JPEG stegaonography with side information from the processing pipeline

Rémi Cogranne

SELECTION-CHANNEL-AWARE REVERSE JPEG COMPATIBILITY FOR HIGHLY RELIABLE STEGANALYSIS OF JPEG IMAGES

Rémi Cogranne

ROSAS Dept. - LM2S Lab. - FRE 2019 CNRS - Troyes University of Technology

ABSTRACT

This paper deeply studies the principle of the recent reverse JPEG compatibility attack [1]. This analysis allows us to cast the problem of hidden data detection in DCT coefficients within hypothesis testing theory. The optimal LR test, thought efficient, is rather computationally expensive. Therefore, mild assumptions are used to simplify the detection problem dramatically and design a test that is simple yet extremely efficient and reliable. It is shown that the proposed detector is way more efficient than the original test [1], and allows highly reliable detection of data hidden within JPEG images.

Index Terms— Steganalysis, Hypothesis testing, JPEG compression, Statistical methods, Reliable Detection.

1. INTRODUCTION

Steganography and steganalysis form a cat-and-mouse game in which steganography aims at hiding data within innocuous-looking digital images. On the opposite, steganalysis aims at detecting images that contain hidden data. Over the past decades, steganography has been improved by the use of coding methods [2] that allows the hiding of a secret message almost as efficiently as the optimal rate-distortion bound. On the other hand, steganalysis has been developed by the use of machine learning method. Very large features sets [3] along with dedicated machine learning algorithms [4, 5] have been specifically designed to perform hidden information detection. However, it has also been observed that those machine learning methods may be dramatically impacted by the so-called cover-source mismatch: when the dataset used for learning only slightly differ from the (testing) dataset of interest, the performance may significantly drop.

This phenomenon poses a severe issue in an operational context where a highly reliable detection is crucial in order to avoid false alarm. The problem of reliable steganalysis has been a topic of research since almost two decades [6]: supervised learning methods have been designed with this goal [5, 7]. On the opposite to tackle this open problem of (highly) reliable steganalysis, a few attempts have been proposed using hypothesis testing theory [8]. However such methods usually have much lower statistical performances than those achieved by supervised learning techniques because a very accurate yet simple model of the cover signal is very difficult to obtain. Very recently, a promising test has been proposed for JPEG images compressed with high quality factor based on the so-called Reverse JPEG compatibility [1]. While extremely efficient, such test is not very reliable since (1) it does not achieve a high detection accuracy for a very low false-alarm rate and (2) its statistical performance remain analytically unknown. Based on the very same approach, the present paper extends the prior work [1] in order to address the aforementioned limitation regarding reliability.

The rest of the paper is organized as follows: Section 2 recalls the principle of JPEG compression and Reverse JPEG compatibility attack [1]. Then Section 3 formally states the problem of hidden data detection within hypothesis testing theory to derive the Most Powerful Likelihood Ratio Test (LRT). This test is hardly applicable in practice and, hence, simplified in Section 4 which leads to a selection-channel aware detector. Eventually, Section 5 presents numerical results that support the relevance of the present approach and the sharpness of the proposed detector.

The present work has been funded, in part, through French National Research Agency: ANR-18-ASTR-0009-02, ALASKA project https://alaska.utt.fr

2. REVERSE JPEG COMPATIBILITY TEST

In order to understand reverse JPEG compatibility [1], let us briefly recall how JPEG compression works. The main steps of JPEG compression are recalled in Figure 1; the reader is referred to [9] for details. In brief a color image is first converted into the so-called YCbCr color space which separates luminance (Y channel) and chrominance (CbCr). Then the Discrete Cosine Transform (DCT) is applied block-wise on each color channel independently over blocks of size 8×8. The ensuing DCT coefficients are eventually quantized adaptively using a quantization matrix; each DCT coefficient is divided with a different factor prior to the rounding operation. The DCT coefficients are eventually lossless compressed (typically using Huffman coding). It is important to note that all those operations can be implemented on floating point variables at a price of slower computation. Therefore, DCT is applied on integers which implies that pixels values in YCbCr spatial domain are rounded to the nearest integer prior application of the DCT. The very fact that standard JPEG compression libraries accept integer-valued variable has been recently exploited in image forensics [10, 11] yet remained unnoticed in steganalysis.

During decompression, DCT coefficients are lossless decompressed then each block values are "scaled", or more precisely multiplied with the quantization factor; the inverse DCT is used to convert coefficients back into the spatial domain before ultimate conversion from YCbCr to RGB color space to get the final decompressed image.

The rounding of the value right before the application of DCT is extremely important in the present work. Indeed, the very last lossy steps of JPEG compression are roughly, rounding (in spatial domain), DCT and quantization (division and rounding in DCT domain). As already explained, the DCT works on blocks of 8×8; hence, let us denote $\mathbf{X}_1, \dots, \mathbf{X}_N$ the N blocks from an image. The DCT can be denoted as a linear change of basis:

$$\mathbf{Z}_n = \mathbf{D}^\top \mathbf{X}_n \mathbf{D}, \tag{1}$$

where matrix $\mathbf{D}$ is made of orthonormal vectors: $\mathbf{D}^\top \mathbf{D} = \mathbf{I}$;
For simplicity and clarity, first we will focus on a single block and

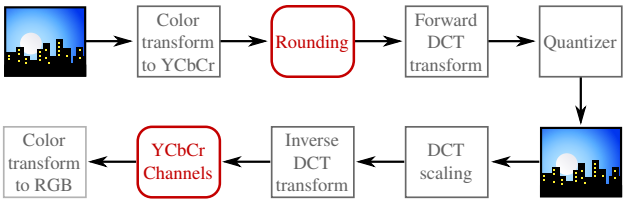

Fig. 1. Illustration of the main steps from JPEG compression and decompression; the quantization on which this paper focuses is highlighted in red.

drop the index n and, second, we will rewrite the DCT linear transformation, putting both matrices $\mathbf{Z}$, $\mathbf{X}$ into column vectors of 64 elements, as:

$$\mathbf{z} = \mathbf{D}^\star \mathbf{x}, \tag{2}$$

where $\mathbf{z}$ and $\mathbf{x}$ correspond to $\mathbf{Z}$ and $\mathbf{X}$ put into column vectors and $\mathbf{D}^\star$ is a matrix of size 64×64 whose rows are given by the product of rows of $\mathbf{D}$.

The ultimate step of quantization can be expressed as two steps, division and rounding:

$$\overline{\mathbf{z}} = \text{Round}\left(\mathbf{z} \oslash \mathbf{q}\right), \tag{3}$$

with $\overline{\mathbf{z}}$ the quantized DCT coefficients, $\mathbf{q}$ the vector that contains the quantization factor (that generally differ for each coefficient) and the operation $\oslash$ stands for the element-wise division.

During the decompression, the DCT coefficients are scaled before applying the inverse DCT coefficients:

$$\widetilde{\mathbf{x}} = \mathbf{D}^{\star -1}\left(\overline{\mathbf{z}} \odot \mathbf{q}\right), \tag{4}$$

where $\widetilde{\mathbf{x}}$ is the decompressed pixels value, $\mathbf{D}^{\star -1}$ is the matrix that represents the inverse DCT transform, made in a similar fashion as $\mathbf{D}^\star$ and $\odot$ represents the operation of element-wise multiplication.

It is obvious that neglecting the rounding error the JPEG compression is reversible:

$$\mathbf{D}^{\star -1}\Big(\Big(\mathbf{D}^\star \mathbf{x} \oslash \mathbf{q}\Big) \odot \mathbf{q}\Big) = \mathbf{x}. \tag{5}$$

Our ultimate goal is to model statistically the impact of JPEG compression in spatial domain; to this end, let us define the quantization noise $\boldsymbol{\epsilon}$ by:

$$\boldsymbol{\epsilon} = \text{Round}\left(\mathbf{z} \oslash \mathbf{q}\right) - \left(\mathbf{z} \oslash \mathbf{q}\right) = \overline{\mathbf{z}} - \left(\mathbf{z} \oslash \mathbf{q}\right) \tag{6}$$

such that one can redefine the quantized DCT coefficients $\overline{\mathbf{z}}$ as corrupted by an additive noise:

$$\overline{\mathbf{z}} = \text{Round}\left(\mathbf{z} \oslash \mathbf{q}\right) = \mathbf{z} \oslash \mathbf{q} + \boldsymbol{\epsilon}. \tag{7}$$

Using the previous notations for quantization (7) and DCT (3) into pixel decompression (4), it is straightforward to write:

$$\widetilde{\mathbf{x}} = \mathbf{x} + \mathbf{D}^{\star -1}\Big(\boldsymbol{\epsilon} \odot \mathbf{q}\Big). \tag{8}$$

Assuming that the rounding error of DCT coefficients $\boldsymbol{\epsilon}$ can be modeled as a uniform noise the error in spatial domain $\widetilde{\mathbf{x}} - \mathbf{x}$ can be modeled, in virtue of Lindeberg's Central Limit Theorem (CLT) [12, Theorem 11.2.5], as a Gaussian random variable.

The reverse JPEG compatibility as proposed in [1] is based on this observation. While the pixels before quantization $\mathbf{x}$ are not available, one can instead use the difference between $\widetilde{\mathbf{x}}$ and $\text{Round}\left(\widetilde{\mathbf{x}}\right)$ which hence follows a so-called "folded" Gaussian distribution. When the quantization steps $\mathbf{q}$ are important, the distribution of rounding error in spatial domain tends to become uniform; However, for small quantization steps $\mathbf{q}$ (typically JPEG quality factors 100 and 99) the distribution of rounding error in spatial domain allows the detection of increase of variance due to data hiding. This simple observation leads to the following detection statistics in [1]:

$$\Lambda(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{1}{64N}\sum_{n=1}^{N}\left\|\widetilde{\mathbf{x}}_n - \text{Round}\left(\widetilde{\mathbf{x}}_n\right)\right\|_2^2, \tag{9}$$

which corresponds, with N the number of blocks from the given image, to the estimated variance of rounding error in spatial domain.

3. STATEMENT OF REVERSE JPEG COMPATIBILITY WITHIN HYPOTHESIS TESTING

In order to model the impact of hidden data into DCT coefficients $\mathbf{z}$ on the value of pixels decompressed into the spatial domain $\widetilde{\mathbf{x}}$, let us define the distribution of the rounding errors:

$$\widetilde{\mathbf{x}} - \text{Round}\left(\widetilde{\mathbf{x}}\right) \sim \mathcal{N}_F(0, \Sigma_s), \tag{10}$$

where $\mathcal{N}_F$ denotes the folded Gaussian distribution [1] whose probability density function (pdf) is given by:

$$f_{\mu,\sigma}(x) = \begin{cases} \sum_{k\in\mathbb{Z}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x-k-\mu)^2}{2\sigma^2}\right) & \forall x \in (-0.5, 0.5), \\ 0 & \forall |x| > 0.5. \end{cases} \tag{11}$$

It follows from (8) that the covariance matrix Σ_s is diagonal whose element at location (i,i) is $\sum_{k=1}^{64} {d^\star_{i,k}}^{-2} \times q_k^2/12$ with $d^\star_{i,k}$ and q_k the elements from $\mathbf{D}^\star$ and $\mathbf{q}$ respectively.

Denoting $\mathbf{s}$ the additive stego-signal into DCT coefficients, the rounding error into spatial domain becomes, after data hiding:

$$\widetilde{\mathbf{x}} - \mathbf{x} = \mathbf{D}^{\star -1}\Big(\boldsymbol{\epsilon} \odot \mathbf{q}\Big) + \mathbf{D}^{\star -1}\Big(\mathbf{s} \odot \mathbf{q}\Big). \tag{12}$$

It thus follows that the rounding errors of a stego-image follows a shifted Gaussian folded distribution defined by:

$$\widetilde{\mathbf{x}} - \text{Round}\left(\widetilde{\mathbf{x}}\right) \sim \mathcal{N}_F(\mathbf{D}^{\star -1}\mathbf{s} \odot \mathbf{q}, \Sigma_s). \tag{13}$$

Since, during the JPEG compression, the DCT is carried out over each and every blocks of 8×8 pixels separately, one can assume that those blocks are statistically independent. On the top of this model, recent steganographic methods are content adaptive, which means that each and every DCT coefficients have a different probability of being used to hide data. Denoting β_k the probability of modifying k-th DCT coefficient in n-th block, the joint probability of modifying together DCT coefficients within a block is given by $\mathbb{P}[\mathbf{s} = (s_1, \ldots, s_{64})] = \beta_1^{s_1}(1-\beta_1)^{1-s_1} \times \ldots \times \beta_{64}^{s_{64}}(1-\beta_{64})^{1-s_{64}}$ where s_k is a binary variable that indicates a change at k-th DCT coefficient. Even assuming that each DCT coefficient can be changed in only one direction, the probability distribution of the rounding errors in the spatial domain after embedding becomes:

$$f^{\boldsymbol{\beta}}_{\mu,\sigma}(x) = \sum_{\mathbf{s}^\star \in \mathcal{S}} \mathbb{P}[\mathbf{s} = \mathbf{s}^\star] f_{\mu,\sigma}\left(x + \mathbf{D}^{\star -1}\mathbf{s}^\star \odot \mathbf{q}\right). \tag{14}$$

Unfortunately, the exact statement (14) of such probability distribution involved a sum over the set $\mathcal{S}$ all possible changes whose cardinality is 2^{64} terms; which becomes 3^{64} if we assume that DCT

coefficients can be changed in both directions ± 1 (at a cost of more complex notations).

The present paper focuses on reliable detection in the sense that the probability of false-alarm must be controlled and possibly set to a very low value. To this end one can note that in the case where all parameters are known to the detector, the problem is reduced to a test between simple hypotheses: $\mathcal{H}_0 : \{\widetilde{\mathbf{x}} - \text{Round}\,(\widetilde{\mathbf{x}}) \sim \mathcal{N}_F(0, \Sigma_s)\}$ and $\mathcal{H}_1 : \left\{\widetilde{\mathbf{x}} - \text{Round}\,(\widetilde{\mathbf{x}}) \sim \mathcal{N}_F^{\boldsymbol{\beta}}(0, \Sigma_s)\right\}$. The pdf of those distributions are respectively given in Eq. (10) and (14). In such a context, that Neyman-Pearson lemma, see [12, Theorem 3.2.1], states that among the tests δ with a probability of false-alarm (PFA) bounded by α_0:

$$\mathbb{P}[\delta(\mathbf{x}) = \mathcal{H}_1 | \mathcal{H}_0] \leq \alpha_0, \tag{15}$$

the most powerful test, which achieves the highest possible detection power (often also referred to as the detection accuracy), defined as:

$$\mathbb{P}\left[\delta(\mathbf{x}) = \mathcal{H}_1 | \mathcal{H}_1\right], \tag{16}$$

is the Likelihood Ratio (LR) test defined, from the independence of DCT blocks, by:

$$\Lambda(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{n=1}^{N} \Lambda(\mathbf{x}_n) = \sum_{n=1}^{N} \frac{f_{\mu,\sigma}^{\boldsymbol{\beta}}(\mathbf{x}_n)}{f_{\mu,\sigma}(\mathbf{x}_n)} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau, \tag{17}$$

where the decision threshold τ is set, to satisfy the false-alarm constraint (15), as the solution of the equation $\mathbb{P}[\Lambda(\mathbf{X}) > \tau | \mathcal{H}_0] = \alpha_0$.

4. SIMPLIFICATION FOR SELECTION CHANNEL AWARE REVERSE JPEG COMPATIBILITY

The statement of hidden data detection problem, as described in Section 3, clearly shows that the exact formulation of the LR test must be simplified. This section aims at describing a few simplifications that allow the applying of the detection method in a fast and efficient way. The first simplification that we propose is to ignore the "folded" aspect of the distribution; Indeed, this very matter makes the distribution in the LR test (17) very complex to deal with. By assuming that rounding error in spatial domain follows a Gaussian distribution:

$$\widetilde{\mathbf{x}} - \text{Round}\,(\widetilde{\mathbf{x}}) \sim \mathcal{N}\Big(\text{Round}\left(\mathbf{D}^{\star -1}\mathbf{s} \odot \mathbf{q}\right), \Sigma_s\Big), \tag{18}$$

the distribution of rounding error from steganographic images can be modeled separately for each possible DCT coefficient due to the orthnonormality of DCT transform. Using this simplification, one can show that the LR test (17) becomes:

$$\Lambda^{\star}(\mathbf{x}_n) = \sum_{k=1}^{64} \frac{\beta_k}{2} \Bigg(\left\| \widetilde{\mathbf{x}} - \text{Round}\,(\widetilde{\mathbf{x}}) \right\|_2^2 - \left\| \widetilde{\mathbf{x}} + \mathbf{D}^{\star -1}\mathbf{1}_k \odot \mathbf{q} - \text{Round}\left(\widetilde{\mathbf{x}} + \mathbf{D}^{\star -1}\mathbf{1}_k \odot \mathbf{q}\right) \right\|_2^2 \Bigg), \tag{19}$$

with $\mathbf{1}_k$ the vector made of 0 except k-th element whose value is 1; Note that test (19) corresponds to a matched subspace detector [13].

One can note that the LR test (19) weights all possible changes of DCT coefficients by the probability that this coefficient is used during embedding, see proof in [14]. This is known in steganalysis as a "Selection-Channel" approach which consists in taking into account knowledge from the embedding during the detection.

The last simplification we proposed is similar to the one adopted in [1]. It essentially consists in assuming that the quantization step is small with respect to the noise, in which case it has been proved [5, 15] that the above test (19) is asymptotically equivalent to a test (9) on the variance, as originality proposed in [1]. However, when using this approach, the test (19) should be weighted by the probability of using each pixel. This is not straightforward since the test (9) operates on a block, of 64 pixels and DCT coefficients, while each DCT coefficient has a different probability of embedding. To this end, it is proposed to approximate the Selection-Channel approach by the expected number of changes into each block. Such simplified Selection-Channel Aware (sca) test is given by:

$$\Lambda^{\text{sca}}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{n=1}^{N} \left(\sum_{n=1}^{64} \beta_{k,n} \right) \left\| \widetilde{\mathbf{x}}_n - \text{Round}\,(\widetilde{\mathbf{x}}_n) \right\|_2^2. \tag{20}$$

Last, we wish to improve this test which requires the knowledge of the embedding scheme. To tackle this lack of knowledge, we seek at finding a Seletion-Channel Aware approach that approximates the probabilities of using the DCT coefficient that is quite accurate for a vast range of embedding schemes. To this end, we have noted that adaptive steganographic schemes embed more in DCT blocks whose values are large, as opposed to DCT coefficients made of small values that represents generally smooth and simple blocks. Therefore we propose the following weighted test:

$$\Lambda^{\text{w}}(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sum_{n=1}^{N} w_n \left\| \widetilde{\mathbf{x}}_n - \text{Round}\,(\widetilde{\mathbf{x}}_n) \right\|_2^2, \tag{21}$$

in which the weights w_n represents the sum of absolute value of DCT coefficients: $w_n = \sum_{k=1}^{64} |z_{n,k}| = \|\mathbf{z}_n\|_1$.

Last, but not the least, the present work aims at providing a reliable test, whose statistical performance can be established analytically. To this end, we propose to normalize the above test (20)-(21) as:

$$\overline{\Lambda}^{\text{w}}(\mathbf{x}_1, \ldots, \mathbf{x}_N) =, \tag{22}$$

with $\mu_0 \approx 0.0657$ the expected value of the decision statistics $\|\widetilde{\mathbf{x}} - \text{Round}\,(\widetilde{\mathbf{x}})\|_2^2$, under hypothesis $\mathcal{H}_0$, and $\sigma_0^2 \approx 0.0625$ its variance. It is straightforward from the CLT that the normalized test (22) follows a zero-mean Gaussian distribution with unit variance:

$$\overline{\Lambda}^{\text{w}}(\mathbf{x}_1, \ldots, \mathbf{x}_N) \sim \mathcal{N}(0, 1). \tag{23}$$

This statistical distribution allows us to guarantees a prescribed false-alarm rate α_0:

$$\mathbb{P}\left[\overline{\Lambda}^{\text{w}}(\mathbf{x}_1, \ldots, \mathbf{x}_N) > \tau\right] = \alpha_0, \tag{24}$$

by setting the decision threshold as follows:

$$\tau = \Phi^{-1}\left(1 - \alpha_0\right), \tag{25}$$

with Φ^{-1} the inverse of the normal cdf. As we shall see in Section 5, the setting of a PFA is extremely accurate in practice.

5. NUMERICAL RESULTS

In order to show the relevance of the proposed approach, we have carried out extensive numerical evaluations using both BOSSbase [16] and ALASKA base [17] respectively made of $10,000$ and $80,000$ images. We have used four different embedding schemes, from rather rusty nsF5 [18] to state-of-the-art adaptive J-UNIWARD [21] including EBS [19] (in its non-side-informed version) and UED [20].

Payload	Test [1]	Λ^{sca} (20)	Λ^{w} (21)	LR (19)
0.1 (4055)	0.8819	1.0000	1.0000	1.0000
0.06 (2339)	0.8326	0.9999	0.9991	0.9999
0.04 (1512)	0.7824	0.9991	0.9968	0.9987
0.025 (913)	0.7058	0.9975	0.9931	0.9968
0.015 (527)	0.5902	0.9930	0.9796	0.9830
0.01 (340)	0.4352	0.9728	0.9441	0.8876
0.006 (196)	0.1560	0.8374	0.7288	0.5502

Table 1. Comparison of efficiency of prior work and proposed steganalysis tests over BOSSbase [16], compressed with quality factor 100, against UED embedding scheme. Test performance is measured using detection power (16), i.e. true positive rate, for PFA of 0.1%.

Embedding	Test [1]	Λ^{sca} (20)	Λ^{w} (21)	LR (19)
UED (822)	0.0994	0.5973	0.4998	0.5624
EBS (509)	0.0428	0.2356	0.1854	0.0894
J-UNIWARD (606)	0.0489	0.3728	0.2108	0.2653

Table 2. Comparison of efficiency of prior work and proposed steganalysis tests over BOSSbase [16], compressed with quality factor 99, for various embedding scheme with payload 0.025 bpnzAC. Test performance is measured using detection power (16), i.e. true positive rate, for PFA of 0.5%.

Embedding	Test [1]	Λ^{sca} (20)	Λ^{w} (21)	LR (19)
UED (1364)	0.1297	0.7214	0.6352	0.6580
EBS (886)	0.0513	0.3825	0.2923	0.1074
J-UNIWARD (1020)	0.0553	0.4744	0.3068	0.2892

Table 3. Comparison of efficiency of prior work and proposed steganalysis tests over BOSSbase [16], compressed with quality factor 99, for various embedding scheme with payload 0.04 bpnzAC. Test performance is measured using detection power (16), i.e. true positive rate, for PFA of 0.1%.

First of all, to show the improvement in terms of detection accuracy, Tables 1-3 contrast the detection power (16) of the original detector, as proposed in [1], and the test proposed in Section 4. Table 1 shows the empirical detection power at PFA set to $\alpha_0 = 0.001$ obtain with UED [20] embedding scheme at various embedding payload over BOSSbase [16]. First, one can note that the proposed detectors dramatically improve the detection accuracy, especially for very low embedding payload; note that numbers in bracket represent the mean number of DCT coefficients changed. The proposed detectors maintained a very high detection power around or above 0.9 for as low as 340 changed DCT coefficients. Surprisingly, the simplified LR test (19) has a lower detection accuracy despite a much larger computational cost.

Next, Tables 2-3 show similar results, in terms detection power, for JPEG quality factor 99 and for several embedding schemes. Table 2 shows the detection accuracy for PFA of $\alpha_0 = 0.005$ and embedding rate of 0.025 bpnzAC (bits per non-zero AC coefficients). Table 3 shows the detection accuracy for PFA set to $\alpha_0 = 0.001$ and embedding rate of 0.04 bpnzAC. Again one can note the very large improvement especially for lowest number of changed DCT coefficients. One can also note that the Selection-Channel Aware detector (20) always reaches the highest accuracy; however, the proposed weighted detector achieves very competitive detection performance and has the advantages of not depending on the embedding scheme.

To conclude with the comparison of detectors' accuracy, Figure 2, proposes a ROC curve including results from prior work [1] along with Selection-Channel aware (20) and weighted detectors (21). Those results have been obtained with UED embedding scheme [20] (straight lines) and J-UNIWARD [21] (dashed lines)

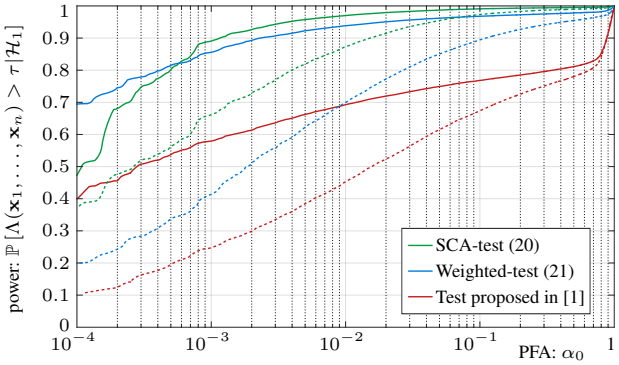

Fig. 2. Comparison of detectors performance through ROC curves; results obtained with UED (solid lines) [20] and J-UNIWARD (dashed lines) [21] embedding schemes over ALASKA base [17].

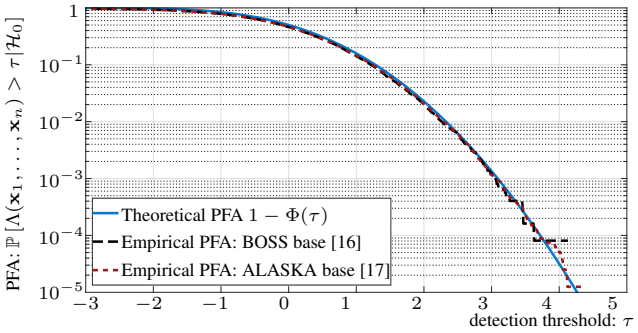

Fig. 3. Comparison between theoretical and empirical false alarm rate as a function of decision threshold for BOSSbase [16] and ALASKA base [17]

with payload 0.01 bpnzAC over the $80,000$ images from ALASKA base [17]. This large dataset of images allows us to draw with higher accuracy detection power for very low false alarm rate (typically up to 10^{-4}). Again, one can note that the proposed detectors allow achieving up to twice higher detection accuracy over prior work [1].

Eventually, besides improvements of detection accuracy, the second main contribution of present paper lies in the control of the false-alarm probability (22). To show the relevance of the proposed methodology, Figure 3 contrasts the theoretical false alarm rate and the empirical ones over two different datasets of images, BOSSbase [16] and ALASKA base [17]. One can note that the theoretical false-alarm rate deduced from CLT matches very well with empirical false-alarm rate up to below 10^{-4}. Those results show both the relevance of the proposed approach, which allows setting a decision threshold as a function of the desired false-alarm rate, as well as the sharpness of the statistical model.

6. CONCLUSION

The present work aims at extending the recent Reverse JPEG compatibility [1] for steganalysis of JPEG images compressed with highest quality factor. We have proposed an approach based on hypothesis testing theory. The present work shows the relevance of the approach proposed in [1] since highest detection accuracy is obtained using the same approach, which subpar the results obtained using simplified LR test. We show, however, that testing theory allows the designing of a Selection-Channel Aware test that achieves a much higher detection performance as well as a false alarm rate that can be controlled with high accuracy, which is of crucial importance in an operational context.

7. REFERENCES

[1] J. Butora and J. Fridrich, "Reverse JPEG compatibility attack *(DOI:10.1109/TIFS.2019.2940904, available as Early Access)*," *IEEE Transactions on Information Forensics and Security*, 2019.

[2] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *Information Forensics and Security, IEEE Transactions on*, vol. 6, no. 3, pp. 920–935, Sept 2011.

[3] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *Information Forensics and Security (WIFS), IEEE 6th International Workshop on*, December 2014, pp. 48–53.

[4] R. Cogranne, V. Sedighi, J. Fridrich, and T. Pevný, "Is ensemble classifier needed for steganalysis in high-dimensional feature spaces?," in *Information Forensics and Security (WIFS), IEEE 7th International Workshop on*, November 2015, pp. 1–6.

[5] R. Cogranne and J. Fridrich, "Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory," *Information Forensics and Security, IEEE Transactions on (in press)*, 2015.

[6] J. Fridrich, M. Goljan, and R. Du, "Reliable detection of LSB steganography in color and grayscale images," *IEEE Multimedia*, vol. 8, pp. 22–28, 2001.

[7] Y. Miche, P. Bas, A. Lendasse, C. Jutten, and O. Simula, "Reliable steganalysis using a minimum set of samples and features," *EURASIP Journal on Information Security*, vol. 2009, no. 1, pp. 901381, 2009.

[8] R. Cogranne, C. Zitzmann, L. Fillatre, I. Nikiforov, F. Retraint, and P. Cornu, "A cover image model for reliable steganalysis," in *Information Hiding, 13th International Workshop*, Prague, Czech Republic, May 18–20, 2011, Lecture Notes in Computer Science, pp. 178 – 192, LNCS vol.6958, Springer-Verlag, New York.

[9] ISO 12232:2006(E), "Photography – Digital still cameras – Determination of exposure index, ISO speed ratings, standard output sensitivity, and recommended exposure index," Standard, International Organization for Standardization, Geneva, CH, April 2006.

[10] T. H. Thai and R. Cogranne, "Estimation of primary quantization steps in double-compressed JPEG images using a statistical model of discrete cosine transform," *IEEE Access*, vol. 7, pp. 76203–76216, 2019.

[11] C. Pasquini and R. Bohme, "Towards a theory of JPEG block convergence," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 550–554.

[12] E.L. Lehmann and J.P. Romano, *Testing Statistical Hypotheses, Second Edition*, Springer, 3rd edition, 2005.

[13] L.L. Scharf and B. Friedlander, "Matched subspace detectors," *Signal Processing, IEEE Transactions on*, vol. 42, no. 8, pp. 2146 –2157, aug 1994.

[14] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *Information Forensics and Security, IEEE Transactions on*, vol. 11, no. 2, pp. 221 – 234, February 2016.

[15] R. Cogranne and F. Retraint, "An asymptotically uniformly most powerful test for LSB matching detection," *Information Forensics and Security, IEEE Transactions on*, vol. 8, no. 3, pp. 464–476, March 2013.

[16] P. Bas, T. Filler, and T. Pevný, "Break Our Steganographic System — the ins and outs of organizing BOSS," in *Information Hiding, 13th International Workshop*, Prague, Czech Republic, May 18–20, 2011, Lecture Notes in Computer Science, pp. 59–70, LNCS vol.6958, Springer-Verlag, New York.

[17] R. Cogranne, Q. Giboulot, and P. Bas, "The alaska steganalysis challenge: A first step towards steganalysis *into the wild*," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, New York, NY, USA, 2019, IH&MMSec'19, pp. 125–137, ACM.

[18] J. Fridrich, T. Pevný, and J. Kodovský, "Statistically undetectable JPEG steganography: Dead ends challenges, and opportunities," in *Proceedings of the 9th Workshop on Multimedia & Security*, New York, NY, USA, 2007, MM&Sec '07, pp. 3–14, ACM.

[19] C. Wang and J. Ni, "An efficient JPEG steganographic scheme based on the block entropy of dct coefficients," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 1785–1788.

[20] L. Guo, J. Ni, and Y.Q. Shi, "An efficient JPEG steganographic scheme using uniform embedding," in *Information Forensics and Security (WIFS), 2012 IEEE International Workshop on*, Dec 2012, pp. 169–174.

[21] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 2014, no. 1, pp. 1–13, 2014.