A Benford's Law Based method to Detect Manipulated Digital Photos

Pedro Fernandes[14]
pedro.a.fernandes@ipleiria.pt,pedro.fernandes@tus.ie
Mário Antunes[123]
mario.antunes@ipleiria.pt

[1] Polytechnic of Leiria
School of Technology and Management, Leiria, Portugal
[2] Computer Science and Communication Research Centre, Polytechnic of Leiria - Portugal
[3] INESC-TEC, CRACS, Porto - Portugal
[4] Technological University of Shannon, Limerick, Ireland

Abstract

The automatic detection of manipulated digital photos has challenged criminal investigation. There is a wide range of techniques for detecting manipulations in digital photos, supported mainly by a set of machine learning methods. However, these techniques require substantial computational resources and make digital forensic analysis processes expensive.

This paper describes a statistical model based on Benford's Law and the results obtained with a dataset of 560 digital photos, of which 280 were authentic, and the remaining were manipulated. Benford's law returns the frequency with which the first digits occur on a logarithmic scale and can be applied to any database. The method was applied to a set of features (colours, textures) extracted from digital photos. It extracts the first digits, that is the frequency with which they occurred in the set of features extracted for each photo. The detection method focus on the behaviour with which the frequency of each digit occurred in comparison with the frequency expected by Benford's law.

The proposed method integrates *Pearson's* and *Spearman's* correlations and *Cramér-Von Mises* (CVM) fitting model, applied to the first digit of a number consisting of several digits, obtained by extracting digital photos features through *Fast Fourier Transform* (*FFT*) method. The global results obtained are promising, but worst than those obtained with machine learning techniques. An F1 value of 64.74%, with a recall of 91.19% were obtained, using the CVM model.

Keywords: Benford's law, digital forensics, first digits law, statistical coefficient correlation

1 Introduction

Cybercrime is dynamic and affects all the EU Member States. Cybercriminals take advantage of the sufficiently robust Internet infrastructure, the users' negligence in making online payments and their high exposure about what they do online, [3]. Cybercriminals have found their way to perpetuate illicit activities, using powerful tools such as *Photoshop* to manipulate multimedia content, namely digital photos, by using splice and copy-move techniques

The motivation for crimes that include photos manipulation is diverse, whether personal or political. Generally, revenge pornography or paedophilia involving people in a more vulnerable context, [5], and blackmail for ransom are the most prominent, leading to severe multi-level implications in people's lives.

This paper describes the application of Benford's Law to detect tampered digital photos by splicing and copy-move techniques. The operation of the proposed model is based on the extraction of a differentiated set of features from the digital photos, calculated by the Fast Fourier Transform (FFT) method. The proposed model requires less CPU and memory processing, as it does not require the use of data for training, does not require specific hardware to produce results and the possibility of creating lightweight modules that can be included in the most diverse digital forensic tools is an asset that is worth exploring. In order to evaluate the reliability of the results, statistical correlations have been used, such as Pearson's chi-square correlation coefficient, Spearman´s correlation coefficient and Cramer-Von Misses (CVM) goodness of fit test [6].

The paper starts with the fundamentals behind Benford's Law operation, the proposed method and the dataset used, and ends with the results obtained and the main conclusions.

2 Benford´s Law fundamentals

Suppose we are in the presence of an independent and identically distributed random variable, $X = (X_1, X_2, ..., X_i), i = 1, 2, ..., n,\ \forall n \in \mathbf{N}$, and $D_i(X)$ represents the i^{th} significant decimal digit of X.

The probability mass function that best describes Benford's law is given by equation $P(D_i(X)) = \log\left(1 + \frac{1}{d}\right)$, if $d = \{1, 2, 3, ..., 9\}$, [1].

From the probability mass function, we can calculate the empirical frequency of each digit appearance. In this way, the probability of the number 1 is given by 0.301, the number 2 is given by 0.176 and so forth, until you get the probability of $d = 9$.

3 Benford's law based method

The proposed model is based on the analysis of the first digit extracted from the characteristics of digital photos, following a line of investigation aimed at answering the question: *"If we are faced with a database containing digital photos, it is possible to detect whether there are authentic or manipulated photos and which ones?* According to Benford's law, if there is a manipulation in the first digit, the graph will produce a curve different from the curve produced by Benford's law.

Figure 1 illustrates the overall architecture designed to apply Benford's law under the context of manipulated digital photos detection. It is based on the following three main building blocks: preprocessing, processing and analysis of the results.

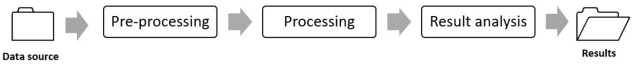

Figure 1: General architecture of the proposed method.

The preprocessing phase consists in extracting a set of n features (that refer to colours, textures and the shape of objects and their relationship) from the photos by applying the FFT (Fast Fourier Transform) method. The extracted data is stored in a vector of features, where the first digit of all the obtained values is further extracted and subsequently stored in a matrix of digits. The resulting vector is labelled (1-original; 0-manipulated). At the end of preprocessing phase, a labelled dataset is available to apply a set of hypothesis tests based on Pearson, Spearman and Cramer-Von Mises statistical models.

The processing phase is depicted in Figure 2 and consists of two steps. The first step counts the first digits from the values obtained in the preprocessing phase for each photo. The second step calculates the absolute frequency of each digit in the instances of the dataset. Then, the relative frequency of the values obtained in the two previous steps is calculated, consisting of the quotient between the absolute frequency of each digit and the sum of the total number of digits of each photo under study, allowing the subsequent comparison with Benford's law. Finally, the calculated relative frequency is stored in a dataset, allowing the calculation of Pearson's and Spearman's correlation coefficients, and the calculation of the Cramer-Von Mises-based goodness of fit test. For a better visual understanding of the data, graphs were generated from the relative frequencies of each photo and compared with the graph produced by Benford's Law.

The equations that allow the calculation of Pearson's correlation and Spearman's correlation are $r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X - \overline{x})^2 \sum_{j=1}^{n}(Y - \overline{Y})^2}}$ and $r(x,y) = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n^3 - n}$ where n is the length of each col-

umn, [2]. The Crámer-Von Mises criterion is defined by $W^2 = \int (F^*(t) - F_0(t))^2 dF_0(t)$ where $F^*(t) = \frac{k}{N}$ with k observations.

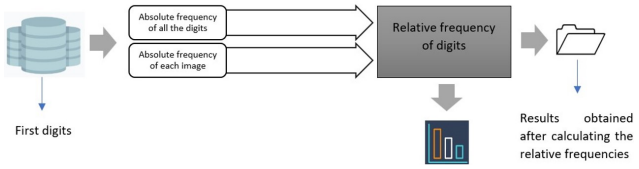

Figure 2: Processing stage

Figure 3 illustrates the processing performed by the hypothesis tests. Three hypothesis tests were introduced from the relative frequencies based on three different models: Pearson, Spearman, and Cramer-Von Misses. Each model allowed the generation of labels related to the evaluation, indicating 1 if the photo is genuine or 0 if the photo was manipulated. These labels are stored according to the statistical model used, and compared with the labels obtained in the preprocessing of the photos.

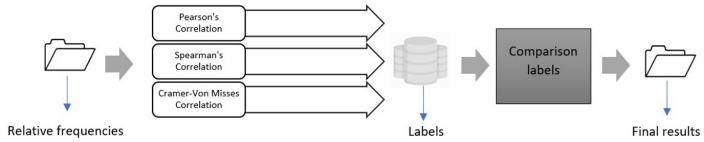

Figure 3: Processing performed by the hypothesis tests

The dataset used in the experimental tests comes compiles two datasets containing both authentic and manipulated photos, including splicing to copy-move manipulations. One of the datasets derives from Columbia Image Splicing Dataset, having 180 fake and 180 real photos. The second dataset comes from the Coverage Dataset, containing 100 fake and 100 real photos, for a total of 280 fake and real photos.

4 Results

The experiments analysed 280 authentic photos and 280 fake photos, for a total of 560 photos, starting with the extraction of 200 features, then 500 and finally 1000 features for each photo, by applying the FFT method in the pre-processing phase, and referring to colours, textures, and shapes. The main goal of the experiment focused on the possibility of detecting the manipulated and authentic photos present in the dataset. Tables 1, 2 and 3 contain the results obtained after extracting 200, 500 and 1000 features from the authentic and manipulated photos of the dataset.

	TP	TN	FP	FN	PR	RE	F1	AC
200	167	179	101	113	0.6231	0.5964	0.6095	0.6179
500	168	177	103	112	0.6199	0.6000	0.6098	0.6161
1000	166	179	101	114	0.6217	0.5929	0.6069	0.6161
Mean	167	178	101	113	0.6215	0.5964	0.6087	0.6167

Table 1: Results obtained after extracting 200, 500 and 1000 features from the photo dataset, using Pearson $\alpha = 0.001$

	TP	TN	FP	FN	PR	RE	F1	AC
200	221	75	205	59	0.5188	0.7893	0.6261	0.5286
500	218	75	205	62	0.5154	0.7786	0.6202	0.5232
1000	218	75	205	62	0.5154	0.7786	0.6202	0.5232
Mean	219	75	205	61	0.5165	0.7821	0.6221	0.525

Table 2: Results obtained after extracting 200, 500 and 1000 features from the photo dataset, using Spearman $\alpha = 0.001$

Comparing the average values obtained in the Tables 1, 2 and 3, it is possible to verify that Pearson model produces the better accuracy, 61.67%, concerning the other models evaluated. We may observe a high number of misclassified photos covering false positives, and false negatives. In comparison, the number of samples misclassified in the Spearman and CVM models is low for false negatives, 61 and 25, but high

	TP	TN	FP	FN	PR	RE	F1	AC
200	253	25	255	27	0.4980	0.9036	0.6421	0.4964
500	259	27	253	21	0.5059	0.9250	0.6540	0.5107
1000	254	28	252	26	0.5020	0.9071	0.6463	0.5036
Mean	255	27	253	25	0.5019	0.9119	0.6474	0.5035

Table 3: Results obtained after extracting 200, 500 and 1000 features from the photo dataset, using CVM with $\alpha = 0.001$

for false positives, 205 and 253; the number of samples classified as manipulated when they were real is relatively high. Comparing the results obtained in the tables by applying unusual statistical-based methods over conventional machine learning-based models [4], the current model is not competitive and produces worst results. For example, in [4] the F1 score of Support Vector Machines reaches 99.8%, considerably higher than the best result obtained by the Benford's Law based model.

The research was based on statistical models without training data as with machine learning-based methods, limiting the analysis to only the first digit of the data extracted from each photo.

A possible justification for these results being worse than those obtained by learning processes may be related to how the researcher got the photos. The extraction of the characteristics of a photo is performed from its pixels, and a care must be taken to clearly identify the quantity and quality of the extracted pixels. It is essential that the data can be obtained directly from the source devices (cameras, sensors), avoiding the possibility of noise (heating of a sensor, for example) that may alter the pixels of the photo and thus affect the results of the proposed model.

5 Conclusions

The present research consisted of creating a model based on Benford's law that allowed a fast response in the detection of manipulated photos, which could be an important help in criminal investigation and digital forensics. The proposed model has the added value of not being demanding in terms of computational resources, making the resulting digital forensic investigation process less expensive.

The method based on Benford's law was applied to a set of features (colours, textures) extracted from digital photos, through the extraction of the first digit. This procedure allowed using a set of unusual statistical techniques, generating a set of results that, despite not being better than those obtained by conventional methods, are promising. The possible causes for a lower performance of the proposed model, when compared with models based on machine learning, may be related to how the photos were obtained, the possibility of being faced with photos with low resolution, or in the low sensitivity of the correlation coefficients used in the present investigation. In this sense, it may be necessary to obtain all the photo characteristics and apply the proposed model to the remaining digits.

References

[1] Theodore P. Hill Arno Berger. *An Introduction to Benford's Law*. Princeton University Press, 2015. ISBN 0691163065.

[2] D. J. Best and D. E. Roberts. Algorithm AS 89: The upper tail probabilities of spearman's rho. *Applied Statistics*, 24(3):377, 1975. doi: 10.2307/2347111.

[3] Europol. Cybercrime. URL `https://www.europol.europa.eu/crime-areas-and-statistics/crime-areas/cybercrime`. Accessed 30 June 2022.

[4] Sara Ferreira, Mário Antunes, and Manuel E. Correia. Exposing manipulated photos and videos in digital forensics analysis. *Journal of Imaging*, 7(7):102, jun 2021. doi: 10.3390/jimaging7070102.

[5] Douglas A. Harris. Deepfakes: False pornography is here and the law cannot protect you. *Duke law and technology review*, 17:99–127, 2019.

[6] Neetu Singh and Rishab Bansal. Analysis of benford's law in digital image forensics. mar 2015. doi: 10.1109/icspcom.2015.7150688.