CAPACITY-SECURITY ANALYSIS OF DATA HIDING TECHNOLOGIES

S. Voloshynovskiy and T. Pun

University of Geneva, 24 rue Général-Dufour, CH-1211 Geneva 4, Switzerland

ABSTRACT

In this paper we consider the problem of joint capacity-security analysis of data hiding technologies from the communications point of view. First, we formulate data hiding as an optimal encoding problem for different operational regimes, that include both robust digital watermarking and steganography. This provides the corresponding estimation of the hidden data statistics, as well as of the rates approaching embedding capacity. Secondly, we formulate the problem of blind stochastic hidden data detection based on the developed watermark statistics. Finally, we estimate the error of watermark detection and the variance of the watermark estimation that determine the system security.

1. INTRODUCTION

The commonly accepted requirements for data hiding technologies such as digital watermarking, authentication, tamper proofing, self-recovering watermarks, and document security are perceptual invisibility, capacity, and robustness to some types of attacks. However, it is not easy to simultaneously satisfy all of the above contradictory requirements. Therefore, a number of practical data hiding systems either reduce the requirements or completely neglect some of them. Another very important requirement both for watermarking and for steganography systems is the *statistical* invisibility of the hidden data. This requirement is often overlooked by watermarking developers and may have an important impact on the reliability of the whole system in general. Therefore, we analyze in this paper a possibility of blind or unauthorized detection of hidden data assuming the behavior of a common data hider. Our assumption is based on the desire of the data hider to provide the maximum rate of reliable communications under certain attacks. Moreover, we consider in this paper both low-WNR (watermark-to-noise) regime of data encoding typical for robust watermarking and high-WNR (steganography) regime. The goal of unauthorized detection of hidden data in robust watermarking applications is different from that in steganography, since in the later the simple fact of detecting secret communications is considered to be enough to break the protocol. In robust digital watermarking, the unauthorized detection can be used for three purposes. First, it can provide protocol assistance to robust watermarking. Secondly, it can provide a fair evaluation of the robustness and security of a given watermarking technology during benchmarking. Thirdly, it can identify which watermarking technology is being used and help select an appropriate attacking strategy. Therefore, the security of data hiding system greatly depends on the ability of the attacker to perform an unauthorized blind detection of hidden data.

2. DATA HIDING GAMES

The unauthorized detection of hidden data strictly depends on the stochastic model of the watermark relatively to the stochastic model of the host (cover) data. The particular stochastic models of watermarks depend on four main factors: the statistics of the encrypted and encoded message, the projection function, the perceptual mask, and the modulation or embedding function. We consider here only modulation schemes with appropriate encoding that can approach the channel capacity, and their corresponding embedding strategies. For simplicity we will assume additive linear watermarking:

$$y[k] = x[k] + w[k] \quad (1)$$

where $y[k]$ is the stego image, $x[k]$ the cover image and $w[k]$ is the watermark. The p.d.f. of the stego image is $p_{\mathbf{y}}(y) = p_{\mathbf{x}}(x) * p_{\mathbf{w}}(w)$. Therefore, the goal of the data hider from the security point of view is to ensure the stochastic invisibility of the watermark in terms of an information-theoretic measure (such as the the *Kullback-Leiber distance* (KLD)), resulting in the condition $p_{\mathbf{y}}(y) \approx p_{\mathbf{x}}(x)$. This means that the influence of the watermark embedding on the smoothing of the stego image p.d.f. should be minimized. This can be achieved when $p_{\mathbf{w}}(w)$ is a d-like function with its origin near zero, when $p_{\mathbf{x}}(x)$ is considerably wider than $p_{\mathbf{w}}(w)$ or when appropriate mapping of resulted $p_{\mathbf{y}}(y)$ is performed.

2.1. Data hider strategy

The data hider strategy from the communication perspective consists of two tasks:

- to approach channel capacity for given attacks, keeping image distortions in acceptable ranges;
- to stay secure (undetectable).

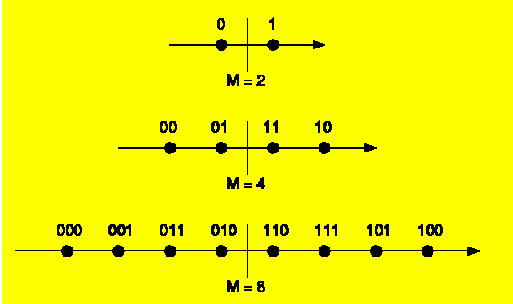

Figure 1. M-PAM constellations.

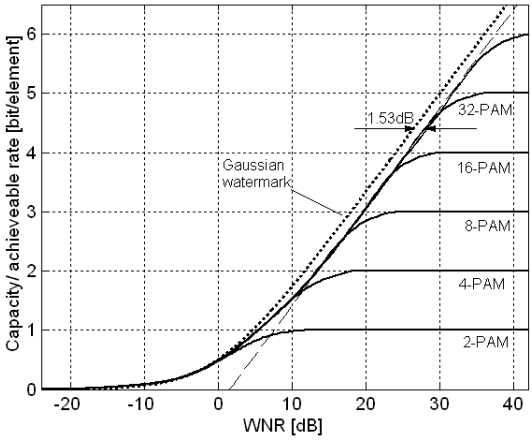

Figure 2. Capacity of the ideal AWGN channel with Gaussian watermark and with equiprobable M-PAM watermark.

We consider the link between these tasks for the simple case of an additive white Gaussian noise (AWGN) channel. The capacity of this channel is determined as:

$$C = \frac{1}{2}\log_2\left(1+\frac{s_w^2}{s_n^2}\right) \qquad (2)$$

where the watermark is assumed to be Gaussian $\mathbf{w} \sim N\left(\mathbf{0}, s_w^2\mathbf{I}\right)$ and $\mathbf{WNR} = 10\log_{10} s_w^2/s_n^2$ is the watermark-to-noise ratio. The capacity of the AWGN channel is shown in Figure 2 in dotted line. To justify the data hider strategy we also consider an M-PAM (M-ary Pulse Amplitude Modulation) signaling that is quite often used by the watermarking community. The M-PAM signal constellation consists of $M \geq 2$ equidistant real symbols centered on the origin, i.e., $\mathbf{w} = \frac{d_0}{2}\{-M+1,-M+3,...,M-1\}$ (Figure 1), where d_0 is the minimum distance between symbols. For equiprobable symbols the average symbol energy is $E_w = \left(M^2-1\right)d_0^2/12$. Figure 2 shows the AWGN channel capacity and the achievable rates with equiprobable M-PAM, for M=2,4,…,64 [1]. The highest rate for the uncoded M-PAM is $R = \log_2 M$.

2.1.1. Host interference cancellation (HIC)

To achieve the capacity of the AWGN channel one should suppress the interference with the host signal:

$$\mathbf{y} = \mathbf{w} + \mathbf{x} + \mathbf{n}. \qquad (3)$$

In this case, the resulting capacity is reduced to:

$$C = \frac{1}{2}\log_2\left(1+\frac{s_w^2}{s_x^2+s_n^2}\right). \qquad (4)$$

We will refer to (3) as a direct spread spectrum (SS) technique. Obviously, the higher the variance of the host signal s_x^2 (as is the case for textures and edges), the higher is the reduction in capacity. Figure 3 shows the capacity of the SS technique for WIR=-16dB where $\mathbf{WIR} = 10\log_{10} s_w^2/s_x^2$ is the watermark-to-host-image ratio. The SS technique approaches the capacity as $s_x^2 \to 0$, i.e., for the flat areas.

There are two main approaches that provide HIC considering the data hiding problem in the framework of communications with side information (SI). The first approach is based on the so-called *dirty codes* that utilize the SI at the encoder and includes quantization index modulation (QIM) [2], scalar Costa scheme (SCS) [3] and latice codes as particular cases. We have chosen the binary SCS proposed by Eggers *et al* [3] as the reference technique due to its superior performance in this class of methods and to the simplicity of its codebook (Figure 3). The behavior of this scheme in the case of M-ary signaling is asymptotically similar to the M-PAM due the equidistant constellations and to the final uniform distribution of the watermark. One can conclude based on Figure 3 that in the low-WNR regime (WNR<0dB) this technique does not approach the capacity of the AWGN channel.

The second approach overcomes the HIC using the SI at the decoder. This approach estimates the interference signal and subtracts it from the stego data. Thus, the name of this method is *estimation-substracting* (ES) [4]. Assuming a maximum a posteriori (MAP) estimate of the host data, one can derive the resulting capacity:

$$C = \frac{1}{2}\log_2\left(1+\frac{s_w^2}{s_e^2+s_n^2}\right) \qquad (5)$$

where $s_e^2 = \frac{s_x^2 s_z^2}{s_x^2+s_z^2}$, and $s_z^2 = s_w^2 + s_n^2$. This capacity is also shown in Figure 3 for WIR=-16dB. The ES scheme has higher capacity for very low-WNR regimes and the difference is negligible for WNR<-10dB. Therefore, this scheme has superior performance in comparison to the SCS in this regime. Moreover, this scheme approaches capacity for the flat image regions. Taking into account the empirical observation that real world images contain roughly 10-20% of edges and textures (we determined these numbers based on the integration of long tails in the

p.d.f.s of real images in wavelet subbands), one can conclude that the ES is nearly optimal for the low-WNR regime.

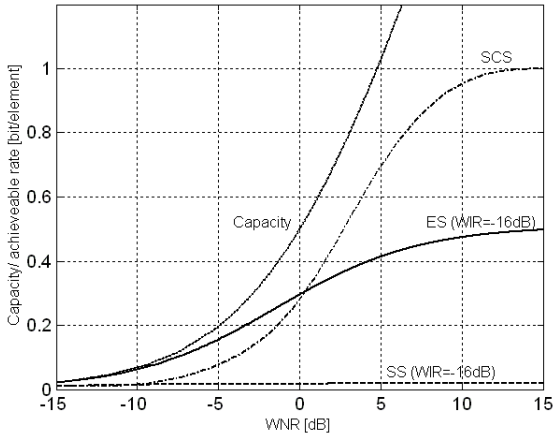

Figure 3. Capacity of the direct spread spectrum (SS), scalar Costa scheme (SCS) and estimation-substracting (ES) techniques.

2.1.2. Watermarking in low-WNR regime

The AWGN channel capacity for the low-WNR regime (WNR<0dB) is less than 0.5bit/pixel. Therefore, one can exploit this regime only for applications that do not require high embedding rates. This is typical for robust watermarking or for secret communications where the embedding rate is artificially decreased to reduce the success of steganalysis. Moreover, in the low-WNR regime an equiprobable binary alphabet is nearly optimal, providing a negligible reduction in capacity (Figure 2). Therefore, the most appropriate model of watermark signaling for this regime is the 2-PAM.

2.1.3. Watermarking in high-WNR regime

In the high-WNR regime, the capacity of equiprobable M-PAM constellations asymptotically approaches a straight line parallel to the capacity of the AWGN channel (Figure 2). The asymptotic loss of 1.53dB is due to using a uniform rather than the Gaussian distribution. Therefore, non-binary signal constellations must be used. Finally, to approach capacity, coding techniques that use constellation-shaping have to be used.

Therefore, for optimal signaling, the final stochastic model of the watermark for the low-WNR and the high-WNR regimes should differ. One can use low-rate binary codes with soft decoding for the low-WNR regime. Therefore, the appropriate model of the watermark is the equiprobable 2-PAM. Oppositely, the model of the watermark for the high-WNR regime will be more Gaussian-like. The reason for that is based on the desire to overcome the 1.53dB shaping gap for the uniform constellations. Therefore, one can assume two major categories for the watermark detection problem:

- detection of a *binary known* watermark;
- detection of an *unknown (random)* watermark with Gaussian p.d.f..

2.2. Attacker strategy

According to our two main applications we consider only steganography and robust watermarking. The attacker strategy in these two applications might be different. In the case of steganography, the attacker applies a steganalysis. The mere fact of detecting some hidden data is already enough to completely break the established secret communication protocol. In the case of robust watermarking, the attacker assumes that the image contains some watermark either with certainty, or with high probability. However, the technology used for the watermark embedding could be unknown. Therefore, the goal of the attacker is to decrease the rate of reliable communication, while preserving the image quality so that the image could be reused further in some prohibited way.

3. WATERMARK DETECTION

We consider two watermark detection problems for the two watermark models considered above. The first detector is a MAP detector for a known watermark. The second detector is an estimation-correlator detector for a random watermark.

3.1. MAP detector

The simplest watermark detection scheme can be derived using multiple hypothesis testing. It is a common practice to use for this purpose a MAP detector that minimizes the probability of error (although a Neyman-Pearson (NP) detector can be used as well) assuming known signaling. In this case, the problem is similar to the M-ary pulse amplitude modulation (PAM). Consider the simplest common case of 3-PAM that is typical of LSB embedding (level "0" assumes no embedding). The hypotheses under test are:

$$\begin{aligned} H_0 &: y[k] = x[k] - d_0/2 \\ H_1 &: y[k] = x[k] \\ H_2 &: y[k] = x[k] + d_0/2 \end{aligned} \qquad (6)$$

where d_0 can be chosen to provide maximum robustness for the given image quality. The MAP detector takes the decision:

$$H = \arg\max_j p(H_j)\, p(y_n | H_j) \text{ for all } k. \qquad (7)$$

We have investigated this type of detection for different cover image models. The probability of error $P_e = \frac{4}{3} Q\left(\sqrt{\frac{N d_0^{\,2}}{4 s_x^{\,2}}}\right)$ is a well known expression for M-ary PAM (LSB watermark with a constant mask) and the

stationary Gaussian cover image model. In our case M=3 and we have six types of errors. N=1 for the detection on the sample level (pixel-wise detection) and N can be increased, if the watermark bit allocation scheme is known and the watermark is repeated. In the more general case of M-PAM signaling: $P_e = \frac{2(M-1)}{M} Q\left(\sqrt{\frac{Nd_0^2}{4s_x^2}}\right)$. This also indicates a possible security leakage in the data hiding system expressed as the knowledge of the spatial watermark allocation. Therefore, the reliable detection is only possible for relatively small image variance, that is the case for the flat image regions and/or relatively large sample space N.

3.2. Estimation-correlation detector

In the case of a random watermark, i.e. a watermark weighted by an unknown perceptual mask or encoded/shaped to approach channel capacity, we can assume that the watermark is a zero mean Gaussian random process with a known covariance (i.e., some information about its periodicity might be available). The presence of the watermark is then detected according to the following hypothesis testing:

$$\begin{aligned} H_0 &: y[k] = x[k] \\ H_1 &: y[k] = x[k] + w[k] \end{aligned} \tag{8}$$

In this case, a NP detector decides about the presence of the watermark, if the likelihood ratio exceeds a threshold:

$$L(\mathbf{y}) = \frac{p(\mathbf{y}; H_1)}{p(\mathbf{y}; H_0)} > g \tag{9}$$

The typical assumptions can be $\mathbf{y} \sim N\left(\mathbf{0}, s_x^2 \mathbf{I}\right)$ under H_0 and thus $\mathbf{y} \sim N\left(\mathbf{0}, s_x^2 \mathbf{I} + \mathbf{C}_w\right)$ under H_1. The resulting detector will consist of the so-called *estimator-correlator structure* (Figure 4).

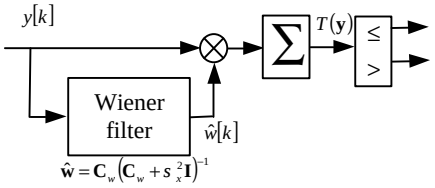

Figure 4. The estimation-correlation detector**.**

4. WATERMARK ESTIMATION

One can also first estimate the parameters of the watermark and then make the decision or more generally classify the watermark. The MAP estimator is well suited for this goal:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p_{y|\mathbf{w}}(\mathbf{y}|\mathbf{w}) p_{\mathbf{w}}(\mathbf{w}) \tag{10}$$

If we assume $\mathbf{x} \sim N(\bar{\mathbf{x}}, \mathbf{C}_\mathbf{x})$ and $\mathbf{w} \sim N\left(\mathbf{0}, s_w^2 \mathbf{I}\right)$, where $\mathbf{C}_\mathbf{x}$ is diagonal, we obtain the well-known Wiener filter as the solution of the MAP estimation problem:

$$\hat{w}[k] = \frac{s_w^2}{s_w^2 + s_{x[k]}^2} (y[k] - \bar{y}[k]) \tag{11}$$

The pixel-wise variance of this estimate is:

$$\mathbf{Var}[\hat{w}[k]] = \frac{s_w^2 s_{x[k]}^2}{s_w^2 + s_{x[k]}^2} \tag{12}$$

that again supports the conclusion that the accurate estimation of a watermark is only possible for relatively small s_x^2, i.e., in flat image areas.

5. CONCLUSION

In this paper we theoretically analyze the trade-off between capacity and security of data hiding technologies. The stochastic image and watermark models play a crucial role in this analysis. In the simplest case of Gaussian image model, it is evident from the presented results that the image regions with higher variance provide higher security (both the probability of detection error and the estimator variance are increased). Oppositely, the capacity of the SS and ES techniques is decreased in these regions especially for the high-WNR regime.

6. ACKNOWLEDGMENT

The authors are thankful to Joahim Eggers, Frederic Deguillaume, Yuriy Rytsar and Oleksiy Koval for many interesting and fruitful discussions. This work has been partially supported by the European Certimark project and SNF grant No 21-064837.01.

7. REFERENCES

[1] G. Ungerboek, "Trellis Coded Modulation with Redundant Signal Sets, Parts. I+II, *IEEE Comm. Mag*., 25, 2, 5-21, 1987.

[2] B. Chen and G.W. Wornell, "Provably robust digital watermarking", *Proc. SPIE*, Vol. 3845, 43-54, Boston, USA, Sept. 1999.

[3] J.J. Eggers, J.K. Su and B. Girod, "A Blind Watermarking Scheme Based on Structured Codebooks," *IEE Coll.: Secure Images and Image Authentication*, London, UK, Apr. 2000.

[4] S. Voloshynovskiy, F. Deguillaume, T. Pun, "Content adaptive watermarking based on a stochastic multiresolution image modeling", *EUSIPCO'2000*., Sept. 5-8, 2000, Tampere, Finland.