PROVABLY SECURE STEGANOGRAPHY: ACHIEVING ZERO K-L DIVERGENCE USING STATISTICAL RESTORATION

K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran

Dept. of Electrical and Computer Engineering
University of California at Santa Barbara
Santa Barbara, CA 93106

ABSTRACT

In this paper, we present a framework for the design of steganographic schemes that can provide provable security by achieving zero Kullback-Leibler divergence between the cover and the stego signal distributions, while hiding at high rates. The approach is to reserve a number of host symbols for statistical restoration: host statistics perturbed by data embedding are restored by suitably modifying the symbols from the reserved set. A dynamic embedding approach is proposed, which avoids hiding in low probability regions of the host distribution. The framework is applied to design practical schemes for image steganography, which are evaluated using supervised learning on a set of about 1000 natural images. For the presented JPEG steganography scheme, it is seen that the detector is indeed reduced to random guessing.

1. INTRODUCTION

Security and secrecy of information has always been important to people, organizations, and governments. In this paper, we consider the problem of *steganography*: a *message* is to be embedded into an innocuous looking *host* or *cover* to get a *stego* or *composite* signal, such that the presence of hidden data is invisible, both *perceptually* and *statistically*.

We build-upon a framework, called *statistical restoration* (proposed in our previous work [1]), for the design of embedding schemes that can evade statistical steganalysis while hiding at high rates, and achieve robustness against attacks. We are motivated by the notion of ϵ-secure steganography proposed by Cachin [2], in which the relative entropy (also called Kullback-Leibler or K-L divergence) between the cover and stego distributions is less than or equal to ϵ. Techniques proposed in [1] can achieve a small ϵ using *statistical restoration*: a portion of the data-hider's "distortion budget" is spent in repairing the damage done to the host statistics by the embedding process.

In the framework presented here, one can achieve *provable security* by having **zero** K-L divergence between the cover and the stego signal distributions, while hiding at high rates. The probability density function (pdf) of the stego signal *exactly* matches that of the original cover, and hence no statistical steganalysis can detect the presence of embedded data. This result, however, must be used with caution, since it requires that dependencies of all orders be restored to match the original. For example, a JPEG steganography scheme, designed to match the histogram of discrete cosine transform (DCT) coefficients, can still be detected by steganalysis techniques that exploit cover memory. The good news, however, is that the framework is general, and can be applied to restore higher-order statistics as well.

Another advantage of the proposed framework is that it allows design of robust techniques that are not fragile against attacks, unlike many other steganographic methods. Also, one can match continuous statistics using the proposed approach, not just discrete (or quantized) statistics. The techniques do not rely on accurate modeling of the host statistics. Mathematical analysis to estimate the allowable embedding rate for the proposed zero-divergence steganography is provided in [3]. This analysis takes into account the fact that there are only a finite number of host samples available to the hider and the detector, an aspect sometimes ignored by prior works.

In order to demonstrate the practical applicability of our framework, we implement a JPEG steganography scheme that perfectly restores the DCT coefficient distribution while hiding several thousand bits into images (e.g., 25000 bits in a 512x512 image). In our evaluation, we assume that the adversary knows the embedding algorithm, and can train a detector specifically tuned to our embedding schemes using supervised learning. In the experiments, a set of 1000 natural images is used to train a support vector machine (SVM). In spite of such stringent steganalysis, we find that detection is reduced to *random guessing* for our proposed high-capacity JPEG steganography technique.

2. RELATED WORK

Provos' Outguess [4] is an early attempt at restoring the stego distribution to the cover empirical distribution. This method was followed up later by Eggers et al [5], with a more mathematical formulation denoted histogram-preserving data mapping (HPDM), and Franz, with work in matching the message data to the cover distribution [6]. All of these schemes are designed for compensating discrete-valued hiding medium, and are also fragile against attacks.

We know of only two methods that can potentially achieve zero K-L divergence for continuous host statistics. Guillon et al [7] suggest transforming the source to a uniform distribution, and then embedding data using quantization index modulation (QIM). QIM is known not to change the probability mass function (PMF) of uniform sources. However, apart from difficulties in practical implementation (in companding to a uniform distribution), the method is not likely to be robust, and also, there is no systematic way to control the embedding distortion. Another approach called stochastic QIM [8] can potentially achieve zero K-L divergence. However, because of the stochastic nature of the hiding process, the method is likely to

This research is supported in part by a grants from ONR # N00014-01-1-0380 and #N00014-05-1-0816.

yield high error rates when embedding large volumes of data.

Other methods such as stochastic modulation [9], Fridrich et al's JPEG perturbed quantization (PQ) [10], and Sallee's model based embedding [11] accept a change of stego distribution from the original cover, but attempt to match a different distribution, which is close to a plausible cover distribution. It is difficult to define what is "plausible enough", and in some cases (e.g. [12]) a steganalyst can exploit the divergence from the original. Additionally, these approaches are very fragile to any interference between sender and receiver. Note that our methods do not rely on accurate modeling of the host statistics (to define a plausible distribution). Moreover, the proposed framework allows design of robust techniques that are not fragile against attacks, unlike prior approaches such as OutGuess [4], HPDM [5], PQ [10], and model based methods [11].

3. PROVABLY SECURE STEGANOGRAPHY

In this section, we describe our approach for secure steganography. We start by presenting a brief review of the concept of statistical restoration initially proposed in [1].

3.1. Statistical Restoration

In the game between the steganographer and the steganalyst, the advantage with the steganographer is that he or she is 'informed' of the cover signal statistics. Thus, he or she can be assured of perfectly secure communication simply by sending a composite signal whose statistics resemble that of the original cover. A natural way to accomplish this is to spend a part of the allocated distortion budget to *restore* the statistics.

In the statistical restoration framework, the host symbols are divided into two *streams*: an embedding stream, and a compensation stream. The goal is to match the continuous probability density function (pdf) of the cover signal. We use QIM with dithering to embed the data into host symbols in the embedding stream, thus making sure that we do not leave any "gaps" in the stego pdf. Next, the host symbols in the compensation stream are modified to match the original, while incurring minimum mean-squared error. This design ensures that the robustness properties of the employed embedding algorithm remain intact. Note that previous compensation approaches use entropy codecs [5, 11], and hence, are fragile against attacks.

In real-world systems, the steganalyst does not have the perfect knowledge of the cover signals (i.e., the continuous pdfs). Moreover, only a finite number of host samples are available for analysis. From the available host samples, the steganalyst must calculate a histogram approximation of the cover distribution, using a bin size w. Our data hiding is secure if we match the stego histogram to the cover histogram with the bin size, w.

We seek to maximize the ratio of symbols used for hiding, denoted $\lambda \in [0, 1]$, for a given cover distribution. Denoting the cover PMF as $P_X[i]$, the standard (uncompensated) stego PMF as $P_S[i]$, the achievable embedding rate (derived in [1]) is given by $\lambda^* = \min_i \frac{P_X[i]}{P_S[i]}$. If we apply this constraint to typical PMFs, we run into erratic behavior in the low-probability tails. The ratio $\frac{P_X[i]}{P_S[i]}$ can vary widely here, from infinitesimally small to huge. In the work presented in [1], this problem is solved by relaxing exact equality constraint: a small low-probability region is ignored for compensation. Using this approach, we can communicate at high rates, but there is always a low non-zero divergence between cover and stego signals, which can be exploited by the steganalyst.

3.2. Zero Divergence Steganography

The idea for achieving zero K-L divergence is quite simple. As seen in the previous section, since the low-probability region is hard to compensate, we just avoid embedding in that region. Thus, our new hiding strategy is not to hide in any symbol whose magnitude is greater than a predefined threshold[1] T. We define a hiding region $\mathcal{H}$ as $\mathcal{H} \triangleq [-T, T]$. The net rate, denoted R, can be given as, $R = \lambda^* \sum_{i \in \mathcal{H}} P_X[i]$, where λ^* is now defined over the hiding region: $\lambda^* = \min_{i \in \mathcal{H}} \frac{P_X[i]}{P_S[i]}$.

Changing the threshold affects the net rate in two ways: (i) if the threshold is reduced, there is a reduction is number of host samples used for embedding, which reduces the effective rate, and (ii) a smaller hiding region can lead to a higher λ^*, since the minimization of $\frac{P_X[i]}{P_S[i]}$ occurs over a narrower high-probability region. A detailed analysis of the optimum threshold and achievable rates is presented in [3].

In practical systems, the choice of threshold cannot be arbitrary, since we must make sure that the embedded data is decodable at the receiver. For QIM embedding, we can get around this problem by choosing the threshold to be an integer multiple of the quantization interval Δ. In the presence of attacks, the dynamic embedding strategy can potentially lead to desynchronization of the decoder. Thus, if attacks are anticipated, we use the coding framework proposed in [13], which allows the encoder to choose the embedding locations without explicitly sending that information to the decoder.

Figure 1 shows a zero-divergence steganography example for a Gaussian host $\mathcal{N}(0, 1)$. QIM embedding with $\Delta = 2$ is used (so that $\sigma/\Delta = 0.5$). The bin-width is $w = 0.05$, and threshold is $T = 1$. In this example, we hide 33,242 bits in 100,000 host samples and achieve perfect restoration.

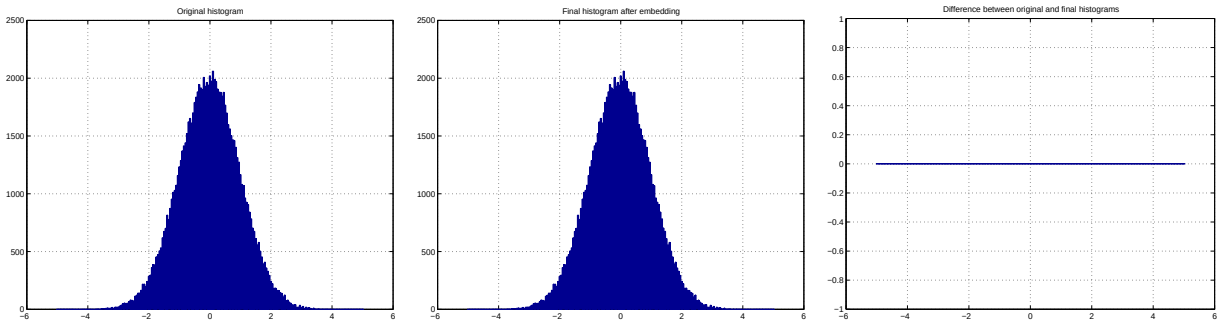
3.3. Variable Bin Size

When the involved distributions are continuous, a fixed bin size w is used for the analysis of the statistics. It is natural to ask what happens if the steganalyst analyzes the statistics with a finer bin size. Note that when there are finite number of samples, a finer bin-size does not guarantee a better observation (see [14] for a discussion on optimal bin sizes).

A potential solution is to employ a system with variable bin-size: instead of fixing the bin-width w, we fix the number of host symbols n_h in every bin. Thus, the bin-width gets automatically adjusted in such a way that it is finer in the high probability regions, and wider in the low-probability regions. The idea here is to match the original histogram more precisely in the high probability regions compared to the low-probability parts.

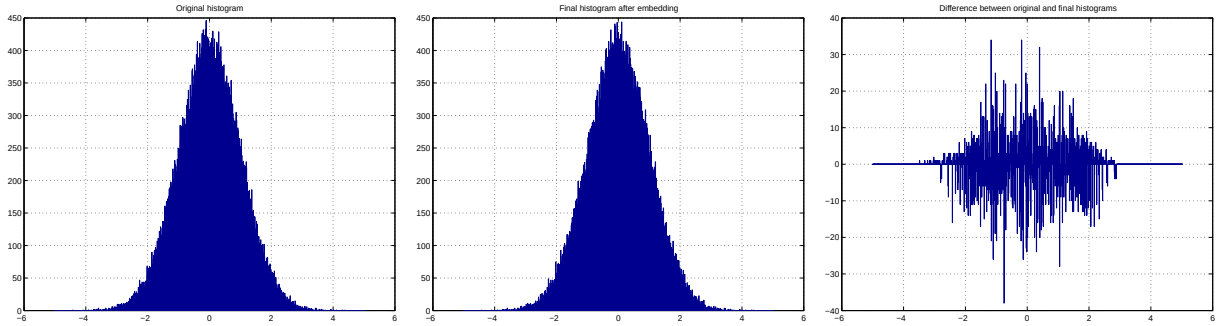
Figure 2 shows variable bin-size embedding example for a Gaussian cover $\mathcal{N}(0, 1)$. The bins have exactly 250 host symbols. The $\sigma/\Delta = 0.5$, number of samples are 100,000, the λ is 0.45, and the threshold is $T = 1$. The histograms are displayed for a bin-width of 0.01. Notice that in spite of such finer analysis (five times smaller bin size than previous example of Figure 1) and high rate of embedding, the difference between cover and stego histograms is very small.

[1]Note that a threshold can be used to define high-probability region for all peaked unimodal distributions (e.g., Gaussian, generalized Laplacian, or generalized Cauchy).



(a) Original histogram (b) Final histogram after embedding and compensation (c) Difference between original and final histograms

Fig. 1. Simulations for a Gaussian cover: Zero K-L divergence can be achieved while hiding at high rates (0.33 bits/symbol). Figure (c) above shows that the difference between cover and stego distributions is *exactly* zero.



(a) Original histogram (b) Final histogram after embedding and compensation (c) Difference between original and final histograms

Fig. 2. Variable bin-size compensation for a Gaussian cover: The original and final histograms, and their differences. Even with five times finer histogram analysis than in Figure 1, the difference, as shown in Figure (c) above, is very small.

4. APPLICATION TO IMAGE STEGANOGRAPHY

In this section, we apply the zero-divergence statistical restoration framework to design practical methods.

4.1. JPEG Steganography

Here we describe an adaptation of our zero K-L divergence framework for a JPEG steganography scheme. The goal here is to embed in a JPEG compressed image at a particular quality factor, such that the stego image is also a JPEG image at the same quality factor with *exactly* the same distribution of the DCT coefficients.

The host image is divided into 8×8 non-overlapping blocks and its 2-d DCT is taken. Those coefficients that lie in a low frequency band of 21 coefficients are considered to be eligible for data embedding or compensation. Now, out of all eligible coefficients, a fixed percentage (we use 25-40% in our experiments) are set aside for hiding and the rest are used for compensation. The hiding and compensation locations are pre-determined based on a secret key shared between the encoder and the decoder. We then embed data using $\pm k$ LSB steganography (with $k = 1$) into the quantized DCT coefficients that are in the hiding stream. Note that QIM cannot be used because the coefficients here are already quantized. As prescribed by the zero-divergence framework of Section 3.2, only those coefficients whose magnitude is smaller than a positive integer threshold T, are used for data embedding. The coefficients in the compensation stream are modified as per the minimum mean-squared error (MMSE) algorithm presented in [1].

Using the above approach, we ensure that the DCT coefficient distribution achieves exactly zero K-L divergence. This scheme is still prone to steganalysis techniques that bank on the increase in the blockiness due to block-DCT embedding [15]. In [1], we propose a method to restore the pixel value differences within the blocks as well as along the block boundaries, so as to survive these blockiness based steganalysis techniques.

4.2. Flexible Image Steganography

Here we describe a flexible technique that accepts two quality factors (QFs): a design QF, and an advertised QF. The idea is to "advertise" the images at a higher quality factor, but still design to survive compression at lower QFs. As an example, when the scheme has advertised QF of 75 and design QF of 50, the released stego images' DCT coefficients resemble the PMF of 75 QF compressed cover images. However, the images can survive attacks of up to 50 QF. Also note that a JPEG attack would not compromise the steganographic security of the image.

The implementation of this technique is similar to the above JPEG steganography scheme, except for the use of a dithered QIM embedding. The employed QIM's quantizer corresponds to the de-

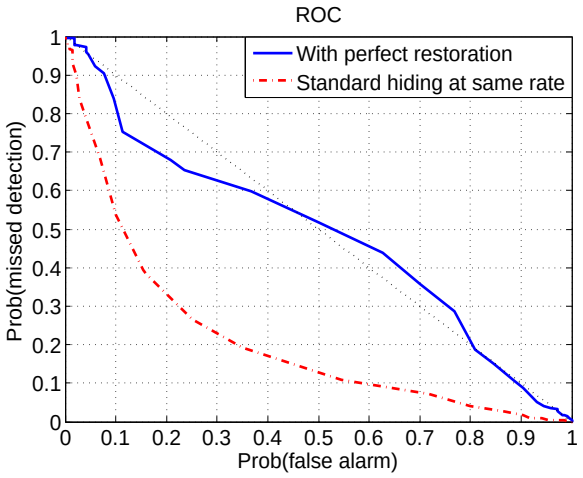


Fig. 3. Detection of JPEG steganography with standard QIM verses perfect restoration QIM. As expected, the detection for perfect-restoration JPEG scheme is random. However, the standard QIM at same rate is detectable.

sign QF. The dither pattern is also quantized using intervals defined by the advertised QF.

5. RESULTS

We now present the results for our JPEG steganography technique discribed in Section 4.1. We use supervised learning on about 1000 natural images to test the system. Parameters used for hiding are fixed for all the images in the test set: QF = 75, $T = 30$, and $\lambda = 0.4$. A SVM classifier is trained and tested on the first order statistics of the DCT coefficients. We here compare the perfect restoration JPEG steganography with the standard QIM. Same rate and same images are used in both the cases. Figure 3 plots the probability of missed detection verses probability of false alarm for both the schemes. As expected, the detector performance is random for the JPEG steganography scheme with perfect restoration.

Experiments with flexible steganography of Section 4.2 provide similar results. The images were designed to survive QF of 50 and were advertised at QF of 75. The embedded messages can be decoded with low error rates (less than 2%) from the images after JPEG attacks up to QF of 50. Note that error correcting codes must be employed to ensure perfect recovery.

6. CONCLUSION

In this paper, we demonstrate a provably secure steganography framework that can achieve zero K-L divergence between cover and stego distributions while embedding at high rates. Key to our efforts is the fact that we do not attempt to model the host statistics, but rather match the empirical density of the cover signal. Practical application of this framework to JPEG steganography indeed shows that the detector can be reduced to random guessing. The method, when integrated with the coding framework of [13], survives several attacks such as compression, additive noise, and tampering. The effect of such attacks on the detectability of the presence of embedded data has not been investigated and is an avenue of future work.

Other avenues for future work include applying the framework to other embedding methods (such as spread spectrum), or to other host signals (such as audio and video). Challenges also remain in dealing with steganalysis methods that consider several different statistical measures. We finally note the close relationship shared by this method with the earth-mover's distance (see [16]), a popularly used distance metric in computer vision applications.

References

[1] K. Solanki, K. Sullivan, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Statistical restoration for robust and secure steganography," in *Proceedings of ICIP*, Genoa, Italy, Sept. 2005.

[2] C. Cachin, "An information theoretic model for steganography," *LNCS: 2nd Int'l Workshop on Information Hiding*, vol. 1525, pp. 306–318, 1998.

[3] K. Sullivan, K. Solanki, U. Madhow, S. Chandrasekaran, and B. S. Manjunath, "Determining achievable rates for secure, zero divergence, steganography," in *Proceedings of ICIP*, Atlanta, GA, USA, Oct. 2006.

[4] N. Provos, "Defending against statistical steganalysis," in *In 10th USENIX Security Symp.*, Washington DC, USA, 2001.

[5] J. J. Eggers, R. Bauml, and B. Girod, "A communications approach to image steganography," in *Proceedings of SPIE: Security, Steganography, and Watermarking of Multimedia Contents IV*, San Jose, CA, Jan. 2002.

[6] E. Franz, "Steganography preserving statistical properties," in *5th International Working Conference on Communication and Multimedia Security*, 2002.

[7] P. Guillon, T. Furon, and P. Duhamel, "Applied public-key steganography," in *Proceedings of SPIE: Security, Steganography, and Watermarking of Multimedia Contents IV*, San Jose, CA, Jan. 2002.

[8] P. Moulin and A. Briassouli, "A stochastic QIM algorithm for robust, undetectable image watermarking," in *Proceedings of ICIP*, Singapore, Oct. 2004.

[9] J. Fridrich and M. Goljan, "Digital image steganography using stochastic modulation," in *Proceedings of SPIE: Security, Steganography, and Watermarking of Multimedia Contents IV*, Santa Clara, CA, USA, Jan. 2002, pp. 191–202.

[10] J. Fridrich, M. Goljan, P. Lisoněk, and D. Soukal, "Writing on wet paper," in *ACM workshop on Multimedia and Security*, Magdeburg, Germany, Sept. 2004.

[11] P. Sallee, "Model-based steganography," in *IWDW 2003, LNCS 2939*, Oct. 2003, pp. 154–167.

[12] R. Bohme and A. Westfeld, "Breaking Cauchy model-based JPEG steganography with first order statistics," *P. Samarati et al (Eds.): ESORICS 2004, LNCS 3193*, pp. 125–140, 2004.

[13] K. Solanki, N. Jacobsen, U. Madhow, B. S. Manjunath, and S. Chandrasekaran, "Robust image-adaptive data hiding based on erasure and error correction," *IEEE Trans. on Image Processing*, vol. 13, no. 12, pp. 1627 –1639, Dec. 2004.

[14] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[15] J. Fridrich, M. Goljan, and D. Hogea, "Steganalysis of JPEG images: Breaking the F5 algorithm," in *Lecture notes in computer science: 5th Int'l Workshop on Information Hiding*, 2002, vol. 2578, pp. 310–323.

[16] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth movers distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.