Are we there yet?

Mehdi Boroumand,[+] *Jessica Fridrich,*[+] *and Remi Cogranne*[×]
[+] *Department of ECE, SUNY Binghamton, NY, USA, {mboroum1,fridrich}@binghamton.edu*
[×] *Troyes University of Technology (UTT), Lab. of System Modeling and Dependability (LM2S), France, Remi.Cogranne@utt.fr*

Abstract

The purpose of this study is to prepare a source of realistic looking images in which optimal steganalysis is possible by enforcing a known statistical model on image pixels to assess the efficiency of detectors implemented using machine learning. Our goal is to answer the questions that researchers keep asking: "Are our empirical detectors close to what can be possibly detected? How much room is there for improvement?" or simply "Are we there yet?" Our goal is achieved by applying denoising to natural images to remove complex statistical dependencies introduced by processing and, subsequently, adding noise of simpler and known statistical properties that allows deriving the likelihood ratio test in a closed form. This theoretical upper bound informs us about the amount of further possible improvement. Three content-adaptive stego algorithms in the spatial domain and non-adaptive LSB matching are used to contrast the upper bound with the performance of two modern detection paradigms: a convolutional neural network and a classifier with the maxSRMd2 rich model. The short answer to the posed question is "We are much closer now but there is still non-negligible room for improvement."

Motivation

Steganography is the art of covert communication in which messages are hidden in cover objects so that the very existence of the secrets cannot be established. The objective of steganalysis is to detect the usage of steganography and do so as reliably as possible. A popular choice for cover objects today are digital multi media files, such as digital images, audio, and video. Such objects are ideal for covert communication for two reasons. They contain an indeterministic component, the acquisition noise, that helps mask the presence of steganographic embedding changes. Additionally, the inherent complexity of these objects is hard to capture using tractable and estimable statistical models, which further complicates detection. Steganographers finetune their embedding algorithms to locally adapt to content complexity since complicated textures and small-scale details are extraordinarily difficult to model statistically. This forced steganalysts to use complex high-dimensional (rich) media models [17, 27, 9, 18, 21, 13, 36, 35, 11, 12, 1] coupled with low-complexity classifiers, such as the FLD-ensemble [28] and the Low-Complexity Linear Classifier (LCLC) [10], possibly boosted by non-linear preprocessing [6, 7]. Recently, further progress has been achieved with non-linear differentiable hierarchical models with a large number of parameters, deep neural networks [32, 33, 41, 40, 42, 43, 46, 5, 39, 45].

It should be stressed that, fundamentally, it is the unavailability of statistical models for natural images that is responsible for this seemingly never ending spiral development. Steganography in artificial sources (sources with a known statistical model) can be perfectly secure[1] as covers can be synthesized [2] to communicate at a positive rate (payload whose size is linear w.r.t. the number of cover elements) [31, 38]. Likewise, optimal detectors of imperfect steganography methods in artificial sources can be constructed and their performance computed.

The situation is quite different for empirical sources that lack description using tractable and estimable statistical models. All steganographic methods inevitably become imperfect, the size of their secure payload sublinear in the number of pixels [26, 15, 25, 24], and detectors can be built that can distinguish between cover and stego objects better than randomly guessing. Without a cover model, however, we are unable to assess how good our steganography methods are and how well our detectors perform.

This paper is an attempt to address this problem by forming an artificial source of realistically looking images while forcing a known statistical model on pixels to allow derivation of optimal statistical tests for benchmarking empirical detectors built using machine learning. While it is entirely possible to synthesize artificial images for this purpose, the authors believe that it is valuable to keep a more realistic dataset with images visually similar to popular sources, such as BOSSbase 1.01 [3], in which content adaptive schemes execute changes with a similar selection channel as in the original source. We also need to avoid sources in which steganography would be too easy or too hard to detect while making sure that an optimal detector can still be derived. Since these requirements are in conflict, preparing a suitable source of both cover and stego images is quite challenging.

The idea for the cover source proposed in this paper was inspired by the experiment reported in Fig. 5 of [34]. The authors selected one BOSSbase image, denoised it, and then created 10,000 different versions of the same image by adding to it 10,000 independent realizations of a heteroscedastic sensor acquisition noise. Steganalysis in such a homogeneous cover source with the spatial rich model (SRM) [17] and MiPOD embedding algorithm [34] was reported to be rather close in terms of the Receiver Operating Characteristic (ROC) to the optimal statistical test designed for the noise component. However, for a *heterogeneous* source with images of diverse content, the SRM

[1] In Cachin's sense [8].

detector lagged behind the optimal Likelihood Ratio Test (LRT) quite a bit most likely due to the inability of the empirical detector to deal with the diversity of natural images (Fig. 6 in [34]).

The strategy adopted in this paper is to start with an existing dataset, apply a denoising filter to all images to remove complex noise introduced during acquisition and the subsequent development of the image from the raw sensor capture to a viewable form. Then, independent realizations of a Gaussian noise whose variance was estimated per pixel from the original images is reintroduced to force a known and tractable noise model in the cover source. This needs to be executed with care to prevent introducing dependencies among stego pixels. In particular, the pixels costs cannot be computed from the cover itself as the stego pixels would be locally dependent, which would prevent derivation of a closed-form LRT.

The proposed cover source dataset is described in the next section. Optimal statistical test and its properties appear in Section "Optimal test." In Section "Justification", we provide reasoning and experimental evidence for the choices made for the preparation of the cover and stego source. The results of our investigation and their interpretation are described in the subsequent Section "Experiments." The paper is concluded in the last section.

Cover and stego sources

We now describe the process for building the source of cover and stego images. The specific choices made are further explained, discussed, and experimentally justified in Section "Justification." The formation of the dataset proceeds in five steps and an additional sixth step for creating the stego images in a way that allows derivation of a closed-form optimal statistical test.

A simple way to describe the procedure for building the proposed dataset of cover images is to say that we start with an existing dataset that we modify to enforce the statistical model of pixels used in MiPOD [34]: the individual pixels in a cover image will be independent realizations of Gaussian random variables with a known mean and variance that depends on local content complexity.

The cover source was generated from the union of BOSSbase 1.01 [3] and BOWS2 [4] grayscale images resized from their original 512×512 size in Matlab using '`imresize.m`' with default parameters to a total of 20,000 256×256 grayscale images. This dataset will be denoted $\mathcal{B}$. The smaller image size was selected in anticipation that the best empirical detectors will be deep convolutional neural networks (CNNs), which typically require smaller images for effective training to fit reasonable size mini batches to the memory of current GPUs. We note that the latest designs for the spatial domain, the YeNet [44], the Yedroujd-Net [45], and the SRNet [5], were trained and benchmarked on this same database. We note that the methodology explained below can be applied to other, bigger datasets that may be needed for training deeper architectures in the future.

Step 1: Estimate pixel variance

We estimate the pixel variance σ_{ij}^2 from the original images in $\mathcal{B}$ using MiPOD's variance estimator as explained in Section V in [34]. This estimator was purposely designed to capture both the indeterministic Gaussian acquisition noise [16, 37, 23, 19] as well as local content complexity. For a given image and its pixel (i,j), $1 \le i,j \le 256$, we denote its estimated variance σ_{ij}^2. Note that the output of MiPOD's estimator is lower bounded: $\sigma_{ij} \ge 0.1$ for all i,j.

Step 2: Denoising

All images from $\mathcal{B}$ were first denoised to remove complex dependencies among pixels introduced by the RAW developer and subsequent processing. We used the wavelet denoising method described in [30] with Daubechies 8-tap wavelets and standard deviation of the removed i.i.d. Gaussian noise $\sigma_{\text{den}} = 10$. The pixel values in the denoised image were left in their non-rounded form but were clipped to the interval corresponding to 8-bit grayscale images $[0,255]$.

Step 3: Narrowing dynamic range

As the third step, the dynamic range of each denoised and clipped image was narrowed to the range $[15,240]$ by linearly mapping the interval $[0,255]$ to $[15,240]$ using :

$$g(x) = 15 + \frac{225}{255}x. \tag{1}$$

The scaled values were also rounded to integers, which we will denote $\mu_{ij} \in \{15,\dots,240\}$. The resulting 8-bit grayscale image with a narrower dynamic range will next be noisified with the variances estimated in Step 1 and further adjusted in Step 4.

Step 4: Adjusting the variance

The estimated pixel variances σ_{ij}^2 were adjusted so that the probability of a pixel getting out of the 8-bit dynamic range $[0,255]$ after noisification is equal to the probability of a one-sided 5σ Gaussian outlier (2.87×10^{-7}). This was done by making sure that σ_{ij} is smaller or equal to one fifth of the distance between the pixel mean μ_{ij} and the dynamic range boundary (0 or 255) :

$$\underline{\sigma}_{ij} = \min\Big\{\frac{1}{5}\min\{\mu_{ij}, 255-\mu_{ij}\}, \sigma_{ij}\Big\}. \tag{2}$$

Step 5: Noisifying

The noisified cover pixel c_{ij} is obtained by adding to μ_{ij} a sample ξ_{ij} from $\mathcal{N}(0,\underline{\sigma}_{ij}^2)$, rounding to an integer and clipping to $[1,254]$ to make sure the embedding will be free to modify all pixels by ±1 without getting out of the dynamic range. As explained above, we impose the cover image model from MiPOD – pixels are realizations of independent Gaussian variables $\mathcal{N}(\mu_{ij},\underline{\sigma}_{ij}^2)$ that are rounded to integers (rounding denoted with the square bracket $[\cdot]$), and then clipped to a finite dynamic range :

$$c_{ij} = [\mu_{ij} + \xi_{ij}] \tag{3}$$

$$c_{ij} = \begin{cases} c_{ij} & \text{if } 1 \le c_{ij} \le 254 \\ 1 & \text{if } c_{ij} \le 0 \\ 254 & \text{if } c_{ij} \ge 255. \end{cases} \tag{4}$$

The ij-th cover image pixel thus follows a probability mass function (p.m.f.) p_{ij} on $\{0,\ldots,255\}$, $c_{ij} \sim p_{ij}$:

$$p_{ij}(m) = \begin{cases} 0 & m = 0 \\ Q_{ij}\left(m - \frac{1}{2}\right) & m = 254 \\ Q_{ij}\left(m - \frac{1}{2}\right) - Q_{ij}\left(m + \frac{1}{2}\right) & 1 < m < 254 \\ 1 - Q_{ij}\left(m + \frac{1}{2}\right) & m = 1 \\ 0 & m = 255 \end{cases} \tag{5}$$

with $Q_{ij}(x)$ defined as the tail probability of $\mathcal{N}(\mu_{ij}, \underline{\sigma}_{ij}^2)$:

$$Q_{ij}(x) \triangleq \Pr\{\mathcal{N}(\mu_{ij}, \underline{\sigma}_{ij}^2) > x\}. \tag{6}$$

The values c_{ij} form the first data source used in our experiments, which we denote $\mathcal{B}(\sigma)$. We also report the results on a less noisy source obtained by multiplying the standard deviation σ_{ij} outputted by MiPOD's estimator in Step 1 by 1/2 (and then carrying out Steps 2–5 as above). This source will be denoted $\mathcal{B}(\sigma/2)$.

In Figure 1, we show a few examples of images from $\mathcal{B}(\sigma)$ for the reader to get a sense of how the images look. The viewer is encouraged to zoom into the pdf document to better see the small scale differences.

Stego images

All stego methods used in this study were ternary embedding algorithms that change each pixel i,j by ±1 with equal probability β_{ij}. Since we curbed the cover values in Step 5 to the interval $[1,254]$, the embedding does not need to be constrained in any way – all pixels can be changed by 1 or -1. For content adaptive steganography studied in this paper, the change rates β_{ij} are determined from pixel costs ρ_{ij} that in turn depend on a local neighborhood of pixel i,j. This dependence is quite complicated as the costs are usually computed in a non-linear fashion from outputs of several high-pass filters (e.g., as in S-UNIWARD [22], HILL [29], and WOW [20]). Consequently, computing β_{ij} from the noisified cover c_{ij} would create dependencies among stego pixels that are too complex to derive a closed-form expression for the distribution of stego pixels and tractable evaluation of the associated likelihood ratio test.

We resolved this problem by computing the change rates β_{ij} from the corresponding original image from $\mathcal{B}$. Another possibility is to compute the costs from a different independent noisification of the image, c'_{ij}. Since both versions gave similar results in our experiments, we opted for the former as a default for the rest of this paper.

Since β_{ij} does not depend on the specific noisification of the image and since the embedding changes are executed independently, the stego pixel p.m.f. is factorizable. In particular, it is a product of the following Gaussian mixtures q_{ij} over all pixels i,j :

$$q_{ij}(m) = \begin{cases} \beta_{ij} p_{ij}(254) & m = 255 \\ (1 - 2\beta_{ij}) p_{ij}(m) & \\ +\beta_{ij} p_{ij}(m-1) & \\ +\beta_{ij} p_{ij}(m+1) & 1 \le m \le 254 \\ \beta_{ij} p_{ij}(1) & m = 0. \end{cases} \tag{7}$$

For S-UNIWARD, HILL, and WOW, the change rates were obtained from an embedding simulator (e.g., assuming optimal source coding [14]).

Optimal test

Given an image with pixels s_{ij}, the steganalyst is facing the following statistical hypothesis test for all i,j :

$$\begin{aligned} \mathcal{H}_0 &: s_{ij} \sim p_{ij} \\ \mathcal{H}_1 &: s_{ij} \sim q_{ij}. \end{aligned} \tag{8}$$

We will assume that the parameters of the added MVG noise, the mean μ_{ij}, and the variance $\underline{\sigma}_{ij}^2$, are known. We also assume that the change rates β_{ij} are known. Under these assumptions, the test is simple, and, by the statistical independence of pixels, the optimal statistic is the log-likelihood ratio

$$\Lambda(\mathbf{s}) = \sum_{i,j} \Lambda_{ij}(s_{ij}) = \sum_{i,j} \log\left(\frac{q_{ij}(s_{ij})}{p_{ij}(s_{ij})}\right) \tag{9}$$

where $\Lambda_{ij}(m) = q_{ij}(m)/p_{ij}(m)$, $m \in \{0,\ldots,255\}$. For convenience, we will use the following normalized form of the log-LRT :

$$\Lambda^{\star}(\mathbf{s}) = \frac{\sum_{i,j} \Lambda_{ij}(s_{ij}) - E_{\mathcal{H}_0}[\Lambda_{ij}]}{\sqrt{\sum_{i,j} Var_{\mathcal{H}_0}[\Lambda_{ij}]}}, \tag{10}$$

where

$$E_{\mathcal{H}_0}[\Lambda_{ij}] = \sum_m p_{ij}(m)\Lambda_{ij}(m) \tag{11}$$

$$Var_{\mathcal{H}_0}[\Lambda_{ij}] = \sum_m p_{ij}(m)\Lambda_{ij}^2(m) - \left(E_{\mathcal{H}_0}[\Lambda_{ij}]\right)^2. \tag{12}$$

Under the fine quantization limit, $1 \le \sigma_{ij}$ for all i,j, and as the number of pixels approaches infinity, the Lindeberg's version of the Central Limit Theorem implies

$$\Lambda^{\star}(\mathbf{s}) \rightsquigarrow \begin{cases} \mathcal{N}(0,1) & \text{under } \mathcal{H}_0 \\ \mathcal{N}(\varrho,1) & \text{under } \mathcal{H}_1 \end{cases}, \tag{13}$$

where $\rightsquigarrow$ means convergence in distribution and $\varrho = \sum_{i,j} \underline{\sigma}_{ij}^{-4} \beta_{ij}^2 > 0$ is the deflection coefficient.

We note that technically the fine quantization assumption is not satisfied for all pixels because the standard deviation outputted by MiPOD's variance estimator is only

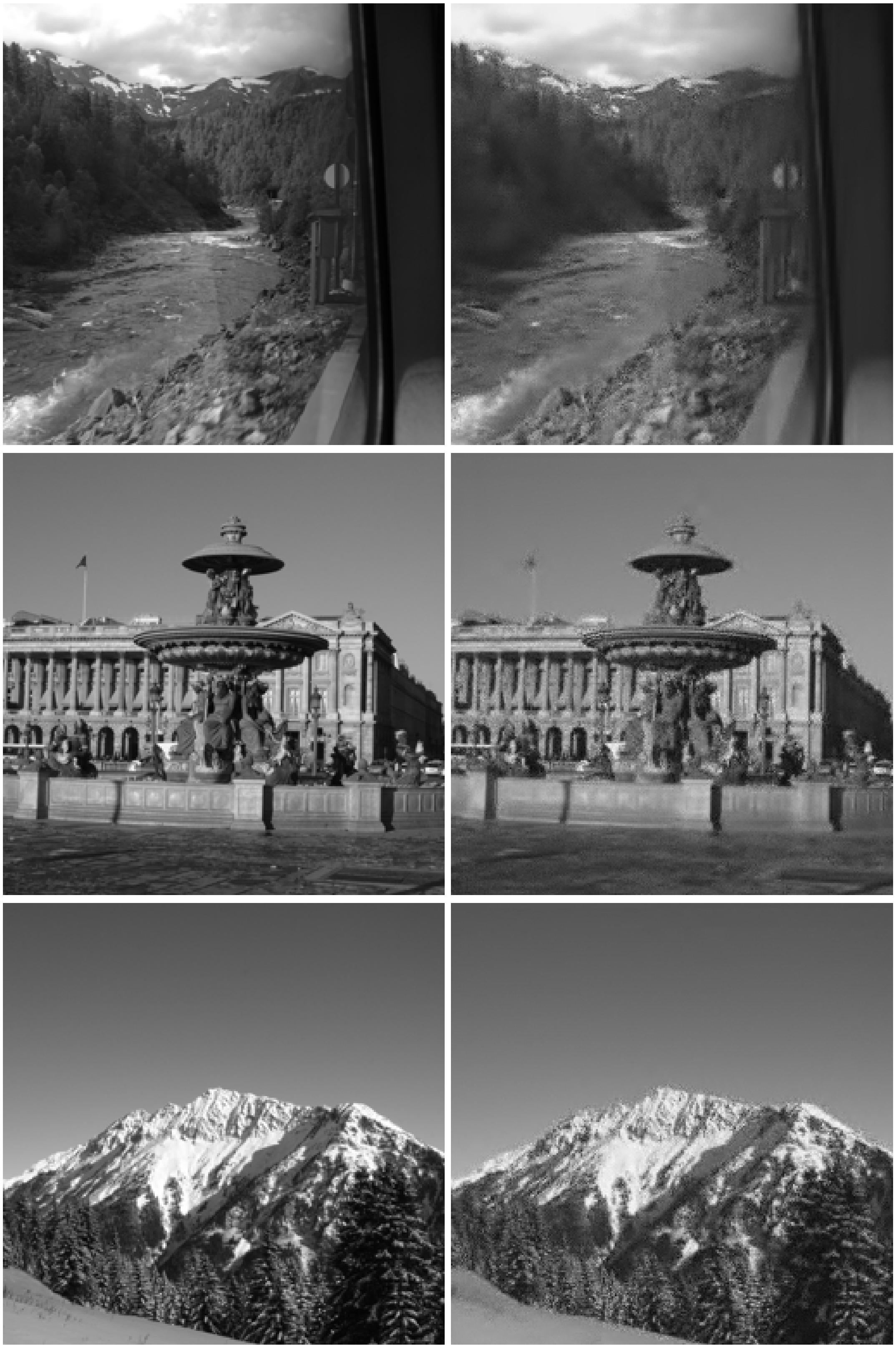


Figure 1. *Examples of images from $\mathcal{B}(\sigma)$ (right) prepared from original images (left). Top down: BOSSbase images '280.pgm', '2840.pgm', and '2814.pgm'.*

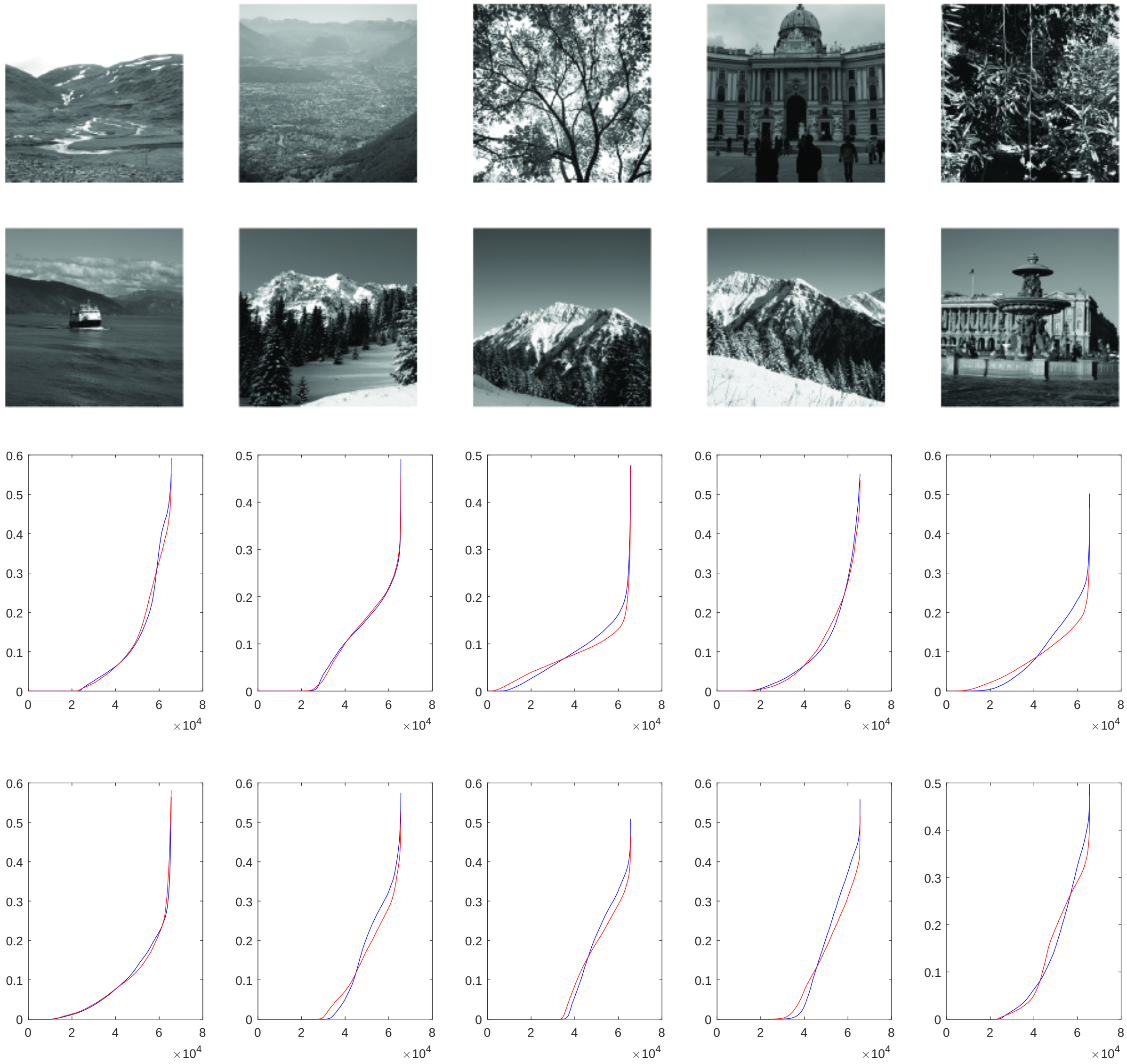


Figure 2. *Sorted change rates computed for ten images (above) from $\mathcal{B}(\sigma)$ for HILL at 0.4 bpp. Blue: change rates computed from the original images from $\mathcal{B}$, Red: from noisified covers c_{ij} from $\mathcal{B}(\sigma)$.*

guaranteed to be larger than 0.1 for $\mathcal{B}(\sigma)$ and 0.05 for $\mathcal{B}(\sigma/2)$. This topic is discussed in more detail in the next section in which we provide justification for the cover source design choices made above.

Justification

In this section, we provide a justification for the choices made when creating the dataset in the previous two sections. First, we discuss the importance of the denoising Step 2, the choice of the variance of the added Gaussian noise in Step 1, and narrowing of the dynamic range in Step 3. Finally, we study the effect of the failure to comply with the fine quantization assumption.

The denoising step in the preparation of the cover source is essential because the means μ_{ij} are assumed to be known for the optimal test. By adding the Gaussian noise to the denoised image, we force the cover complexity to be primarily in the noise component. This way, both the optimal test and the network need to deal with content complexity due to indeterminism, the added noise. Of course, the network still needs to learn a model for the (heterogeneous) denoised content.

Because the added noise mimics the content complexity of the original image (a combination of the indeterminism in the original image and texture), when the change rates β_{ij} are computed from the noisified cover c_{ij}, the change rates of the three tested content-adaptive algorithms closely match the change rates computed from the original image from $\mathcal{B}$ (see Figure 2). This justifies our choice of computing the change rates from the original image.

Note that since $p_{ij}(0) = p_{ij}(255) = 0$, the boundary values do not occur in covers from $\mathcal{B}(\sigma)$ and $\mathcal{B}(\sigma/2)$. Thus, whenever the embedding produces these "forbidden" values, the LRT Λ_{ij} becomes infinity for such pixels, arranging for a perfect detection in this case. Fortunately, due to our choice of the standard deviation $\underline{\sigma}_{ij}$ (2), this occurs with very small probability. For example, for S-UNIWARD at 0.4 bpp across the entire test set $\mathcal{B}^{(\mathrm{TST})}(\sigma)$ we saw only two such stego images.

The scaling of the dynamic range in Step 3 and adjusting the variance of the added Gaussian noise in Step 4 are necessary to avoid covers containing pixels with the boundary values 0 and 255. If this had not been done, the embedding would have to be adjusted at the boundary, which would reintroduce a dependence between a specific noisification c_{ij} and the change rates β_{ij} and complicate the derivation of a closed-form expression for the LRT. The specific choice of the narrower dynamic range $[15, 240]$ was dictated by our desire to avoid lowering the estimated variance in Eq. (2) too much as this would undermine the fine quantization assumption needed for the asymptotic result (13).

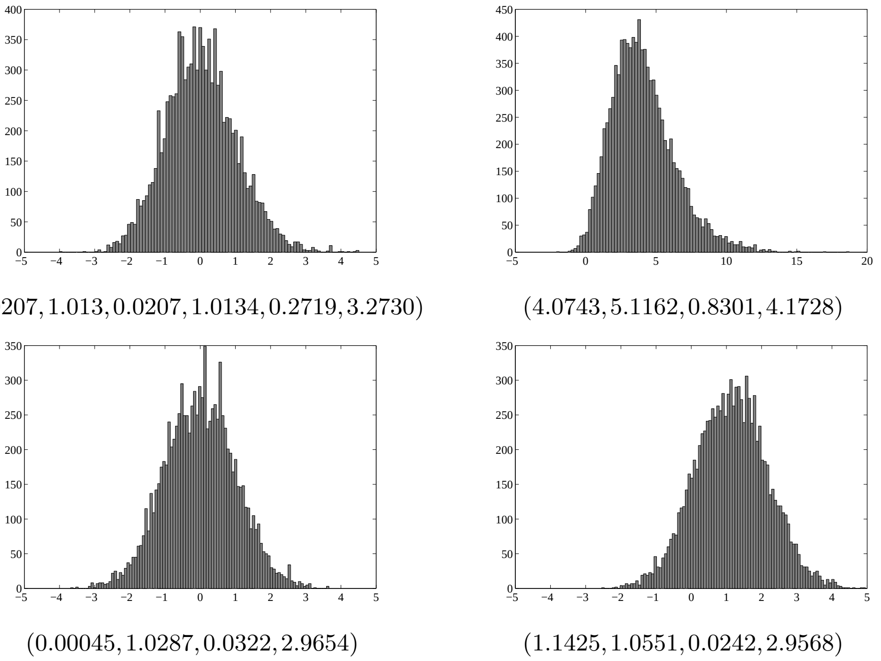
The presence of a large number of pixels with a small σ_{ij} (e.g., $\sigma_{ij} < 0.5$) may introduce large deviations from the asymptotic distribution of the log-LRT (13), especially for the alternative (stego) hypothesis. To demonstrate this, we selected BOSSbase image '280.pgm' shown in Figure 1 top, which contains many pixels with a small noise variance σ_{ij}^2 (31% of pixels have $\sigma_{ij} < 1$ and 21% $\sigma_{ij} < 1/2$), and executed the following experiment. The image was independently processed $10,000$ times and embedded as described in Section "Cover and stego sources" for $\mathcal{B}(\sigma/2)$. Figure 3 top shows the distribution of the normalized log-LRT $\Lambda^\star$ for the $10,000$ noisifications for both the cover (left) and stego (right) versions of this image with S-UNIWARD at 0.4 bpp. The bottom figures show the same distributions when additionally flooring the estimated variance returned by MiPOD by 1: $\sigma_{ij}^2 \to \max\{1, \sigma_{ij}^2\}$. Note that when not enforcing the fine quantization limit, the distribution of $\Lambda^\star$ under $\mathcal{H}_1$ fails to follow (13) – the distribution is asymmetrical with a positive skewness and a thick right tail. The distribution under $\mathcal{H}_0$ is also positively skewed and slightly leptokurtic but overall affected to a much smaller degree. When flooring the variance to enforce fine quantization, both distributions become compatible with the expected asymptotic limits (13).

To study the impact of the slightly thicker tail of the test statistic $\Lambda^\star$ under $\mathcal{H}_0$, in Figure 4 we plot the distribution of $\Lambda^\star$ under $\mathcal{H}_0$ across the entire test set $\mathcal{B}^{(\mathrm{TST})}(\sigma/2)$ (see the next section for details), whose first four moments are compatible with $\mathcal{N}(0,1)$. This indicates that the impact of the slightly thicker right tail across the entire test dataset is small. Technically, however, the distribution of $\Lambda^\star$ under $\mathcal{H}_0$ and the increased thickness of the right tail depends on the image and thus the decision threshold needs to be adjusted for each image separately.

To investigate this further, we redrew the ROC for S-UNIWARD at 0.4 bpp without assuming the asymptotic result $\Lambda^\star \sim \mathcal{N}(0,1)$ under $\mathcal{H}_0$ for all images from the test set $\mathcal{B}^{(\mathrm{TST})}(\sigma/2)$. As this experiment is extremely computationally demanding, we only investigated one stego algorithm and one payload. Referring the reader to Appendix A for details, we generated 10,000 noisifications of each image from the test set (and their stego versions), computed $\Lambda^\star$ for each noisification, fit a scaled χ^2 distribution to the samples using the method of moments, and calculated the decision threshold for a range of false alarm rates to draw the ROC without the asymptotic approximation of the LRT. The maximal difference in probability of missed detection across all false alarms between this ROC and the ROC drawn based on the asymptotic distribution (13) was smaller than 0.26%. This provides additional evidence and justification for drawing the ROC for the LRT from the asymptotic approximation.

Experiments

The source of $20,000$ cover images $\mathcal{B}(\sigma)$ (and $\mathcal{B}(\sigma/2)$) created in Section "Cover and stego sources" was randomly split into three disjoint parts, $\mathcal{B}^{(\mathrm{TRN})}(\sigma)$, $\mathcal{B}^{(\mathrm{VAL})}(\sigma)$, and $\mathcal{B}^{(\mathrm{TST})}(\sigma)$, each with $14,000$, $1,000$, and $5,000$ images, for training, validation, and testing, respectively. The following three detectors were included in our study: the optimal LRT (10), the convolutional neural network SRNet [5] as an example of a leading deep learning architecture for the spatial domain, and the maxSRMd2 [13] feature transformed using random conditioning (RC) [6] with the low-



$(0.0207, 1.013, 0.0207, 1.0134, 0.2719, 3.2730)$ $(4.0743, 5.1162, 0.8301, 4.1728)$

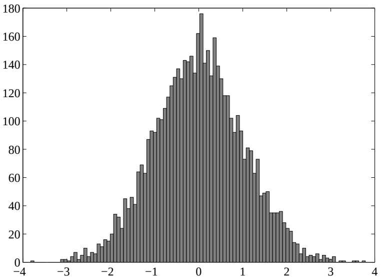
$(0.00045, 1.0287, 0.0322, 2.9654)$ $(1.1425, 1.0551, 0.0242, 2.9568)$

Figure 3. *Distribution of* $\Lambda^\star$ *and its mean, variance, skewness, and kurtosis computed from 10,000 different noisifications of BOSSbase image '280.pgm' from dataset* $\mathcal{B}(\sigma/2)$*. Left: covers, Right: stego images for S-UNIWARD* 0.4 *bpp. Top: cover and stego images as described in Section "Cover and stego sources," Bottom: when flooring the estimated variance* σ_{ij} *with 1.*

complexity linear classifier [10] representing the detection paradigm based on rich media models.

The more powerful version of the SRNet aware of the selection channel, the change rates β_{ij}, has not been used in our experiments because there is no "correct" way of incorporating the selection channel within our setup. If the network is given change rates always computed from the analyzed image, it is essentially misled because it is not given the correct change rates even for covers as β_{ij} are (and need to be as explained above) computed from the original images from $\mathcal{B}$. This was experimentally confirmed through an experiment with S-UNIWARD at 0.6 bpp – the performance of the SCA-SRNet on $\mathcal{B}(\sigma/2)$ was worse by almost 1% compared to the regular SRNet. On the other hand, giving the true change rates used for embedding (computed from images in $\mathcal{B}$) for both cover and stego images would correspond to an unrealistic situation as the steganalyst will never have access to them. Finally, giving the true change rates for covers and the change rates computed from the analyzed image for stego images is not an option either because in this case the network detects the inconsistency in the supplied selection channel and the image instead of detecting steganographic changes.

The SRNet was trained on $\mathcal{B}^{(\mathrm{TRN})}(\sigma)$ with $\mathcal{B}^{(\mathrm{VAL})}(\sigma)$ used for validation while the maxSRMd2 was trained on the union $\mathcal{B}^{(\mathrm{TRN})}(\sigma) \cup \mathcal{B}^{(\mathrm{VAL})}(\sigma)$. Because of problems with convergence of the stochastic gradient descend on $\mathcal{B}^{(\mathrm{TRN})}(\sigma)$, the SRNet was first trained on the less noisy dataset $\mathcal{B}(\sigma/2)$ from a random initialization. Then, the trained SRNet was used as a seed for training on $\mathcal{B}(\sigma)$. The LRT's performance was computed on $\mathcal{B}^{(\mathrm{TST})}(\sigma)$.



$(0.0054, 1.0228, 0.0415, 3.0069)$

Figure 4. *Normalized log-LRT* $\Lambda^\star$ *and its four moments across the test set* $\mathcal{B}^{(\mathrm{TST})}(\sigma/2)$ *for S-UNIWARD at* 0.4 *bpp.*

Figure 5 and Table 1 show the comparison in terms of the ROCs and three scalar performance descriptors – the minimum total classification error under equal class priors $P_{\mathrm{E}} = \frac{1}{2}(P_{\mathrm{FA}} + P_{\mathrm{MD}})$, the false-alarm rate at 50% detection, $P_{\mathrm{FA}}(0.5)$, and the missed-detection rate for 5% false alarm, $P_{\mathrm{MD}}(0.05)$, all on the test set.

The SRNet is better than the detector with maxSRMd2 but the amount of improvement depends on the embedding algorithm. For LSBM, the performance of the empirical detectors appears closer to the optimal LRT than for adaptive algorithms. For HILL and S-UNIWARD, the gap between the bound and the ROC of the empirical detector was "cut roughly by half" in terms of P_{E}. Both empirical detectors seem to struggle to detect WOW, which is, curiously, the least detectable algorithm in both our datasets by all three detectors.

The less noisy source $\mathcal{B}(\sigma/2)$ seems a better choice for a benchmark dataset because of the potential problems with convergence of detectors built as CNNs in the noisier $\mathcal{B}(\sigma)$. Using an even smaller variance for noisification, however, would further violate the fine quantization assumption and our ability to leverage the asymptotic distribution of the test statistic (13). This is not an insurmountable problem, though, as the ROC of the optimal detector can still be built from parametric models of the distribution of $\Lambda^\star$ for each image as explained in Appendix A. This is, however, a computationally very demanding procedure.

In Figure 6, we show additional comparisons between the LRT and SRNet for S-UNIWARD for three different payloads. The SRNet for payloads 0.2 and 0.6 bpp was trained by seeding with the detector built for 0.4 bpp.

In summary, our analysis shows that, while the SRNet is most of the time markedly better than maxSRMd2, a large space for improvement still exists especially for low false alarms.

Conclusions

With the recent progress in steganalysis due to deep learning, researchers started asking the obvious question – are such advanced detectors, which can be easily scaled up, as good detectors as they can be? If not, how much is there in terms of further possible improvement? Answering this question is, however, very difficult because the noise component of natural images is too complex to capture using tractable and estimable statistical models.

In this paper, we describe a methodology for creating a dataset of natural-looking cover and stego images with a known statistical model that allows derivation of a closed form of an optimal statistical test to establish theoretical bounds on detection performance and thus find answers to the questions posed above. We start with an existing dataset, denoise the images, and reintroduce Gaussian noise whose variance depends on the noise component of the original image as well as its local content complexity. With three content adaptive embedding schemes and LSBM, we contrast the performance of the optimal test, a deep neural network called the SRNet, and a detector built with the maxSRMd2 rich model.

The model parameters, the values of the denoised pixels and the variance of the added Gaussian noise, as well as the knowledge of the embedding change rates is given to the optimal detector, which is a likelihood ratio test with known asymptotic distribution (in large data sample and fine quantization limits). The empirical detectors thus face a rather challenging task as they need to learn the model parameters from examples.

SRNet generally offers much better detection than the detector with maxSRMd2. The amount of improvement depends on the embedding algorithms and is larger for the non-adaptive LSBM. For S-UNIWARD, the gap between the bound and the empirical detector got cut roughly by one half in terms of the minimal total detection error P_{E}.

The results reported in this paper should not be over-interpreted as it is certainly possible to prepare the dataset in other ways that may lead to different conclusions. While there still appears a rather large space for improvement, the current rapid improvement of detectors will likely further close this gap due to scaling up current architectures and increasing the size of the training sets as well as due to novel architectures and further advancements in machine learning. We finally note that while our study is limited to the spatial domain, in principle a similar approach could be taken for the JPEG domain as well.

All code used to produce the results in this paper, including the network configuration files are available from `http://dde.binghamton.edu/download/`.

Acknowledgments

The work on this paper was supported by NSF grant No. 1561446 and by DARPA under agreement number FA8750-16-2-0173. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of DARPA or the U.S. Government.

Appendix A: Drawing ROC without asymptotic approximation

We now explain how to draw the ROC of the optimal test without using the asymptotic distribution (13). This is needed when the fine quantization assumption fails to hold, i. e., when the variance of the added noise in Step 5 in Section "Cover and stego sources" is small.

Since the distribution of the normalized LRT $\Lambda^\star$ is generally different for each image, we fit a parametric model to the distribution of $\Lambda^\star$ for each image individually and then compute the decision threshold for a given false alarm rate from the model. This requires that for each image in the test set, we noisify it n times and compute n normalized LRT values $\Lambda^\star$ for the parametric fit. The thick right tail and thin left tail indicate that a good model may be the scaled chi-square density. Formally, $Y_{k,s,t} = sX_k + t$ follows the scaled chi-square distribution when $X_k \sim \chi_k^2$, the chi-square distribution with k degrees of freedom,

$$f_k(x) = \begin{cases} \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)} & x > 0 \\ 0 & x \leq 0. \end{cases} \tag{14}$$

and s and t are real parameters. The p.d.f. of $Y_{k,s,t}$ is

$$Y_{k,s,t} \sim \frac{1}{s} f_k\left(\frac{x-t}{s}\right). \tag{15}$$

Since

$$E[X_k] = k \tag{16}$$

$$Var[X_k] = 2k \tag{17}$$

$$\gamma_1 \triangleq \frac{E[(X_k - E[X_k])^3]}{(Var[X_k])^{3/2}} = \sqrt{\frac{8}{k}} \tag{18}$$

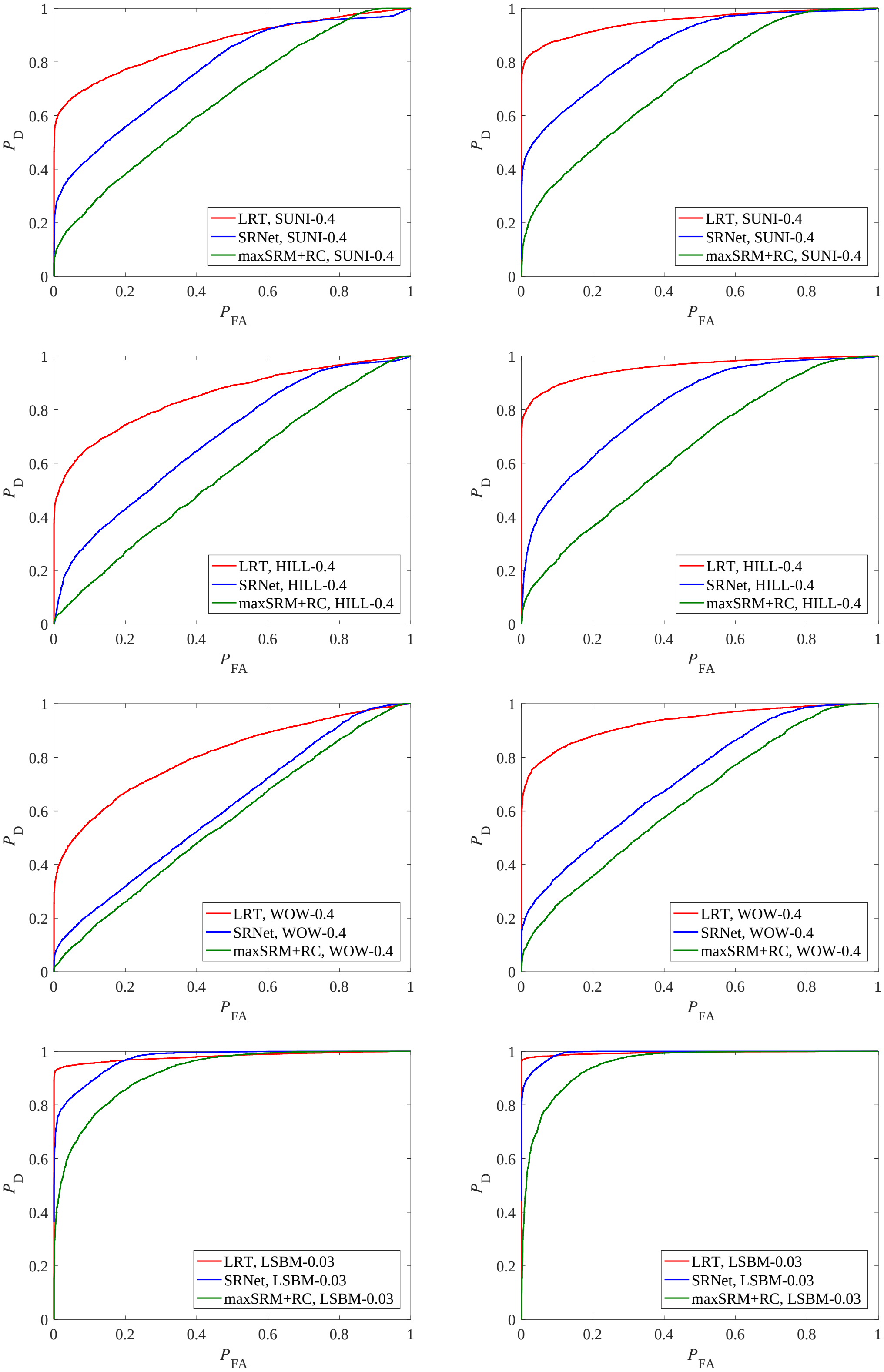


Figure 5. *ROCs for (top down) S-UNIWARD, HILL, WOW at 0.4 bpp, and LSBM for total change rate $\beta = 0.03$ for the optimal test (LRT), SRNet, and maxSRMd2 with the low-complexity linear classifier with the left and right part columns corresponding to datasets $\mathcal{B}(\sigma)$ and $\mathcal{B}(\sigma/2)$, respectively.*

	P_{E}				$P_{\mathrm{FA}}(0.5)$				$P_{\mathrm{MD}}(0.05)$			
	HILL	SUNI	WOW	LSBM	HILL	SUNI	WOW	LSBM	HILL	SUNI	WOW	LSBM
LRT	.2186	.1920	.2638	.0378	.0148	.0006	.0594	7.75×10^{-14}	.4102	.3358	.5142	.0522
SRNet	.3763	.3181	.4368	.1072	.2680	.1452	.3748	.0000	.7699	.6227	.8450	.1714
maxSRM+RC	.4572	.4005	.4596	.1692	.4194	.3092	.4204	.0182	.9140	.8120	.9076	.3620

Table 1. Performance of three empirical detectors in terms of P_{E}, $P_{\mathrm{FA}}(0.5)$, and $P_{\mathrm{MD}}(0.05)$.

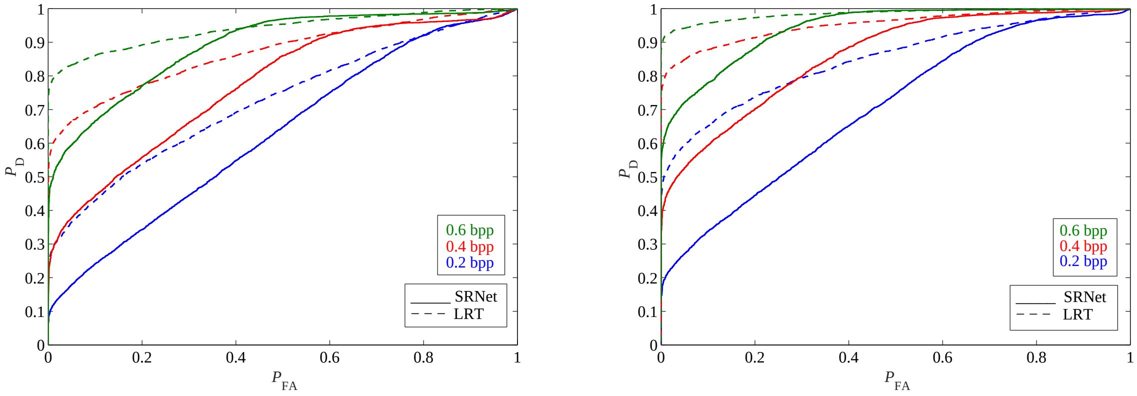


Figure 6. *Contrasting the performance of the LRT (dashed line) and SRNet (solid line) for payloads $0.2-0.6$ bpp for S-UNIWARD on $\mathcal{B}(\sigma)$ and $\mathcal{B}(\sigma/2)$.*

we have

$$E[Y_{k,s,t}] = sk + t \tag{19}$$

$$Var[Y_{k,s,t}] = 2ks^2 \tag{20}$$

$$E[(Y_{k,s,t} - E[Y_{k,s,t}])^3] = (Var[Y_{k,s,t}])^{3/2}\sqrt{\frac{8}{k}} \tag{21}$$

$$= 2\sqrt{2}k\sqrt{k}s^3\frac{\sqrt{8}}{\sqrt{k}} = 8ks^3. \tag{22}$$

A particularly simple method for estimating the parameters s,k,t from n independent realizations $y_1,\ldots,y_n$ of $Y_{k,s,t}$ (n independent noisifications of the image and evaluations of $\Lambda^\star$) is the method of moments. First, compute the three sample statistics – the sample mean $\hat{c}_1$, sample variance $\hat{c}_2$, and sample centralized but non-centered third moment $\hat{c}_3$, which are the sample quantities of the expectations on the left hand side of (19)–(21):

$$\hat{c}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{23}$$

$$\hat{c}_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{c}_1)^2 \tag{24}$$

$$\hat{c}_3 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{c}_1)^3. \tag{25}$$

From (19)–(21), the parameters can be progressively calculated :

$$\hat{s} = \frac{\hat{c}_3}{4\hat{c}_2} \tag{26}$$

$$\hat{k} = \frac{\hat{c}_2}{2\hat{s}^2} \tag{27}$$

$$\hat{t} = \hat{c}_1 - \hat{s}\hat{k}. \tag{28}$$

To determine the decision threshold τ for a given false alarm α, realize that

$$\alpha = \Pr\{Y_{\hat{k},\hat{s},\hat{t}} > \tau\} = 1 - F_{\hat{k}}\left(\frac{\tau - \hat{t}}{\hat{s}}\right) \tag{29}$$

where F_k is the c.d.f. of χ^2_k, and thus

$$\tau = \hat{t} + \hat{s}F_{\hat{k}}^{-1}(1-\alpha). \tag{30}$$

Having estimated the parameters $\hat{t}_j, \hat{s}_j$, and $\hat{k}_j$, $j = 1,\ldots,N$, for all N images in the test set, the ROC is drawn in the following manner. Let us denote the log-LRT (10) for jth stego image as $\Lambda^\star(\mathbf{s}^{(j)})$. To get a point on the ROC for false alarm α, we first compute the decision thresholds for all N images

$$\tau_j(\alpha) = \hat{\mu}_j + \hat{s}_j F_{\hat{k}_j}^{-1}(1-\alpha), \quad \forall j. \tag{31}$$

The point on the ROC curve $(\alpha, P_{\mathrm{D}}(\alpha))$ is for

$$P_{\mathrm{D}}(\alpha) = \frac{1}{N}\sum_{j=1}^{N}[\Lambda^\star(\mathbf{s}^{(j)}) > \tau_j(\alpha)]_I, \tag{32}$$

where $[.]_I$ is the Iverson bracket.

References

[1] H. Abdulrahman, M. Chaumont, P. Montesinos, and M. Baptiste. Color image steganalysis based on steerable Gaussian filters bank. In F. Perez-Gonzales, F. Cayre, and P. Bas, editors, *The 4th ACM Work-*

shop on Information Hiding and Multimedia Security, pages 109–114, Vigo, Spain, June 20–22, 2016.

[2] R. J. Anderson and F. A. P. Petitcolas. On the limits of steganography. *IEEE Journal of Selected Areas in Communication*, 16(4):474–481, 1998.

[3] P. Bas, T. Filler, and T. Pevný. Break our steganographic system – the ins and outs of organizing BOSS. In T. Filler, T. Pevný, A. Ker, and S. Craver, editors, *Information Hiding, 13th International Conference*, volume 6958 of Lecture Notes in Computer Science, pages 59–70, Prague, Czech Republic, May 18–20, 2011. Springer Berlin Heidelberg.

[4] P. Bas and T. Furon. BOWS-2. `http://bows2.ec-lille.fr`, July 2007.

[5] M. Boroumand, M. Chen, and J. Fridrich. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5):1181–1193, May 2019.

[6] M. Boroumand and J. Fridrich. Non-linear feature normalization in steganalysis. In R. Bohme and C. Pasquini, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, June 20–22, 2017.

[7] M. Boroumand and J. Fridrich. Applications of explicit non-linear feature maps in steganalysis. *IEEE Transactions on Information Forensics and Security*, 13(4):823–833, April 2018.

[8] C. Cachin. An information-theoretic model for steganography. *Information and Computation*, 192(1):41–56, July 2004.

[9] L. Chen, Y.-Q. Shi, and P. Sutthiwan. Variable multi-dimensional co-occurrence for steganalysis. In *Digital Forensics and Watermarking, 13th International Workshop, IWDW*, volume 9023, pages 559–573, Taipei, Taiwan, October 1–4 2014. Springer.

[10] R. Cogranne, V. Sedighi, T. Pevný, and J. Fridrich. Is ensemble classifier needed for steganalysis in high-dimensional feature spaces? In *IEEE International Workshop on Information Forensics and Security*, Rome, Italy, November 16–19 2015.

[11] T. Denemark, M. Boroumand, and J. Fridrich. Steganalysis features for content-adaptive JPEG steganography. *IEEE Transactions on Information Forensics and Security*, 11(8):1736–1746, August 2016.

[12] T. Denemark and J. Fridrich. Improving selection-channel-aware steganalysis features. In A. Alattar and N. D. Memon, editors, *Proceedings IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2016*, San Francisco, CA, February 14–18, 2016.

[13] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich. Selection-channel-aware rich model for steganalysis of digital images. In *IEEE International Workshop on Information Forensics and Security*, Atlanta, GA, December 3–5, 2014.

[14] T. Filler, J. Judas, and J. Fridrich. Minimizing additive distortion in steganography using syndrome-trellis codes. *IEEE Transactions on Information Forensics and Security*, 6(3):920–935, September 2011.

[15] T. Filler, A. D. Ker, and J. Fridrich. The Square Root Law of steganographic capacity for Markov covers. In N. D. Memon, E. J. Delp, P. W. Wong, and J. Dittmann, editors, *Proceedings SPIE, Electronic Imaging, Media Forensics and Security*, volume 7254, pages 08 1–11, San Jose, CA, January 18–21, 2009.

[16] A. Foi, M. Trimeche, V. Katkovnik, and K. Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, Oct. 2008.

[17] J. Fridrich and J. Kodovský. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, June 2011.

[18] M. Goljan, R. Cogranne, and J. Fridrich. Rich model for steganalysis of color images. In *Sixth IEEE International Workshop on Information Forensics and Security*, Atlanta, GA, December 3–5, 2014.

[19] G. E. Healey and R. Kondepudy. Radiometric CCD camera calibration and noise estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(3):267–276, March 1994.

[20] V. Holub and J. Fridrich. Designing steganographic distortion using directional filters. In *Fourth IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5, 2012.

[21] V. Holub and J. Fridrich. Low-complexity features for JPEG steganalysis using undecimated DCT. *IEEE Transactions on Information Forensics and Security*, 10(2):219–228, February 2015.

[22] V. Holub, J. Fridrich, and T. Denemark. Universal distortion design for steganography in an arbitrary domain. *EURASIP Journal on Information Security, Special Issue on Revised Selected Papers of the 1st ACM IH and MMS Workshop*, 2014:1, 2014.

[23] J. R. Janesick. *Scientific Charge-Coupled Devices*, volume Monograph PM83. Washington, DC: SPIE Press - The International Society for Optical Engineering, January 2001.

[24] A. D. Ker. The square root law of steganography: Bringing theory closer to practice. In R. Bohme and C. Pasquini, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, June 20–22, 2017.

[25] A. D. Ker. On the relationship between embedding costs and steganographic capacity. In *The 6th ACM Workshop on Information Hiding and Multimedia Security*, Innsbruck, Austria, June 20–22, 2018.

[26] A. D. Ker, T. Pevný, J. Kodovský, and J. Fridrich. The Square Root Law of steganographic capacity. In A. D. Ker, J. Dittmann, and J. Fridrich, editors, *Proceedings of the 10th ACM Multimedia & Security Workshop*, pages 107–116, Oxford, UK, September 22–23, 2008.

[27] J. Kodovský and J. Fridrich. Steganalysis of JPEG images using rich models. In A. Alattar, N. D. Memon, and E. J. Delp, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2012*, volume 8303, pages 0A 1–13, San Francisco, CA, January 23–26, 2012.

[28] J. Kodovský, J. Fridrich, and V. Holub. Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on Information Forensics and Security*, 7(2):432–444, 2012.

[29] B. Li, M. Wang, and J. Huang. A new cost function for spatial image steganography. In *Proceedings IEEE, International Conference on Image Processing, ICIP*, Paris, France, October 27–30, 2014.

[30] M. K. Mihcak, I. Kozintsev, K. Ramchandran, and P. Moulin. Low-complexity image denoising based on statistical modeling of wavelet coefficients. *IEEE Signal Processing Letters*, 6(12):300–303, December 1999.

[31] P. Moulin and Y. Wang. New results on steganographic capacity. In *Proceedings of the Conference on Information Sciences and Systems, CISS*, Princeton, NJ, March 17–19, 2004.

[32] Y. Qian, J. Dong, W. Wang, and T. Tan. Deep learning for steganalysis via convolutional neural networks. In A. Alattar and N. D. Memon, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2015*, volume 9409, pages 0J 1–10, San Francisco, CA, February 8–12, 2015.

[33] Y. Qian, J. Dong, W. Wang, and T. Tan. Learning and transferring representations for image steganalysis using convolutional neural network. In *IEEE International Conference on Image Processing (ICIP)*, pages 2752–2756, September 25–28, 2016.

[34] V. Sedighi, R. Cogranne, and J. Fridrich. Content-adaptive steganography by minimizing statistical detectability. *IEEE Transactions on Information Forensics and Security*, 11(2):221–234, 2016.

[35] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang. Steganalysis of adaptive JPEG steganography using 2D Gabor filters. In A. Alattar, J. Fridrich, N. Smith, and P. Comesana Alfaro, editors, *The 3rd ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec '15, Portland, OR, June 17–19, 2015.

[36] W. Tang, H. Li, W. Luo, and J. Huang. Adaptive steganalysis against WOW embedding algorithm. In S. Katzenbeisser, R. Kwitt, and A. Piva, editors, *The 2nd ACM Workshop on Information Hiding and Multimedia Security*, pages 91–96, Salzburg, Austria, June 11–13, 2014.

[37] T. H. Thai, R. Cogranne, and F. Retraint. Camera model identification based on the heteroscedastic noise model. *IEEE Transactions on Image Processing*, 23(1):250–263, Jan 2014.

[38] Y. Wang and P. Moulin. Perfectly secure steganography: Capacity, error exponents, and code constructions. *IEEE Transactions on Information Theory, Special Issue on Security*, 55(6):2706–2722, June 2008.

[39] G. Xu. Deep convolutional neural network to detect J-UNIWARD. In R. Bohme and C. Pasquini, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, June 20–22, 2017.

[40] G. Xu, H.-Z. Wu, and Y. Q. Shi. Ensemble of CNNs for steganalysis: An empirical study. In F. Perez-Gonzales, F. Cayre, and P. Bas, editors, *The 4th ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec '16, pages 5–10, Vigo, Spain, June 20–22, 2016.

[41] G. Xu, H. Z. Wu, and Y. Q. Shi. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, 23(5):708–712, May 2016.

[42] J. Yang, K. Liu, X. Kang, E. Wong, and Y. Shi. Steganalysis based on awareness of selection-channel and deep learning. In *Internartional Workshop on Digital Forensics and Watermarking*, volume 10431 of *LNCS*, pages 263–272, 2017.

[43] J. Yang, Y.-Q. Shi, E.K. Wong, and X. Kang. JPEG steganalysis based on densenet. *CoRR*, abs/1711.09335, 2017.

[44] J. Ye, J. Ni, and Y. Yi. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11):2545–2557, November 2017.

[45] M. Yedroudj, F. Comby, and M. Chaumont. Yedroudj-net: An efficient CNN for spatial steganalysis. Alberta, Canada, April 15–20, 2018.

[46] J. Zeng, S. Tan, B. Li, and J. Huang. Large-scale JPEG image steganalysis using hybrid deep-learning framework. *IEEE Transactions on Information Forensics and Security*, 13(5):1200–1214, 2018.

Author Biography

Mehdi Boroumand received the B.S. degree in electrical engineering from the K. N. Toosi University of Technology, Iran, in 2004, and the M.S. degree in electrical engineering from the Sahand University of Technology, Iran, in 2007. He is currently pursuing the Ph.D. degree in electrical engineering with Binghamton University. His areas of research interest include digital image steganalysis and steganography, digital image forensics, image processing and computer vison, and machine learning.

Jessica Fridrich is Distinguished Professor of Electrical and Computer Engineering at Binghamton University. She received her PhD in Systems Science from Binghamton University in 1995 and MS in Applied Mathematics from Czech Technical University in Prague in 1987. Her main interests are in steganography, steganalysis, and digital image forensics. Since 1995, she has received 20 research grants totaling over $12 mil that lead to more than 200 papers and 7 US patents.

Remi Cogranne is an Associate Professor at Troyes University of Technology (UTT), France, since 2013. He has regularly been a visiting scholar at Binghamton University between 2014 and 2017. He received his PhD in Systems Safety and Optimization from UTT in 2011, since on, his research focuses on hypothesis testing applied to image forensics, steganalysis, steganography, and computer network anomaly detection, which lead to more than 60 papers and 3 International patents.