IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE.Please read the form carefully and keep a copy for your files.

Author Online Use

6. Personal Servers. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore.Authors shall not post the final, published versions of their papers.

7. Classroom or Internal Training Use. An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the authors personal web site or the servers of the authors institution or company in connection with the authors teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.

RELIABLE DETECTION OF HIDDEN INFORMATION BASED ON A NON-LINEAR LOCAL MODEL

Rémi Cogranne, Cathel Zitzmann, Lionel Fillatre, Igor Nikiforov, Florent Retraint and Philippe Cornu

ICD - LM2S - Université de Technologie de Troyes - UMR STMR CNRS
12, rue Marie Curie - B.P. 2060 - 10010 Troyes cedex - France
E-mail : name.surname@utt.fr

ABSTRACT

This paper investigates the reliable detection of information hidden in natural images. It is aimed to design a test with analytically predictable probabilities of error. To this end, the problem of hidden information detection is cast in the framework of hypothesis testing. The optimal test solving the decision problem of steganalysis requires image parameters which are not available in practice. To design a feasible test, a non-linear locally-adapted model of natural images is proposed. This model is linearized to allow an efficient and simple estimation of image parameters which leads to the design of an almost optimal test. Numerical results on a large number of natural images show the relevance of the theoretical findings.

Index Terms— Statistical hypothesis test, LSB steganalysis, piecewise non-linear approximation.

1. INTRODUCTION

It is a crucial and useful challenge for security forces to reliably detect in a huge set of media (image, audio, or video) those which contain hidden information (like a text). In such an operational context, an efficient detector might not be enough. Indeed, the most important challenge is to get a detection algorithm with analytically predictable probabilities of false alarm and non detection. Moreover, this detection scheme should be immediately applicable without any training phase like needed for supervised learning methods.

In this paper, it is assumed that the embedding scheme is a priori unknown but belongs to the commonly used family of LSB replacement steganographic schemes. Certainly, such steganographic algorithms are not extremely efficient but they are simple, popular, downloadable on the Internet and within the reach of anyone. Moreover, the capacity to detect very sophisticated but seldom used stegosystems is not very important in the framework of the above mentioned scenario.

The recently proposed *ad hoc* Weighted Stego-image (WS) detector (and its improvements) is certainly very interesting and efficient [1, 2, 3] but it has been designed empirically. Thus, its performances are not theoretically established. They are only evaluated by using large databases of media and numerical simulations. An alternative approach is to design a detector with clearly established theoretical properties by using the hypothesis testing theory with model of cover media. The first step in this direction has been done in [4] with an independent and identically distributed (i.i.d) samples model. In the present paper, the direction started in [4] is extended to take into account the specific content of natural images, especially when the images are noisy and the embedding rate is small.

Hence, the current paper proposes 1) to locally model the content of a cover image by describing the optical system which gives birth to a natural image, 2) to exploit, as simply as possible, this model of natural image to design an almost optimal test, 3) to establish the theoretical performances of the proposed test and 4) to numerically show the relevance of the proposed detection scheme in comparison with other steganalysis methods.

The paper is organized as follows. Section 2 starts with the problem statement. Section 3 presents the cover image model. Section 4 proposes the model-based statistical test and establishes its theoretical performances. Section 5 studies the numerical performances of the proposed test. Finally, Section 6 concludes this paper.

2. DETECTION PROBLEM STATEMENT

Let $\mathbf{c}=\{c_m\}_{m=1}^{M}$ be a vector representing a natural cover-image of $M=M_x\times M_y$ grayscale level pixels. The set of grayscale levels is denoted $\mathcal{Y}=\{0,\dots,2^b-1\}$. A cover pixel c_m results from the quantization

$$c_m = Q_1[y_m], \tag{1}$$

where $y_m \in \mathbb{R}_+$ denotes the raw pixel intensity recorded by the digital camera and $Q_1[y_m] = \lfloor y_m \rfloor$ is the operation of uniform quantization (integer part of y_i). The recorded pixel value y_m can be decomposed as [5]

$$y_m = \theta_m + \xi_m \tag{2}$$

This work is supported by the French National Agency (ANR) through ANR-CSOSG program (project ANR-07-SECU-004).

where θ_m is the mathematical expectation of y_m and ξ_m is a random variable representing all the noise corrupting the signal during acquisition (see Section 3). The ξ_m's are independent Gaussian random variables satisfying $\xi_m \sim \mathcal{N}(0, \sigma_m^2)$ where σ_m^2 is assumed to be known. The mean vector $\boldsymbol{\theta} = \{\theta_m\}_{m=1}^M$ is assumed to belong to a compact set Θ.

The probability mass function (pmf) of c_m is denoted:

$$Q_{Q_1}(\theta_m) = \{q_0(\theta_m), \dots, q_{2^b-1}(\theta_m)\}$$

where, for all $k \in \mathcal{Y}$, $k \neq 0$, $k \neq 2^q-1$, one get

$$q_k(\theta_m) = \frac{1}{\sigma_m} \int_{(k-\frac{1}{2})}^{(k+\frac{1}{2})} \varphi\left(\frac{x-\theta_m}{\sigma_m}\right) dx \qquad (3)$$

and $\varphi(x)$ is defined in (6). In this paper, quantizer saturation effects which arrise for $k = 0$ and $k = 2^q-1$ are neglected. Let the insertion rate R be defined as the number of hidden bits per pixel. A short calculation [4] shows that the pmf of c_m after insertion with embedding rate R is given by $Q_{Q_1}^R(\theta_m) = \{q_0^R(\theta_m), \cdots, q_{2^b-1}^R(\theta_m)\}$ where

$$\forall k \in \mathcal{Y},\ q_k^R(\theta_m) = \left(1-\frac{R}{2}\right) q_k(\theta_m) + \frac{R}{2} q_{(k-\overline{k})}(\theta_m) \quad (4)$$

and $\overline{k}$ indicates the integer k with LSB flipped [1], i.e., $\overline{k} = k + 1 - 2\,\mathrm{lsb}(k)$ where $\mathrm{lsb}(k)$ is the LSB of k.

Let $\mathbf{z}=\{z_m\}_{m=1}^M$ denotes an inspected signal, which is either a cover or a stego-signal. Two situations may occur: $\mathcal{H}_0$={ signal is steganography-free } and $\mathcal{H}_1$={ signal contains hidden bits }. Hence, the hypothesis testing problem of steganalysis consists of choosing between:

$$\begin{cases} \mathcal{H}_0=\{\mathbf{z}_m \sim Q_{Q_1}(\theta_m), \forall m=1,\dots,M\} \\ \mathcal{H}_1=\{\mathbf{z}_m \sim Q_{Q_1}^R(\theta_m), \forall m=1,\dots,M, \forall 0<R\leq 1\}. \end{cases} \qquad (5)$$

The goal is to find a test $\delta : \mathcal{Y}^M \mapsto \{\mathcal{H}_0; \mathcal{H}_1\}$ which accepts hypothesis $\mathcal{H}_i$ if $\delta(\mathbf{z}) = \mathcal{H}_i$ (see details in [6]). Let

$$\mathcal{K}_\alpha = \left\{ \delta : \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{P}_{\boldsymbol{\theta}}(\delta(\mathbf{z}) = \mathcal{H}_1) \leq \alpha \right\}$$

be the class of tests with an upper-bounded false alarm probability α. Here $\mathbb{P}_{\boldsymbol{\theta}}(A)$ stands for the probability of the event A when z_m is generated by $Q_{Q_1}(\theta_m)$ for all m. The power function β_δ is the probability of hidden bits detection:

$$\beta_\delta(\boldsymbol{\theta}, R) = \mathbb{P}_{\boldsymbol{\theta},R}(\delta(\mathbf{z}) = \mathcal{H}_1).$$

where $\mathbb{P}_{\boldsymbol{\theta},R}(A)$ stands for the probability of the event A when z_m is generated by $Q_{Q_1}^R(\theta_m)$ for all m.

According to (5), it is crucial to know the M parameters θ_m. In practice, it is necessary to estimate them. Since there are only M pixels, a special attention must be paid to reduce their number by exploiting some redundancies existing between them. Such a parsimonious model is described in the following section.

3. MODEL OF NATURAL COVER-IMAGES

Seeking simplicity, this paper deals with the cover image $\mathbf{c}$ as a set of K statistically independent quantized signals [5] of $N > 0$ samples $\mathbf{c}_k=(c_{k,1}, \dots, c_{k,N})^T$. For example, a digital raw image can be viewed as a set of signals $\mathbf{c}_k$ composed of N pixels which are extracted from image lines. Each signal $\mathbf{c}_k$ is acquired with a digital device and thus, is subject to quantization, sampling and filtering processes with an Impulse Response Function (IRF) $h(x)$ applied to the corresponding source signal $s_k(x)$. It is assumed that $s_k(x)$ is defined on the interval $\mathcal{X}_k = [a_k, a_{k+1}] \subset \mathbb{R}$. The union of $\mathcal{X}_k$ describes the support of the source signal for the complete cover signal. The literature proposes a wide range of IRF models [7] but this paper is restricted to the study of the Gaussian IRF $h(x)$:

$$h(x) = \frac{1}{\varsigma}\varphi\left(\frac{x}{\varsigma}\right) \quad \text{where} \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad (6)$$

and $\varsigma > 0$ is the smoothing parameter ($\varsigma > 1$ in practice). The source signal $s_k(x)$ is assumed to be piecewise continuous and thus, for all $x \in \mathcal{X}_k$, it can be rewritten as

$$s_k(x) = \sum_{j=0}^{p-1} q_{k,j}(x-x_k)^j + \sum_{d=1}^{r_k} u_{k,d}\mathbf{1}(x-t_{k,d}), \qquad (7)$$

where $x_k = (a_k + a_{k+1})/2$ is the center of $\mathcal{X}_k$, r_k is the number of discontinuities in $\mathcal{X}_k$ and $\mathbf{1}(\cdot)$ is the unitary step function. The parameters $u_{k,d}$ and $t_{k,d}$ characterize the intensity and the location of the d-th discontinuity in the signal $s_k(x)$ over $\mathcal{X}_k$. The continuous component of $s_k(x)$ is approximated by a p-order polynomial function. Let $\theta_k(x)$ denotes the source signal $s_k(x)$ filtered by $h(x)$. Few algebra shows that

$$\theta_k(x) = \sum_{j=0}^{p-1} s_{k,j}(x-x_k)^j + \sum_{d=1}^{r_k} u_{k,d} f(x; \eta_{k,d}) \qquad (8)$$

where $\eta_{k,d} = (t_{k,d}, \varsigma)$, the coefficients $s_{k,j}$ depend on $q_{k,j}$'s $f(x; \eta_{k,d}) = \Phi\left(\frac{x-t_{k,d}}{\varsigma}\right)$ and $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$.

Consider K sets of N sampling points $x_{k,1},\dots,x_{k,N}$ such that for all $k \in \{1,\dots,K\}$ and all $i \in \{1,\dots,N\}$, sampling step $x_{k,n+1}-x_{k,n}$ is constant and $x_{k,n} \in \mathcal{X}_k$. Hence, neglecting censoring effects, one get $c_{k,n}=Q_1[\theta_{k,n}+\xi_{k,n}]$ where $\theta_{k,n} = \theta_k(x_{k,n})$. In matrix form, one get

$$\boldsymbol{\theta}_k = \mathbf{H}\mathbf{s}_k + \mathbf{F}(\boldsymbol{\eta}_k)\mathbf{u}_k \qquad (9)$$

where the notations for vectors $\mathbf{s}_k$, $\boldsymbol{\eta}_k$, $\mathbf{u}_k$ and matrices $\mathbf{H}$ and $\mathbf{F}(\boldsymbol{\eta}_k)$ are straightforward and

$$\mathbf{c}_k = Q_1[\boldsymbol{\theta}_k + \boldsymbol{\xi}_k] \qquad (10)$$

where $Q_1[\cdot]$ is applied to each component of $\boldsymbol{\theta}_k + \boldsymbol{\xi}_k$.

4. ALMOST OPTIMAL DETECTION

In [8], an optimal Likelihood Ratio Test (LRT) has been designed to solve the problem (5). In this section, it is proposed to use this LRT by replacing the parameter $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$ by an estimate $\widehat{\boldsymbol{\theta}}$ based on the model (10). Two main difficulties are addressed: 1) to deal with the non-linearity of (9) with respect to $\boldsymbol{\eta}_k$ and 2) to bound the loss of optimality of this algorithm.

Seeking simplicity, it is assumed that each segment $\mathcal{X}_k$ has at most one discontinuity and that an estimate $\widehat{\boldsymbol{\eta}}_k=(\hat{t}_{k,1}, \widehat{\varsigma})^T$ is available for each discontinuity (if present) such that $\|\boldsymbol{\eta}_k - \widehat{\boldsymbol{\eta}}_k\|_1 \leq \vartheta$ where ϑ is a small constant. The literature proposes many methods giving such estimates (see for example [7]). Adapting the methodology from [9], the non-linearity is treated by writing (9) :

$$\boldsymbol{\theta}_k = \mathbf{G}(\widehat{\boldsymbol{\eta}}_k)\mathbf{v}_k + o(\vartheta^2) \text{ with } \mathbf{G}(\widehat{\boldsymbol{\eta}}_k) = \Big(\mathbf{H} \,|\, \mathbf{F}(\widehat{\boldsymbol{\eta}}_k) \,|\, \dot{\mathbf{F}}(\widehat{\boldsymbol{\eta}}_k)\Big),$$

$\dot{\mathbf{F}}(\boldsymbol{\eta}_k)$ is the jacobian $(N \times 2)$ matrix of $F(\boldsymbol{\eta}_k)$ and $\mathbf{v}_k=(\mathbf{s}_k, u_{k,1}, u_{k,1}(\boldsymbol{\eta}_k - \widehat{\boldsymbol{\eta}}_k))^T$.

Assuming that the noise variance is constant in each segment and that the quantization has negligible effects on the estimation, $\boldsymbol{\theta}_k$ can be estimated by using the linear estimate:

$$\widehat{\boldsymbol{\theta}}_k = \mathbf{G}(\widehat{\boldsymbol{\eta}}_k) \left(\mathbf{G}(\widehat{\boldsymbol{\eta}}_k)^T \mathbf{G}(\widehat{\boldsymbol{\eta}}_k)\right)^{-1} \mathbf{G}(\widehat{\boldsymbol{\eta}}_k)^T \mathbf{z}_k. \tag{11}$$

It is assumed that the presence of hidden information $\mathbf{z}_k$ does not modify significantly the estimate $\widehat{\boldsymbol{\theta}}_k$. By dividing $\mathbf{z}$ into K segments, as detailed for the cover $\mathbf{c}$, and by calculating the estimate $\widehat{\boldsymbol{\theta}}_k$ for all $\mathbf{z}_k$, one easily get $\widehat{\theta}_m$ from (11) and:

$$w_m = \frac{\bar{\sigma}}{\sigma_m^2 \sqrt{K(N-p-3r_k)}} \quad \text{with} \quad \frac{1}{\bar{\sigma}^2} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{\sigma_m^2}$$

for all pixels. Once these two quantities are calculated, then the following test can be used to detect hidden information.

Let $r = \sum_{k=1}^{K} r_k$ be the total number of discontinuities. A short algebra shows that the LRT with plugged estimates $\widehat{\theta}_m$, denoted $\widehat{\delta}(\mathbf{z})$, is defined by (see details in [8]):

$$\widehat{\delta}(\mathbf{z}) = \begin{cases} \mathcal{H}_0 & \text{if } \widehat{\Lambda}(\mathbf{z}) < \tau_\alpha, \\ \mathcal{H}_1 & \text{if } \widehat{\Lambda}(\mathbf{z}) \geq \tau_\alpha, \end{cases} \tag{12}$$

$$\text{where } \widehat{\Lambda}(\mathbf{z}) = \sum_{m=1}^{M} w_m (z_m - \overline{z}_m)(z_m - \widehat{\theta}_m). \tag{13}$$

The following proposition gives a good approximation of the power function $\widehat{\beta}(\boldsymbol{\theta}, R)$ of the test $\widehat{\delta}(\mathbf{z})$.

Proposition 1. *Choosing* $\tau_\alpha = \Phi^{-1}(1-\alpha)\sqrt{1+b_{\max}}$, *it follows that* $\widehat{\delta}(\mathbf{z}) \in \mathcal{K}_\alpha$ *and*

$$1 - \Phi\left(\frac{\tau_\alpha - \frac{R}{2\bar{\sigma}}\sqrt{\kappa}}{1+b_{\max}}\right) \leq \widehat{\beta}(\boldsymbol{\theta}, R) \leq 1 - \Phi\left(\tau_\alpha - \frac{R}{2\bar{\sigma}}\sqrt{\kappa}\right)$$

whereb is an bias satisfying $0 \leq b \leq \frac{\varepsilon}{\bar{\sigma}^2(N-p-3)} \stackrel{def.}{=} b_{\max}$ *with* ε *a known (little) positive constant and* $\kappa = K(N-p)-3r$.

In [8], it is established that the power function of the LRT (when $\boldsymbol{\theta}$ is perfectly known) is asymptotically (as the number of pixels grows to infinity) given by:

$$\beta_{\max}(R) = 1 - \Phi\left(\tau_\alpha - \frac{R\sqrt{M}}{2\bar{\sigma}}\right). \tag{14}$$

The comparison between the theoretical upper-bound $\beta_{\max}(R)$ and $\widehat{\beta}(\boldsymbol{\theta}, R)$ shows that the loss of optimality of the later is due to: 1) the reduction of the number of "free parameters" from $M=K \cdot N$ to $\kappa=K(N-p)-3r$ and 2) the unknown bias $b_{\max}$ which is due to linearization of $\mathbf{F}(\boldsymbol{\eta}_k)$ around the estimation values of discontinuity parameter $\widehat{\boldsymbol{\eta}}_k$. Hence, provided that r and $b_{\max}$ are sufficiently small, the test $\widehat{\delta}$ is almost optimal. Values r and ϑ are arbitrary bounded to analytically calculate the power function $\widehat{\beta}(\boldsymbol{\theta}, R)$.

The test $\widehat{\delta}(\mathbf{z})$ is quite similar to the Weighted Stego-image (WS) detector initially proposed by [1] to estimate the payload size and deeply studied by [2]. Hence this paper shows that, under some assumptions, the WS detector coincides with an optimal statistical test provided that both the weights and the estimates of the cover pixels are conveniently chosen.

5. NUMERICAL RESULTS AND COMPARISONS

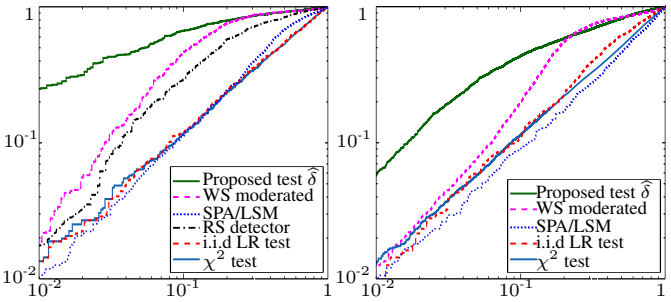
Potentially, there are a large numbers of steganalyzers we can use for comparison [3]. In fact, many of these detectors belongs to the class of *structural detectors* [2] like RS, SPA, Couples/ML. For the purpose of comparison we selected only the two leading competitors, RS and SPA/LSM. In the literature, only few tests rely on the hypothesis testing theory: the "i.i.d LR test" from [4], the χ^2 test.The comparison with these test are interesting as they cast the steganalysis problem in the frame of statistical hypothesis test as the present work.
Hence, it is proposed to compare the LSB followings replacement detectors: the proposed test $\widehat{\delta}$, the Dabeer's test from [4], the χ^2 test, the RS detector (with original mask, see [3]) and the WS [1] with the filter proposed in [2].

For a large scale comparison the 1338 image of UCID[1] and 9000 images from BOSSbase[2] were used with embedding rate $R = 0.05$.

On Figs. 1, the WS have a similar power than the proposed test $\widehat{\delta}$ for high false alarm rate, typically $\alpha \approx 0.2$. In an operational context of digital forensics such high false alarm rate largely unacceptable ; a constraint of $\alpha = 10^{-5}$ is much more realistic. For low false alarm rate, Figs. 1 show that the proposed test outperforms the five other detectors. Moreover, the most serious challenger, the WS, performs better than the others detectors in such conditions. To understand this drastic

[1] http://vision.cs.aston.ac.uk/datasets/UCID/

[2] BOSSBases (v0.93) http://boss.gipsa-lab.grenoble-inp.fr/



(a) ROC Curves for UCID images (b) ROC Curves for BOSS images

Fig. 1: ROC curves : power β as a function of α (log scale).

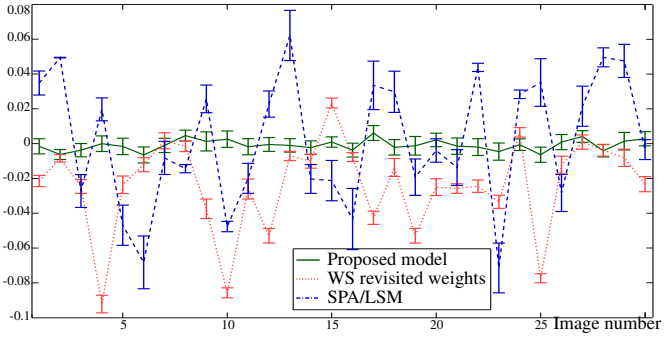
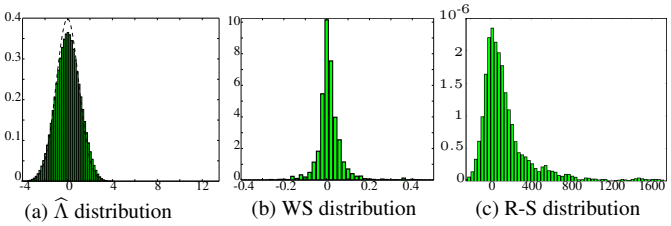


Fig. 3: Monte-Carlo analysis (Mean and standard deviation) of textured images.

loss of power exhibited on Figs. 1 for low false alarm rate α, a thorough comparison of detectors results is necessary. Figs. 2 show the empirical distribution of decision statistics obtained on UCID images for three test: the $\widehat{\Lambda}$, the WS and the RS.

Fig. 2 highlights the importance of the image content model. The WS detector relies on a basic autoregressive model which works efficiently for most images but fails for few. Hence, the distributions of WS residuals exhibit outlier values or heavy tails under both hypothesis of cover or stego images. These outliers values avoid to warrant any false alarm constraint. On the contrary, the proposed model of natural images permits a reliable estimation of the cover ; the decision statistics follow the theoretically expected distribution (dashed line).

Highly textured images are typically difficult to analyze without an accurate model of image content, as shown in Fig. 3. Thirty highly textured images have been analyzed 1000 times by adding a zero-mean Gaussian stationary noise with standard deviation σ=0.5. Fig. 3 shows the mean and standard deviation of the detection statistics. To allow a meaning full comparison of the detection statistics, note that the results have been normalized so that all detectors has the same theoretical mean R, which is set to $R = 0$ for this experimentation. Fig. 3 shows that the variance of WS and proposed decision function $\widehat{\Lambda}$ does not change much whereas the variance of the SPA/LSM is much greater. On the contrary, the mean of each image is not kept under control without an accurate image model ; this later prevent a reliable decision.



(a) $\widehat{\Lambda}$ distribution (b) WS distribution (c) R-S distribution

Fig. 2: Distribution of decision statistics under $\mathcal{H}_0$ for BOSS base.

6. CONCLUSIONS

This paper made a first step to fill the gap between physical model of cover-image and steganalysis. An almost optimal test, which is based on a local non-linear parametric model of natural images, has been proposed in Proposition 1. The proposed test permits to analytically warrant a prescribed false alarm constraint which, together with the accurate image model, allows a reliable steganalysis.

7. REFERENCES

[1] J. Fridrich and M. Goljan, "On estimation of secret message length in LSB steganography in spatial domain," in *Proc. SPIE 5306*, 2004, pp. 23–34.

[2] A. D. Ker and R. Böhme, "Revisiting weighted stego-image steganalysis," in *Proc. SPIE*, 2008, pp. 501–517.

[3] R. Böhme, *Advanced Statistical Steganalysis*, 1st ed. Springer Publishing Company, Incorporated, 2010.

[4] O. Dabeer, & al., "Detection of hiding in the least significant bit," *Signal Processing, IEEE Transactions on*, vol. 52, no. 10, pp. 3046 – 3058, oct. 2004.

[5] A. Foi, & al., "Practical poissonian-gaussian noise modeling and fitting for single-image raw-data," *Image Processing, IEEE Trans*, vol. 17, no. 10, Oct. 2008.

[6] E. Lehman, *Testing Statistical Hypotheses, Second Edition*. Chapman & Hall, 1986.

[7] C. Bruni,& al., "Identification of discontinuities in blurred noisy signals," *Circuit and systems-I, IEEE Trans. on*, vol. 44, no. 5, pp. 422 – 433, May 1997.

[8] C. Zitzmann, R. Cogranne, F. Retraint, I. Nikiforov, L. Fillatre, and P. Cornu, *Hypothesis testing by using quantized observations*, in SSP 2011.

[9] G. Seber and C. Wild, *Nonlinear Regression*. Wiley, 1989.