ANTIFORENSICS ATTACKS TO BENFORD'S LAW FOR THE DETECTION OF DOUBLE COMPRESSED IMAGES

Simone Milani, Marco Tagliasacchi, Stefano Tubaro

DEI - Politecnico di Milano
P.za L. Da Vinci, 32 - 20133 Milano - Italy
e-mail: milani/tagliasa/tubaro@elet.polimi.it

ABSTRACT

Researchers have been recently challenging the robustness of forensic algorithms by designing antiforensic strategies that try to fool them. In this paper, we propose an antiforensic strategy that targets double image compression detectors based on Benford's law (or first digit law). The proposed approach is able to modify the first digit statistics of the considered data (a double compressed image) to fool single/double compression detectors based on Benford's law. In this way, the proposed strategy tries to mimick the effects of a single compression with limited additional distortion. The presented algorithm performs better than previous state-of-the-art antiforensic strategies and can be easily extended to other fraud detection methods.

1. INTRODUCTION

Altering and tampering a digital image is a relatively-easy task considering the wide availability of image editing software and the versatility of digital formats. As a consequence, several image tempering detection approaches have been designed in order to find out whether an image has been altered or not.

Within the research fields connected to image tampering detection, a significant role is played by algorithms that are developed to detect the artifacts left by compression [1]. Most of the digital images are available in compressed format, and therefore, any processing step can be associated with a decoding and a re-encoding operations. Compression operations leave some traces in the reconstructed image (referred as "footprints"). By detecting and identifying these traces, it is possible to understand if an image was compressed once or more [2].

From these premises, several double compression detection algorithms have been recently proposed in literature (e.g., [3]). They are mainly based on the analysis of the statistics related to the quantized transform coefficients. Many of them rely on the so-called *Benford's law* (or *first digit law*), which permit revealing double compression by analyzing the probability mass function (pmf) of the most significant decimal digit (also called *first digit*) of the quantized transform coefficients [2, 4, 5]. Though these forensic techniques are quite suitable for detecting standard image manipulations, they do not account for the possibility that a forensic-aware adversary may enact anti-forensic modifications to hide traces of image manipulation and recompression. A recent work has shown that such operations can be designed to successfully fool existing image forensic techniques [6]. Most of antiforensic strategies targeting double compression operate on the reconstructed image after the first compression altering the data to disguise the traces of the first coding stage [7]. However, attacking a specific detection algorithm permits minimizing its detection accuracy more effectively with respect to the use of a generic antiforensic techniques that tries, for example, to reconstruct the first order statistics of the transformed coefficients.

In this paper, we target the double compression detectors that are based on Benford's law. More precisely, our approach operates after the second compression on the DCT coefficient statistics in order to make the pmf of their first digits (FDs) conforming to distribution expected for a single compressed image. The approach proves to be more effective with respect to the approach in [6] since it is possible to make a double-compressed image look like a single-compressed image at a lower cost in terms of distortion. Moreover, since the approach generically manipulates the data according to their statistics, it can be easily extended to other forensic applications where Benford's law is employed (e.g., financial verifications, election fraud detection, etc..). Moreover, since the approach aims at smoothing the oscillation in FD probability derived from coefficient statistics after double compression, the approach could work for other detectors operating on the coefficient histogram.

In the following, Section 2 briefly describes how double JPEG compression detectors work and how antiforensic algorithms try to fool them. Section 3 presents Benford's law, while Section 4 describes the proposed algorithm. Section 5 compares the performance of the approach with some state-of-the-art solutions and Section 6 draws the final conclusions.

2. ANTIFORENSICS OF DIGITAL IMAGES

Most digital images are available on-line in compressed format, and nearly 90 % of them are coded according to the JPEG standard [8]. After color space conversion, a JPEG coder divides the input image into blocks $\mathbf{X}$ of 8×8 pixels. These are then transformed into blocks $\mathbf{Y}$ of coefficients by applying a bidimensional Discrete Cosine Transform (DCT) on them. Considering a generic transform coefficient block $\mathbf{Y}$, each element is quantized into an integer index $\mathbf{Y}_{\Delta_1}$

$$Y^1_{\Delta_1}(i,j) = \text{sign}(Y(i,j))\text{round}\left(\frac{|Y(i,j)|}{\Delta_1(i,j)}\right), \qquad (1)$$

where the indexes (i,j) denote the position of the elements within the 8×8 block. The values $Y^1_{\Delta_1}(i,j)$ are converted into a binary stream by an entropy coder that follows a zig-zag scan that orders coefficients according to increasing spatial frequencies. The coded block can be reconstructed by applying an inverse DCT transform on the rescaled coefficients $Y^1_r(i,j) = Y^1_{\Delta_1}(i,j) \cdot \Delta_1(i,j)$. Note that the quantization step $\Delta_1(i,j)$ changes according to the index (i,j) of the DCT coefficient and is usually defined by means of a quantization matrix. In the IJG (Independent JPEG Group) implementation, the quantization matrix is selected by adjusting a quality factor (QF), which varies in the range $[0, 100]$. The higher QF, the higher the quality of the constructed image. The block $\mathbf{Y}^1_r$ is then inversely-transformed to the block $\mathbf{X}^1_r$ and the decoded image is then reconstructed.

When the image is encoded a second time, the resulting quantization levels are

$$Y^2_{\Delta_2}(i,j) = \text{sign}(Y^1_{\Delta_1}(i,j))\text{round}\left(\frac{|Y^1_{\Delta_1}(i,j) \cdot \Delta_1(i,j)|}{\Delta_2(i,j)}\right), \qquad (2)$$

where $\Delta_2(i,j)$ are the quantization steps of the second compression stage.

Every quantization step introduces some coding artifacts on the reconstructed coefficients that can be exploited by a forensic analyst to estimate the number of compression stages and the adopted coding parameters (e.g., Δ).

More precisely, the pmf of $Y^1_r(i,j)$ presents a comb-like shape derived from quantization such that the associated non-zero coefficient levels $Y^1_{\Delta_1}(i,j)$ are distributed according to a geometric probability mass function [6]. As for coefficients $Y^2_r(i,j)$, it is possible to verify that they still follows a comb-like distribution, but the pmf of $Y^2_{\Delta_2}(i,j)$ does not follow a geometric probability mass function. Verifying these conditions permits detecting whether an image has been compressed once or twice.

However, it is possible to operate some manipulations on the blocks $\mathbf{Y}^1_r$ or $\mathbf{X}^1_r$ in order to make $\mathbf{X}^1_r$ look like uncompressed.

The antiforensic approach proposed in [6] considers that the pmf of $\mathbf{Y}^1_r$ presents a comb-like shape, and this property allows a forensic analyst to detect that block $\mathbf{X}^1_r$ has been compressed. It is possible to add some dither noise $N(i,j)$ to $Y^1_r(i,j)$ so that the resulting coefficients $Y'^1_r(i,j) = Y^1_r(i,j) + N(i,j)$ present a probability density function close to a Laplacian variable, i.e., $p(a) = \beta/2 \ \exp(-\beta|a|)$. In case the noise $N(i,j)$ is accurately shaped, the block $\mathbf{X}'^1_r$, which has been reconstructed from $\mathbf{Y}'^1_r$, can be compressed a second time into block $\mathbf{X}^2_r$, and the statistics of the associated transform coefficients $\mathbf{Y}'^2_r$ presents a comb-like structure as if a single compression has been applied.

However, it is possible to obtain a better performance by targeting the specific double compression strategy that is adopted by the forensic analyst. Unlike [6], the approach proposed in this paper operates after the second compression stage on the coefficients $\mathbf{Y}^2_{\Delta_2}$ in order to tackle a detector based on the so-called Benford's law (presented in the following section).

3. BENFORD'S LAW FOR DIGITAL COMPRESSED IMAGES

In order to detect double JPEG compression, many approaches have been proposed in literature. Many of these rely on detecting the violation of the so-called Benford's law (also known as first digit law or significant digit law) [9]. Given the most significant digit or first digit m of a strictly-positive integer Y (in base-10 notation)

$$m = \text{FD}(Y) = \left\lfloor \frac{Y}{10^{\lfloor \log_{10} Y \rfloor}} \right\rfloor, \qquad (3)$$

Benford's law [9] is satisfied whenever the probability mass function (pmf) of m can be well-approximated by the equation

$$p(m) = N \log_{10}\left(1 + \frac{1}{m}\right) \qquad \text{or} \qquad (4)$$

$$p(m) = N \log_{10}\left(1 + \frac{1}{\beta + m^\alpha}\right) \quad \text{(generalized)}, \qquad (5)$$

where N is a normalizing factor and α, β are the parameters characterizing the model. This property can be verified for many real-life sources of data and can be used effectively in detecting alterations and frauds, like double and multiple compression in JPEG images [5].

More precisely, the empirical pmf $\hat{p}(m)$ of the first digit is computed for transform coefficients located at the most significant spatial frequencies, and the interpolating Benford's equation $p(m)$ is computed on them. Then, the differences between $p(m)$ and $\hat{p}(m)$ are classified; many solutions employ approaches based on Support Vector Machine (SVM) to evaluate how well $\hat{p}(m)$ is fitted by the statistics of $p(m)$ (see [5, 4, 3]).

It is possible to generalize the classification operated by SVM using the Kullback-Leibler (KL) divergence between

$p(m)$ and $\hat{p}(m)$, i.e.,

$$D_{KL}(p \parallel \hat{p}) = \sum_{m=1}^{9} p(m) \ln \frac{p(m)}{\hat{p}(m)} \quad (6)$$

taking inspiration from [10]. In case the average KL-divergence associated to all the considered DCT coefficients is lower than a given threshold T_{KL} (which is set by the forensic analyst), the image is considered as single compressed. Otherwise, it is deemed to be double compressed.

Even in the case an image is single compressed, the observed KL-divergence is not zero (since fitting could not be perfect). In this case, let D^1_{KL} be the KL divergence between the empirical pmf $\hat{p}(m)$ for a single-coded image and its fitted Benford's equation $p(m)$. Similarly, let D^2_{KL} be the equivalent KL divergence for the same image coded twice. Antiforensic algorithms can modify the reconstructed image after the first compression in order to make D^2_{KL} as close as possible to D^1_{KL} (or at least lower than T_{KL}). The price to be paid for this alteration is an additional distortion in the reconstructed image (which can be measured by quality indexed PSNR or SSIM indexes). As a matter of fact, an effective algorithm should obtain a KL divergence $D^2_{KL} < T_{KL}$ with a minimum quality decrement.

4. THE PROPOSED ALGORITHM

As mentioned in the previous section, an analyzed image can be considered as single compressed if the first digit statistic $\hat{p}(m)$ proves to be close to the fitted Benford's model $p(m)$ of eq. (4). In alternative to the traditional antiforensics approach presented in the last part of the previous section, an image tamperer can alter directly the coefficients $Y^2_{\Delta_2}(i,j)$ of the image after the second quantization in order to shape the resulting $\hat{p}(m)$ such that $D^2_{KL} \rightarrow D^1_{KL}$.

Since most of the double compression detectors based on Benford's law process coefficients separately depending on their spatial frequencies, the proposed antiforensic algorithm groups the coefficients $Y^2_{\Delta_2}(i,j)$ of an image into arrays $\mathbf{c}_{i,j}$ related to position (i,j) in the transform block. In the following we will omit the coordinates (i,j) for the sake of conciseness. Every elements of the array $\mathbf{c} = [c_k]$ can be associated to its first digit m_k (grouped into the array $\overline{\mathbf{m}} = [m_k]$).

For every array $\overline{\mathbf{m}}$ it is possible to compute $\hat{p}(m)$ and compute the Benford's model $p(m)$ that minimizes the difference $d(m) = (\hat{p}(m) - p(m))$ for $m = 1, \ldots, 9$. Then, it is possible to divide the set of possible values for m into two subsets $M_p = \{m | d(m) > 0\}$ and $M_n = \{m | d(m) < 0\}$. The shaping algorithm needs to transfer part of the probabilities associated to $\hat{p}(m)$ in M_p to those in M_n in order to fit $\hat{p}(m)$ to $p(m)$. This can be done converting coefficient values c_k such that $m_k \in M_p$ into values c'_k such that $FD(c'_k) \in M_n$ minimizing the resulting distortion in the reconstructed image.

Let us assume that we want to modify K coefficients. This budget is assigned to different value in M_p. Let $K(m)$ denote the number of coefficients with FD m that we want to modify, i.e., $K = \sum_{m \in M_p} K(m)$. Similarly, let $K'(m)$ be the negative number of coefficients with FD $m \in M_n$ that we want to generate in the array $\mathbf{c}$ ($K = \sum_{m \in M_n} -K'(m)$).

For every value c_n, the algorithm computes the distortion

$$e_{k,m'} = \min\{|c_k - m' 10^{o(k)}|, |c_k - m' 10^{o(k)+1}|\} \quad (7)$$

where $o(nk)$ is the order of the first digit, i.e., $o(k) = \lfloor \log_{10} |c_k| \rfloor$. In this case, $e_{k,m'}$ characterizes the distortion produced by converting c_k with FD m into the closest value with first digit m'.

Given a certain FD value $m \in M_p$ such that $K(m) > 0$, the approach considers the set of coefficient indexes $C_m = \{k | m_k = m\}$ and finds out

$$\begin{aligned} (k^*, m^*) = \arg\min_{k,m'} & e_{k,m'} \\ s.t. \quad & k \in C_m, m' \in M_n \\ & \text{and } K'(m') < 0. \end{aligned} \quad (8)$$

The coefficient c_{k^*} is then converted into either $m^* 10^{o(k^*)}$ or $m^* 10^{o(k^*)+1}$, according to its proximity, and counters $K(m)$ and $K'(m^*)$ are updated as follows, $K(m) \leftarrow K(m) - 1$ and $K'(m^*) \leftarrow K'(m^*) + 1$. The process is iterated as long as there are couples m, m' such that $K(m) > 0$ and $K'(m') < 0$.

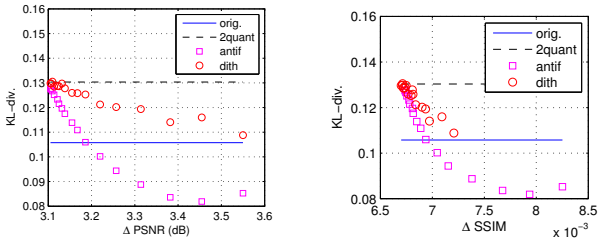
The approach is similar to iterative *water filling* [11] for power allocation in MIMO channels, despite in this case the objective is a redistribution of data with minimum distortion.

In the following section, we will show the performance of this approach both on generic exponential variables and on real images.

5. EXPERIMENTAL RESULTS

The test reported here have been obtained with images from UCID database [12]. We compressed them twice with different couples of quantization factors (QF_1, QF_2), where QF_1 is adopted at the first compression stage and QF_2 is employed in the second compression. The performance was evaluated measuring the average D^1_{KL} and D^2_{KL} over all the possible FD distributions obtained from histograms of coefficients at different spatial frequencies. In this work we limited the antiforensic alteration to a subset of spatial frequencies since most of the double compression detectors limit their analysis to a subset of frequencies (see [4, 3]). However, the processing can be easily extended to all the spatial frequencies at the expense of a stronger distortion and a lower reliability of the forensic detector.

Fig. 1 reports the KL-divergence obtained by our approach (labelled `antif`) and the approach in [6] (labelled `dith`) for the image indexed as 790 in the considered database as a function of the PSNR decrement between the



(a) (b)

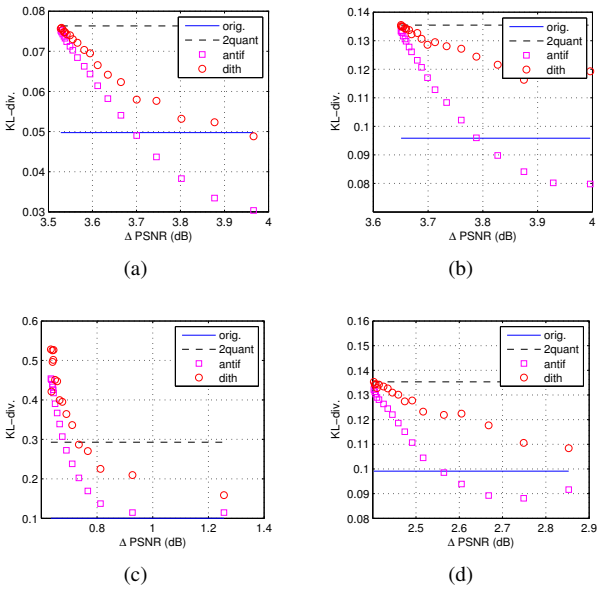
Fig. 1. Comparison performance for algorithms `antif` and `dith` on image 790. The graphs report the values of KL divergence as a function of different metrics. Rate-Distortion performance is reported as well. Quantization steps are $\Delta_1 = 7$ and $\Delta_2 = 9$. (a) KL div. vs. ΔPSNR; (b) KL div. vs. ΔSSIM.

original coded twice image and the one to which antiforensic technique is applied (Fig. 1a) and the SSIM decrement (Fig. 1b). It is possible to notice that the `dith` algorithm,with respect to the proposed algorithm, requires an additional 0.2 dB reduction in the PSNR value to obtain a divergence value equal to 0.08. The same efficiency can be noticed looking at the decrement in the SSIM value (Fig. 1b). It is possible to see that the decrement slope is much steeper for the proposed solution.

Fig. 2 shows the KL-div. vs. Δ PSNR graphs for different images labelled as 656, 121 and 790, respectively. It is possible to notice that the proposed solution outperforms the algorithm `dith` in most of the cases. It is also possible to notice that in some cases (e.g., the image of Fig. 1a) the algorithm `dith` decreases the KL-divergence down to a given limit. No additional distortion on the coefficients can make the KL-divergence closer to that of a single compressed image. On the other hand, the proposed solution can always obtain a KL-divergence lower than that of a single compressed image (and therefore, within the acceptance region of the forensic analyst).

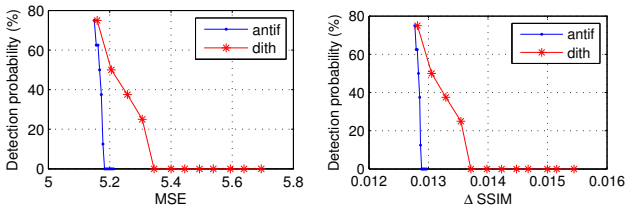
Further experiments were run considering different quantization steps. Fig. 2c and 2d report the KL-div. vs. Δ PSNR curves obtained on the image 790 for $\Delta_1 = 9, \Delta_2 = 7$ and $\Delta_1 = 9, \Delta_2 = 11$. It is possible to notice that the effectiveness of the `antif` algorithm is more evident at higher distortions.

In the end, we report the double compression detection probability vs. MSE and ΔSSIM for both algorithms obtained classifying the doctored image with the classifier in [4]. 100 randomly-selected images from [12] database were coded with $\Delta_2 = 7$ and $\Delta_1 \in [\Delta_2 - 5, \Delta_2 + 5]$ (uniform random variable). These results show that the evidence found for KL-divergence correspond to that obtained for a real detector.



(a) (b) (c) (d)

Fig. 2. KL divergence vs. ΔPSNR for different images and quantization steps. Results refer to $\Delta_1 = 7$ and $\Delta_2 = 9$ for image 656 (a) and image 121 (b); $\Delta_1 = 9$ and $\Delta_2 = 7$ (c) and $\Delta_1 = 9$ and $\Delta_2 = 11$ (d) for image 790.



(a) (b)

Fig. 3. Double compression detection probability vs. MSE (a) and ΔSSIM (b) for `antif` and `dith` approaches on UCID database.

6. CONCLUSIONS

The paper presented an antiforensic strategy targeting forensic methods based on Benford's law. The proposed approach alters the statistics of the first digit by changing the values of the processed data minimizing a distortion function. Experimental results show that the proposed solution permits modulating the introduced distortion and the probability of fooling the forensic detector while requiring a lower amount of distortion. Moreover, the proposed approach is quite general and can be applied to a generic forensic detector based on Benford's law (not only for images).

Acknowledgement

This work was partially supported by the EU FP7 FET project REWIND grant number:268478.

7. REFERENCES

[1] Jan Lukás and Jessica Fridrich, "Estimation of primary quantization matrix in double compressed jpeg images," in *Proc. of DFRWS*, 2003.

[2] D. Fu, Y. Q. Shi, and W. Su, "A generalized Benfords law for JPEG coefficients and its applications in image forensics," in *Proc. of SPIE*, Jan. 28 – Feb. 1, 2009, vol. 6505, pp. 39–48.

[3] T. Pevny and J. Fridrich, "Estimation of primary quantization matrix for steganalysis of double-compressed JPEG images," in *Proc. of SPIE*, San Jose, CA, USA, Jan. 2008, vol. 6819, pp. 11–1 – 11–13.

[4] B. Li, Y. Q. Shi, and J. Huang, "Detecting doubly compressed JPEG images by using mode based first digit features," in *Proc. of MMSP 2008*, Cairns, Queensland, Australia, Oct. 2008, pp. 730–735.

[5] S. Milani, M. Tagliasacchi, and M. Tubaro, "Discriminating multiple jpeg compression using first digit features," in *Proc. of ICASSP 2012*, Kyoto, Japan, Mar. 25 – 30, 2012, pp. 2253–2256.

[6] M. C. Stamm, S. K. Tjoa, W. S. Lin, and K. J. R. Liu, "Anti-forensics of jpeg compression," in *IEEE Int'l Conf. Acoustic, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, Mar. 2010, pp. 1694 –1697.

[7] M.C. Stamm, S.K. Tjoa, W.S. Lin, and K.J.R. Liu, "Undetectable image tampering through jpeg compression anti-forensics," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, Hong Kong, China, Sept. 2010, pp. 2109 –2112.

[8] G.K. Wallace, "The JPEG Still Picture Compression Standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, April 1991.

[9] Frank Benford, "The law of anomalous numbers," *Proceedings of the American Philosophical Society*, vol. 78, no. 4, pp. 551–572, Mar. 1938, JSTOR 984802.

[10] M. Barni, "A game theoretic approach to source identification with known statistics," in *Proc. of ICASSP 2012*, Kyoto, Japan, Mar. 25 – 30, 2012.

[11] M. Tao, Y. C. Liang, and F. Zhang, "Resource allocation for delay differentiated traffic in multiuser ofdm systems," *IEEE Trans. on Wireless Communications*, vol. 7, no. 6, pp. 2190–2201, June 2008.

[12] G. Schaefer and M. Stich, "UCID - An uncompressed colour image database," in *Proc. SPIE*, San Jose, CA, USA, Jan. 2004, vol. 5307, pp. 472 – 480.