Efficient video integrity analysis through container characterization

Pengpeng Yang, Student Member, IEEE, Daniele Baracchi, Massimo Iuliani, Dasara Shullani, Rongrong Ni, Yao Zhao, Senior Member, IEEE, and Alessandro Piva, Fellow, IEEE

Abstract—Most video forensic techniques look for traces within the data stream that are, however, mostly ineffective when dealing with strongly compressed or low resolution videos. Recent research highlighted that useful forensic traces are also left in the video container structure, thus offering the opportunity to understand the life-cycle of a video file without looking at the media stream itself.

In this paper we introduce a container-based method to identify the software used to perform a video manipulation and, in most cases, the operating system of the source device. As opposed to the state of the art, the proposed method is both efficient and effective and can also provide a simple explanation for its decisions. This is achieved by using a decision-tree-based classifier applied to a vectorial representation of the video container structure. We conducted an extensive validation on a dataset of 7000 video files including both software manipulated contents (ffmpeg, Exiftool, Adobe Premiere, Avidemux, and Kdenlive), and videos exchanged through social media platforms (Facebook, TikTok, Weibo and YouTube). This dataset has been made available to the research community. The proposed method achieves an accuracy of 97.6% in distinguishing pristine from tampered videos and classifying the editing software, even when the video is cut without re-encoding or when it is downscaled to the size of a thumbnail. Furthermore, it is capable of correctly identifying the operating system of the source device for most of the tampered videos.

Index Terms—video forensics, video container, social media, integrity, authentication, video tampering, decision trees, machine learning.

This work was supported in part by the National Key Research and Development of China (No. 2016YFB0800404), the National NSF of China (Nos. 61672090, 61532005, U1936212) and the Fundamental Research Funds for the Central Universities (Nos. 2018JBZ001, 2017YJS054). It was also supported in part by the Air Force Research Laboratory and in part by the Defense Advanced Research Projects Agency under Grant FA8750-16-2-0188. Finally it was supported by the Italian Ministry of Education, Universities and Research MIUR under Grant 2017Z595XS. P. Yang, R. Ni, and Y. Zhao are with the Beijing Key Laboratory of Advanced Information Science and Network Technology and the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China.

D. Baracchi, M. Iuliani, D. Shullani, and A. Piva are with the Department of Information Engineering, University of Florence, Via di S. Marta, 3, 50139 Florence, Italy.

M. Iuliani, and A. Piva are with the FORLAB, Multimedia Forensics Laboratory, PIN Scr, Piazza G. Ciardi, 25, 59100 Prato, Italy.

Please address correspondence to Rongrong Ni (email: rnm@bjtu.edu.cn) and Alessandro Piva (email: alessandro.piva@unifi.it).

Digital Object Identifier 10.1109/JSTSP.2020.3008088

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

I. INTRODUCTION

Video Forensics develops algorithms for assessing video integrity and authenticity by looking at the digital traces left during the video life-cycle [1]. Most of the existing video forensic techniques verify the authenticity of a video file by investigating the presence of inconsistencies in pixel statistics. For example, double encoding or manipulation can be detected by analyzing prediction residuals [2] or macroblock types [3, 4, 5]. Similarly, traces of frame rate up-conversion can be used to prove malicious video processing [6, 7]. Recent works have also successfully employed deep-learning techniques to detect video forgeries [8]. Jamimamul et al. [9] focused on in-frame video forgery detection by designing a 3D Convolutional neural network. Verde et al. [10] introduced a CNN-based approach to detect and localize splicing manipulation by learning video codec traces.

A major drawback of most of those techniques is their high computational cost; furthermore, strong compressions and downsampling often hide forensic traces, thus severely restricting the number of scenarios where those methods can be employed.

Recently, a new research branch highlighted that video integrity can be determined using information hidden in the whole video file and not just in the video stream [12]. Video files, in fact, are written to disk using a specific structure called container, comprising multiple streams (video, audio, subtitles) and metadata, which are exploited by decoding software to correctly reproduce the video. Guera et al. [13] showed how to identify forged videos without looking at the pixel space. To

1. https://www.oberlo.com/blog/video-marketing-statistics accessed on March 2020.

2. Note that integrity and authenticity are different concepts. Integrity is proved when the imagery is complete and unaltered, from the time of acquisition or generation through the life of the imagery; indeed, content authentication is used to determine whether the visual content depicted in imagery is a true and accurate representation of subjects and events. More details can be found on the Best Practices for Image Authentication of the Scientific Working Group on Digital Evidences [11].
do so, they extracted high level features (multimedia stream descriptors) related to video coding using ffmpeg. However, the tree-shaped container structure is not taken into account, thus discarding most of the overall available information. Iuliani et al. [14] highlighted that video integrity can be assessed by looking at the video file container structure, since any post-processing operation alter the content and the position of some atoms and field-values. This approach turned out to be also promising in classifying the source brand of native videos.

However, Iuliani et al. [14] merely detects a loss of integrity, without providing a human-interpretable explanation of the reasoning behind its decisions. Furthermore, the method has a linear computational cost since it requires to check the dissimilarity of the probe video with all available reference containers. As a consequence, an increase of the reference dataset size leads to a higher computational effort. Furthermore, both [13] [14] do not provide any characterization of manipulated videos, nor any explainability of the achieved outcome.

In this paper we introduce an efficient method for the analysis of video file containers that allows both to characterize identified manipulations and to provide an explanation for the outcome. The proposed approach is based on Decision Trees [15], a non-parametric learning method used for classification problems in many signal processing fields. Their key feature is the ability to break down a complex decision-making process into a collection of simpler decisions. We enriched the tool with a likelihood ratio framework designed to automatically clean up the container elements that only contribute to source intra-variability.

With respect to the state of the art, the proposed method, simply called EVA from now on, offers new forensic opportunities, such as: identifying the manipulating software (e.g. Adobe Premiere, ffmpeg, . . . ); providing additional information related to the original content history, such as the source device operating system.

The process is extremely efficient since a decision can be taken by checking the presence of a small number of features, independently on the video length or size. Furthermore, EVA can provide a simple explanation for the process leading to an outcome, since container symbols used to take a decision can be inspected. To the best of our knowledge, this is the first video forensic method with all these desirable traits. Experiments have been performed using videos produced by 34 modern smartphones from some of the most popular brands on the market, e.g. Apple, Samsung, LG, Huawei. Tampered contents were generated using both automated processing and manual user operations. This approach allowed us to build a sizeable, realistic dataset. Manipulations include contents generated using Exiftool, ffmpeg, Adobe Premiere, Avidemux and Kdenlive.

Eventually, we investigated whether a container-based approach is effective when dealing with videos exchanged through Facebook, TikTok, Weibo, and YouTube. Overall, the experimental validation involved seven thousands videos. This paper is organised as follows: Section II describes the video container standard; Section III introduces the mathematical tools to represent and analyse the video file container; Sections IV and V are devoted to the experimental validation of the proposed techniques; finally, Section VI draws some final remarks and outlines future works.

II. VIDEO FILE FORMAT

Most smartphones and compact cameras output videos in mp4, mov, or 3gp format. These video packaging refer to the same standard, ISO/IEC 14496 Part 12 [16], that defines the main features of MP4 [17] and MOV [18] containers while leaving a wide margin for those who implement it. In Fig. 1 we provide an example of an MP4-like container, a tree-like structure describing the video file with respect to three aspects: how the bytes are organized (physical aspect); how the audio/video streams are synchronized (temporal aspect); and how the latter two aspects are linked (logical aspect).

Each node (atom) is identified by a unique 4-byte code. It consists of a header which describes its role in the container and possibly some associated data. The first atom to appear in a container has to be ftyp since it defines the best usage and compatibility of the video content. The video structural information is separated from the data itself, indeed the first

![Fig. 1. Pictorial representation of an MP4-like video container structure.](http://www.ftys.com/)

[3]The reader can refer to [http://www.ftys.com/](http://www.ftys.com/) for further details.
one is stored in the movie atom (moov) and the second one in the media data atom (mdat). The moov atom links the logical and timing relationships of the video-samples, and provides pointers to their mdat location. It is worth noting that the moov sub-tree can contain one or moretrak atoms, depending on the number of streams present in a video (i.e. visual-stream and/or audio-stream).

### III. Proposed Approach

We can represent a video container as a labelled tree where internal nodes and leaves correspond to, respectively, atoms and field-value attributes. A video container X can be characterised by the set of symbols \( \{s_1, \ldots, s_m\} \), where \( s_i \) can be: (i) the path from the root to any field (value excluded), also called field-symbols; (ii) the path from the root to any field-value (value included), also called value-symbols. An example of this representation can be:

\[
\begin{align*}
s_1 &= \text{[ftyp/@majorBrand]} \\
s_2 &= \text{[ftyp/@majorBrand/isom]} \\
\ldots \\
s_i &= \text{[moov/mvhd/@duration]} \\
s_{i+1} &= \text{[moov/mvhd/@duration/73432]} \\
\ldots
\end{align*}
\]

Overall, we denote with \( \Omega \) the set of all unique symbols \( s_1, \ldots, s_M \) available in the world set of digital video containers \( \mathcal{X} = \{X_1, \ldots, X_N\} \). Similarly, \( C = \{C_1, \ldots, C_s\} \) denotes a set of possible origins (e.g., Huawei P9, Apple iPhone 6s). Given a container \( X \), the different structure of its symbols \( \{s_1, \ldots, s_m\} \) can be exploited to assign the video to a specific class \( C_u \).

For this purpose binary decision trees \([19]\) are employed to build a set of hierarchical decisions. In each internal tree node the input data is tested against a specific condition; the test outcome is used to select a child as the next step in the decision process. Leaf nodes represent decisions taken by the algorithm. An example is reported in Fig. 3. More specifically, in our approach we adopted the growing-pruning-based Classification And Regression Trees (CART) \([20]\).

Given the size of unique symbols \( |\Omega| = M \), a video container \( X \) is converted into a vector of integers \( X \mapsto (x_1\ldots x_M) \) where \( x_i \) is the number of times that \( s_i \) occurs into \( X \). This approach is inspired by the bag-of-words representation \([21]\) used to reduce variable-length documents to a fixed-length vectorial representation.

Note that \( X \) contains several symbols that are not representative of any class, thus contributing to class intra-variability only (e.g. information related to video length, acquisition date and time). This information is expected to introduce noise in the decision process and it should be possibly removed. Thus, we pre-filtered the data in \( \Omega \) by using the likelihood ratio framework. Given two classes \( C_u, C_v, u \neq v \), and a symbol \( s_i \), the log-likelihood ratio (LLR)

\[
\log L_{u,v}(s_i) = \log \frac{P(s_i|C_u)}{P(s_i|C_v)} \tag{1}
\]

\(^4\)Note that \( \Theta \) is used to identify atom parameters and root is used for visualization purpose but it is not part of the container data.

\(^5\)We avoid null frequencies by adding one to both the numerator and the denominator.

![Fig. 2. LLRs of symbols obtained when comparing native videos with ones altered through Exiftool. Values far from 0 are automatically included in the analysis. Values close to 0 in all compared classes are excluded from the analysis. Note that @ is used to identify a tree leaf.](image-url)
Then, we created 1260 for each device, thus obtaining a total of 140 download of all the videos in VISION, we selected 4 been feasible to perform the editing operations, upload, and smartphones of licly available dataset that includes native videos from collection of videos, starting from VISION [22], a pubrine and tampered videos. To answer that we created a new videos with modified dates through Exiftool

![Decision tree applied to distinguish iOS native videos from iOS smartphones](https://lesc.dinfo.unifi.it/en/datasets)

IV. INTEGRITY VERIFICATION

The first relevant experimental question is whether the proposed approach is capable of distinguishing between pristine and tampered videos. To answer that we created a new collection of videos, starting from VISION [22], a publiccly available dataset that includes native videos from 35 smartphones of 10 different brands. As it would have not been feasible to perform the editing operations, upload, and download of all the videos in VISION, we selected 4 videos for each device, thus obtaining a total of 140 pristine videos. Then, we created 1260 (140 × 9 editing operations) tampered videos, both automatically generated with ffmpeg and Exiftool, and manually created through Kdenlive, Avidemux and Adobe Premiere. More specifically:

- **cut with re-encoding**: each video was cut through ffmpeg and re-encoded
- **cut without re-encoding**: each video was cut through ffmpeg by copying the audio and video coding parameters to minimize the traces left by the operation;
- **speed up**: each video was speeded up through ffmpeg;
- **slow down**: each video was cut through ffmpeg and slowed down;
- **cut + downscale**: each video was cut through ffmpeg and downscaled to the resolution of 320 × 240;
- **cut-kd**: each video was manually cut through Kdenlive (v17.12.3) by keeping 5 to 7 seconds and then the video was saved with the MP4 - the dominating format(H264/AAC) setting;
- **cut-av**: each video was manually cut through Avidemux (v2.7.4) by keeping 5 to 7 seconds and then the video was saved as copy and MP4 Muxer settings;
- **cut-ap**: each video was manually cut through Adobe Premiere Pro CC 2019 by keeping 5 to 7 seconds and by saving as H.264 with medium bitrate setting;
- **date change**: each video was manually processed through Exiftool (v11.37) to change the date information within

The operation is performed with ffmpeg version 3.4.6 through the command

```bash
ffmpeg -i $file -ss 00:00:03 -t 00:00:05 -vcodec libx264 -acodec copy $name
```

The operation is performed with the command ffmpeg -i $file

```bash
-exiftool

```
Table I: Balanced accuracies obtained in the basic scenario for each device.

| Device | Balanced Accuracy | Device | Balanced Accuracy | Device | Balanced Accuracy |
|--------|-------------------|--------|-------------------|--------|-------------------|
| D01    | 1.00              | D13    | 1.00              | D26    | 1.00              |
| D02    | 1.00              | D14    | 1.00              | D27    | 1.00              |
| D03    | 1.00              | D15    | 1.00              | D28    | 1.00              |
| D04    | 0.99              | D16    | 1.00              | D29    | 1.00              |
| D05    | 1.00              | D18    | 1.00              | D30    | 1.00              |
| D06    | 1.00              | D19    | 1.00              | D31    | 1.00              |
| D07    | 1.00              | D20    | 1.00              | D32    | 1.00              |
| D08    | 1.00              | D21    | 1.00              | D33    | 1.00              |
| D09    | 1.00              | D22    | 1.00              | D34    | 1.00              |
| D10    | 1.00              | D23    | 1.00              | D35    | 1.00              |
| D11    | 1.00              | D24    | 1.00              |        |                   |
| D12    | 0.50              | D25    | 1.00              |        |                   |

Table II: Comparison of our method with the state of the art. Values of accuracy and time are averaged over the 34 folds.

| Method            | Balanced Accuracy | Training time | Test time |
|-------------------|-------------------|---------------|-----------|
| Guera et al. [13] | 0.67              | 347 s         | < 1 s     |
| Iuliani et al. [14] | 0.85              | N/A           | 8 s       |
| EVA               | 0.98              | 51 s          | < 1 s     |

It should also be noted that VISION is composed by several iOS/Android devices and a single Windows phone. We removed this latter device (D17) from our tests since it is not a representative sample for Windows Phone devices. For this reason, our approach aims to distinguish between iOS and Android videos only. This is a negligible limitation given that Windows Phone devices represent less than 0.3% of the mobile devices market [24].

In order to estimate the real-world performance of the proposed method we adopted an exhaustive leave-one-out cross-validation strategy. We partitioned our dataset in 34 subsets, each one of them containing pristine, manipulated, and social-exchanged videos belonging to a specific device. We performed each of the experiments hereby described 34 times, each time keeping one of the subsets out as test set, and using the remaining 33 for training our model. In this way, test accuracies collected after each iteration are computed on videos belonging to an unseen device. We reported the mean accuracies obtained among all the iterations as confusion matrices. During the training we assigned to each class a weight inversely proportional to the class frequency. We used the decision trees algorithms included in scikit-learn [23], a freely available Python toolkit for machine learning.

We trained our method to distinguish between the two classes “Pristine” (containing 136 videos) and “Tampered” (containing 1224 videos). We obtained a global balanced accuracy of 98.5%, failing only for videos produced by D12 (see Table I). The low accuracy obtained on such a device is reasonably due to the fact that it is the sole Sony smartphone in our dataset.

As a consequence of our strict leave-one-device-out strategy, we have no videos belonging to a Sony device in our training set when D12 is tested. Thus, our algorithm cannot learn the features needed to correctly classify those videos. This limitation does not always apply as different camera models can exhibit very similar containers. In such a case, a native video can be correctly classified even if the specific originating device is unavailable in the training set. This is the case of the LG D290 (D04) that reaches an accuracy of 0.99.

We also compared our method with two recently proposed algorithms for video integrity [13] [14]. In Table II we report the mean global accuracy and the average runtime per fold for the proposed approach and for those two methods.

A. Discussion

EVA provides several improvements with respect to the state of the art. In comparison with Guera et al. [13] we achieve a higher accuracy. This can be reasonably attributed to their use of a smaller feature space; indeed, only a subset of the available pieces of information are extracted without considering their position within the video container. On the contrary, EVA features also include the path from the root to the value, thus providing a stronger discriminating power. Indeed, this approach allows to distinguish between two videos where the same information is stored in different atoms. When compared with Iuliani et al. [14], EVA is capable of obtaining better classification performance with a lower computational cost. In Iuliani et al. [14] O(N) comparisons are required since all the N reference-set examples must be compared with a tested video; on the contrary, the cost for a decision tree analysis is O(1) since the output is reached in a constant number of steps.

Furthermore, EVA allows a simple explanation for the outcome. For the sake of example, we report in Fig. 4(a) a sample tree from the integrity verification experiment: the decision is taken by up to four checks, just based on the presence of the symbols ftyp/@minorVersion = 0, uuid/@userType, moov/udta/XMP_/moov/udta/auth. We also report in Fig. 4(b) a tree from the blind scenario experiment: in this case the tree needs to check the absence of just one atom to classify a YouTube video; at the same time a series of more complex checks are used to assign a video to other classes.

This shows how a single decision tree can handle both easy- and hard-to-classify cases at the same time. Neither [14] nor [13] provide an equivalent feature. Moreover, EVA is equipped with a formal likelihood ratio framework that can estimate the relevance of symbols for specific tasks. This framework has been used to automatically remove symbols that only contribute to class intra-variability.

V. Manipulation Characterization

We also performed a set of experiments designed to show that the proposed method, as opposed to the state of the art, is also capable of identifying the manipulating software and the operating system of the originating device. More specifically, we tried to answer the following questions:

A. Software identification: Is the proposed method capable of identifying the software used to manipulate a video? If yes, is it possible to identify the operating system of the original video?
### B. Integrity Verification on Social Media

Given a video from a social media platform (YouTube, Facebook, TikTok or WeiBo), can we determine whether the original video was pristine or tampered?

### C. Blind scenario

Given a video that may or may not have been exchanged through a social media platform, is it possible to retrieve some information on the video origin?

#### A. Software identification

In this scenario we only analyze videos that either are native, or that have undergone a manipulation. This time, however, we trained our algorithm to classify which software has been used to tamper the video, if any. Our classes are thus: “native” (136 videos), “Avidemux” (136 videos), “Exiftool” (136 videos), “ffmpeg” (680 videos), “Kdenlive” (136 videos), and “Premiere” (136 videos).

In this experiment EVA obtained a global balanced accuracy of 97.6%; the detailed results reported in Table III show that the algorithm achieved a slightly lower accuracy in identifying ffmpeg with respect to the other tools. This is reasonably due to the fact that ffmpeg library is used by other software and, internally, by Android devices.

We also trained our algorithm to classify both the editing software used to tamper the video, if any, and the operating system of the device originally used for the acquisition. The classes for this scenario are: “Android-native” (84 videos), “iOS-native” (52 videos), “Android-avidemux” (84 videos), “iOS-avidemux” (52 videos), “Android-exiftool” (84 videos), “iOS-exiftool” (52 videos), “Android-ffmpeg” (420 videos), “iOS-ffmpeg” (260 videos), “Android-kdenlive” (84 videos), “iOS-kdenlive” (52 videos), “Android-premiere” (84 videos), and “iOS-premiere” (52 videos).

A summary of the results obtained by this experiment is reported in Table IV. Our approach maintains good performance in correctly identifying the editing software. We notice, however, that the operating system used for videos manipulated with Kdenlive or with Adobe Premiere is often misclassified. At the same time, both those programs are always identified correctly. This indicates that the container’s structure of videos saved by Kdenlive and Adobe Premiere is probably reconstructed in a software-specific way.

#### B. Integrity Verification on Social Media

In this scenario we tested YouTube, Facebook, TikTok and Weibo videos to determine whether they were pristine or manipulated prior the upload.

A summary of the results obtained by our method is reported in Table V. We achieved global balanced accuracies of 0.76, 0.80, 0.79, and 0.60 on Facebook, TikTok, Weibo, and Youtube, respectively. Such results are characterised by low true negative rates, and thus it cannot be considered effective in this scenario, as many tampered videos are incorrectly classified as pristine.

The poor performance are mainly due to the social media transcoding process that flattens the containers almost independently on the video origin. As an example, after YouTube transcoding, videos produced by Avidemux and by Exiftool have exactly the same container representation. We do not know how the videos are processed by the considered platforms due to the lack of public documentation but we can assume that uploaded videos undergo custom/multiple processing. Indeed, social media videos need to be viewable on a great range of platforms, and thus need to be transcoded to multiple video codecs and adapted for multiple resolutions and bitrates. Thus, it seems plausible that those operations could discard most of the original container structure.

#### C. Blind scenario

In this scenario we considered videos that may or may not have been exchanged through a social media platform and we would like to extract the most complete information possible. We used all the videos in our dataset and we trained our
In this paper we proposed an efficient forensic method for checking video integrity. If a manipulation is detected, the proposed method allows to identify the editing software and, in most cases, whether the original video belonged to an Android or iOS device.

This is achieved by exploiting a decision tree classifier applied to a vector based representation of the video container structure, enriched with the likelihood ratio framework that is employed to automatically remove container elements that only contribute to source intra-variability. The proposed method, in case of tampered videos, is able to characterise the software that performed the manipulation with an accuracy of 97.6%, even when the video is cut without re-encoding. Except for manipulations performed with Adobe Premiere and Kdenlive, the proposed method correctly determines the operating system of the video source device.

As opposed to the state of the art, the proposed method is extremely efficient and can provide a simple explanation for its decisions. A new experimental dataset of 7000 videos was also created and shared with the research community, including contents generated with five editing tools (ffmpeg, Exiftool, Adobe Premiere, Avidemux, and Kdenlive) and four social media platforms (Facebook, TikTok, Weibo and Youtube). The current limitation of the method is that a container-based approach can identify whether the video belongs to a social medial platform like YouTube, Facebook, TikTok or Weibo, but it cannot be effectively applied on such contents for authenticity assessment, since the transcoding operation wipes out most of the forensic traces from the video container. Future works will aim to improve our tool by adding handcrafted features to improve the performance on social media contents.

VI. Conclusions

In this paper we proposed an efficient forensic method for checking video integrity. If a manipulation is detected, the proposed method allows to identify the editing software and, in most cases, whether the original video belonged to an Android or iOS device.

This by exploiting a decision tree classifier applied to a vector based representation of the video container structure, enriched with the likelihood ratio framework that is employed to automatically remove container elements that only contribute to source intra-variability. The proposed method, in case of tampered videos, is able to characterise the software that performed the manipulation with an accuracy of 97.6%, even when the video is cut without re-encoding. Except for manipulations performed with Adobe Premiere and Kdenlive, the proposed method correctly determines the operating system of the video source device.

### TABLE IV

|                | Native   | Avidemux | Exiftool | ffmpeg | Kdenlive | Premiere |
|----------------|----------|----------|----------|--------|----------|----------|
| Android        | 0.95     | -        | -        | -      | -        | -        |
| iOS            | -        | 1.00     | -        | -      | -        | -        |
| Avidemux       | -        | -        | 0.95     | 0.05   | -        | -        |
| Android        | -        | -        | -        | 1.00   | -        | -        |
| iOS            | 0.01     | -        | -        | -      | -        | -        |
| Exiftool       | -        | 0.01     | 0.05     | 0.99   | -        | -        |
| Android        | -        | -        | -        | -      | 0.75     | 0.15     |
| iOS            | -        | -        | -        | -      | -        | 0.04     |
| ffmpeg         | 0.01     | -        | -        | -      | -        | -        |
| Android        | -        | 0.01     | 0.05     | -      | -        | -        |
| iOS            | -        | -        | -        | -      | -        | 1.00     |
| Kdenlive       | -        | -        | -        | -      | -        | -        |
| Android        | -        | -        | -        | -      | -        | 0.75     |
| iOS            | -        | -        | -        | -      | -        | 0.25     |
| Premiere       | -        | -        | -        | -      | -        | -        |
| Android        | -        | -        | -        | -      | -        | 0.38     |
| iOS            | -        | -        | -        | -      | -        | 0.62     |
|                | -        | -        | -        | -      | -        | 0.79     |
|                | -        | -        | -        | -      | -        | 0.21     |
|                | -        | -        | -        | -      | -        | 0.37     |
|                | -        | -        | -        | -      | -        | 0.63     |

### TABLE V

Performance achieved for integrity verification on social media contents. We report for each social network the obtained accuracy (ACC), true positive rate (TPR), and true negative rate (TNR). All these performance measures are balanced.

| Social Network | Accuracy | TNR   | TPR   |
|----------------|----------|-------|-------|
| Facebook       | 0.76     | 0.40  | 0.86  |
| TikTok         | 0.80     | 0.51  | 0.75  |
| Weibo          | 0.79     | 0.45  | 0.82  |
| YouTube        | 0.60     | 0.36  | 0.74  |

ACKNOWLEDGMENT

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as having to do with the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory and the Defense Government.

Pengpeng Yang would like to acknowledge the China Scholarship Council, State Scholarship Fund, that supports his joint Ph.D program.

REFERENCES

[1] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, “An overview on video forensics,” APSIPA Transactions on Signal and Information Processing, vol. 1, 2012.

[2] T. Shanableh, “Detection of frame deletion for digital video forensics,” Digital Investigation, vol. 10, no. 4, pp. 350 – 360, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1742287613001102.

[3] D. Vázquez-Padín, M. Fontani, T. Bianchi, P. Comesaña, A. Piva, and M. Barni, “Detection of video double encoding with gop size estimation,” in IEEE 2012 International Workshop on Information Forensics and Security (WIFS), IEEE, 2012, pp. 151–156.
|               | Native | Avidemux | Exiftool | ffmpeg | Kdenlive | Premiere | Facebook | TikTok | Weibo | YouTube |
|---------------|--------|----------|----------|--------|----------|----------|----------|--------|-------|---------|
| Android       | 1.00   | 1.00     | 0.95     | 1.00   | 1.00     | 1.00     | 1.00     |        |       |         |
| iOS           | 0.95   | 0.95     | 1.00     | 0.95   | 1.00     | 1.00     | 1.00     | 0.05   |       |         |

T. Gloe, A. Fischer, and M. Kirchner, “Forensic analysis of video file formats,” *Digital Investigation*, vol. 11, Supplement 1, pp. S68 – S76, 2014, proceedings of the First Annual {DFRWS} Europe. [Online]. Available: [http://www.sciencedirect.com/science/article/pii/S1742287614000140](http://www.sciencedirect.com/science/article/pii/S1742287614000140)

D. Güera, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, “We need no pixels: Video manipulation detection using stream descriptors,” in *International Conference on Machine Learning (ICML), Synthetic Realities: Deep Learning for Detecting AudioVisual Fakes Workshop*, 2019.

M. Iuliani, D. Shullani, M. Fontani, S. Meucci, and A. Piva, “A video forensic framework for the unsupervised analysis of mp4-like file container,” *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 635–645, 2018.

J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

ISO/IEC 14496, “Information technology. coding of audio-visual objects, part 12: Iso base media file format, 3rd ed.” 2008.

———, “Information technology. coding of audio-visual objects, part 14: Mp4 file format,” 2003.

S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

L. Breiman, *Classification and regression trees*. Routledge, 2017.

H. Schütze, C. D. Manning, and P. Raghavan, “Introduction to information retrieval,” in *Proceedings of the international communication of association for computing machinery conference*, 2008, p. 260.

D. Shullani, M. Fontani, M. Iuliani, O. Al Shaya, and A. Piva, “VISION: a video and image dataset for source identification,” *EURASIP Journal on Information Security*, vol. 2017, no. 1, p. 15, 2017.

Apache, “Java mp4 parser,” [http://www.github.com/sannies/mp4parser](http://www.github.com/sannies/mp4parser).

“Statcounter: Globalstats 1999-2020,” [https://gs.statcounter.com/os-market-share/mobile/worldwide](https://gs.statcounter.com/os-market-share/mobile/worldwide).