipulated. Precision is calculated by (2):

$$P = \frac{TP}{(TP + FP)} \tag{2}$$

Recall is the percentage of manipulated images that we could find of the total number of manipulated images.Recall corresponds to the following (3):

$$R = \frac{TP}{(TP + FN)} \tag{3}$$

F1 is an harmonic mean between Precision and Recall. The range for the F1-score is between [0, 1] and measures the preciseness and robustness of the classifier. That is, the number of instances that were correctly classified and those that were misclassified, respectively. F1 measure is calculated by (4):

$$F1 = 2 * \frac{P * R}{(P + R)} \tag{4}$$

5 Results Analysis

This Section describes the results obtained from the experiments and the corresponding analysis. Ten experiments were made and, for each one, the dataset was randomly divided into 33% for testing and 67% for training. As can be seen in Table 1, the results obtained were very satisfactory, with a high number of correctly classified photos and a residual number of FP and FN.

It is possible to observe that we managed to achieve in these tests a precision, recall, and F1-score of approximately 100%. Comparing with the results documented in [6], even enriching the dataset with photos containing objects and other types of manipulation, it was possible to achieve the mean P, R, and F1

Table 1. Results obtained with 10 different runs.

	TP	TN	FP	FN	Precision	Recall	F1-score
Run 1	6629	6646	14	43	0.99789	0.99355	0.99571
Run 2	6580	6698	19	35	0.99712	0.99470	0.99591
Run 3	6636	6633	23	40	0.99654	0.99400	0.99527
Run 4	6644	6618	25	45	0.99625	0.99327	0.99476
Run 5	6621	6648	15	48	0.99774	0.99280	0.99526
Run 6	6713	6554	14	51	0.99792	0.99246	0.99518
Run 7	6674	6608	21	29	0.99686	0.99567	0.99627
Run 8	6584	6683	21	44	0.99682	0.99336	0.99509
Run 9	6600	6680	10	42	0.99849	0.99368	0.99608
Run 10	6640	6635	19	38	0.99715	0.99431	0.99573
Average	6632	6640	18	41	0.99728	0.99378	0.99553

above 99.3% (P = 99.73%, R = 99.38% and F1 = 99.55%). The mean values for FP and FN are residual (FP = 0.13%, FN = 0.3%) and, by analysing the misclassified photos, it is possible to infer that were related to the resolution of the photos. A richer dataset with heterogeneous examples regarding the resolution of the photos would benefit the overall results obtained.

6 Conclusion

This paper described the development of an SVM based method [6] to tackle the detection of manipulated photos. An Autopsy module that incorporates the proposed standalone SVM-based method, was also developed, giving a helping hand to digital forensics investigators and leveraging the use of ML techniques to fight cybercrime activities.

The overall architecture and development make use of two well-known and documented techniques: Discrete Fourier Transform (DFT) technique to extract features from photos; SVM-based method to create a learning model. Both techniques were incorporated in the proposed SVM-based learning standalone application, which was further integrated as an Autopsy module in a digital forensics context. The dataset proposed in [6] was extended with different sources, mainly to accommodate objects and other manipulation types, besides faces and splicing respectively. The final dataset has about 40,000 photos, composed of both faces and objects, where it is possible to find examples of splicing and copy-move manipulations. The results obtained were promising and in line with previous ones documented in the literature. It was possible to achieve a mean F1-score of 99.55% on the detection of manipulated photos.

Future work has two major topics: to enhance the dataset present in this paper, by adding more genuine and manipulated photos; to apply the presented methodology and architecture to detect these types of manipulations in videos.

References

1. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. arXiv preprint arXiv:2001.00179 (2020)
2. Kietzmann, J., Lee, L., McCarthy, I., Kietzmann, T.: Deepfakes: trick or treat? Bus. Horiz. **63**(2), 135–146 (2020)
3. Nguyen, T., Nguyen, C., Nguyen, D.T., Nguyen, D.T., Nahavandi, S.: Deep learning for deepfakes creation and detection 1 (111573) (2019)
4. Spivak, R.: Deepfakes": the newest way to commit one of the oldest crimes. Georget. Law Technol. Rev. **3**(2), 339–400 (2019)
5. Harris, D.: Deepfakes: false pornography is here and the law cannot protect you. Duke Law Technol. Rev. **17**, 99 (2018)
6. Durall, R., Keuper, M., Pfreundt, F.J., Keuper, J.: Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686 (2019)
7. Bissell, K., LaSalle, R.M., Dal Cin, P.: The cost of cybercrime-Ninth annual cost of cybercrime study. Technical report, Accenture, 2019. Independently conducted by Ponemon Institute LLC and jointly developed by Accenture (2019). https://www.accenture.com/_acnmedia/PDF-96/Accenture-2019-Cost-of-Cybercrime-Study-Final.pdf. Accessed 9 Jan 2021
8. Niyishaka, P., Bhagvati, C.: Digital image forensics technique for copy-move forgery detection using DoG and ORB. In: Chmielewski, L.J., Kozera, R., Orłowski, A., Wojciechowski, K., Bruckstein, A.M., Petkov, N. (eds.) ICCVG 2018. LNCS, vol. 11114, pp. 472–483. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00692-1_41
9. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
10. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
11. Wen, B., Zhu, Y., Subramanian, R., Ng, T.T., Shen, X., Winkler, S.: COVERAGE - a novel database for copy-move forgery detection. In: IEEE International Conference on Image processing (ICIP), pp. 161–165 (2016)
12. Hsu, Y.F., Chang, S.F.: Detecting image splicing using geometry invariants and camera characteristics consistency. In: 2006 IEEE International Conference on Multimedia and Expo, pp. 549–552 (2006)
13. Liu, G., Wang, J., Lian, S., Dai, Y.: Detect image splicing with artificial blurred boundary. Math. Comput. Model. **57**(11–12), 2647–2659 (2013)
14. Moore, M.: Top Cybersecurity Threats in 2020. University of Sandiego. https://onlinedegrees.sandiego.edu/top-cyber-security-threats/. Accessed 12 Jan 2021
15. Christian, J.: Experts fear face swapping tech could start an international showdown. The Outline. https://theoutline.com/post/3179/deepfake-videos-are-freaking-experts-out?zd=1&zi=hchawpks. Accessed 13 Jan 2021
16. Roose, K.: Here Come the Fake Videos, Too. The New York Times. https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html. Accessed 13 Jan 2021
17. Shung, K.P.: Accuracy, Precision, Recall or F1?. Towards Data Science.https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9. Accessed 6 Jan 2021
18. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intell. Syst. Appl. **13**(4), 18–28 (1998)