STATISTICAL RESTORATION FOR ROBUST AND SECURE STEGANOGRAPHY

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

We investigate data hiding techniques that attempt to defeat steganalysis by restoring the statistics of the composite image to resemble that of the cover. 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. While statistical restoration has broad applicability to a variety of hiding methods, we illustrate our ideas here for quantization index modulation (QIM) based hiding. We propose a method for significantly reducing the detectability of QIM, while preserving its robustness to attacks. We next use the framework of statistical restoration to develop a method to combat steganalysis techniques which detect block-DCT embedding by evaluating the increase in blockiness of the image due to hiding. Numerical results demonstrating the efficacy of these techniques are provided.

1. INTRODUCTION

In recent years, there has been a great deal of activity in developing data hiding techniques, which have classical applications to steganography, or covert communication, as well as to watermarking for digital rights management. The typical objective in high-volume data hiding is to embed data in a *host* or *cover*, in a manner that is resistant to a number of natural and malicious attacks, and is imperceptible to the casual observer. However, the resulting *stego* signal can be subjected to increasingly sophisticated *steganalysis* techniques for detecting the presence of hidden data. Thus, modern steganography is a game with escalating sophistication between the hider and the steganalyst. One of the first popular steganalysis tools was *Stegdetect* [1], which uses a chi-square statistic on the histogram of transform coefficients to detect least significant bit (LSB) hiding. Stegdetect can be improved upon by more sophisticated detection-theoretic approaches [2]. Such methods, which are based on the histogram of the host coefficients, have spurred the development of hiding techniques that make as little change to the histogram as possible. Provos' Outguess algorithm [3] was an early attempt at histogram compensation for LSB hiding, while Eggers et al [4] suggest a more rigorous approach to the same end, using histogram-preserving data-mapping (HPDM). In turn, steganalysis tools that counter such histogram-preserving hiding methods have been developed, such as detection, for image-based hiding, of block-DCT embedding by evaluation of the increase in blockiness due to hiding [5, 6]. While both HPDM and OutGuess attempt to match the quantized histogram of the discrete cosine transform (DCT) coefficients, more recent proposals [7, 8] try to match the continuous marginal statistics.

In this paper, we propose a framework that allows design of embedding schemes that can evade statistical steganalysis while hiding at high rates, and also achieve robustness against attacks. We are motivated by the notion of ϵ-secure steganography proposed by Cachin [9], 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. Our approach for achieving a small ϵ is to employ *statistical restoration*, wherein a portion of the data-hider's "distortion budget" is spent in repairing the damage done to the image statistics by the embedding process. To ensure that the restoration does not interfere with decoding, a fixed percentage of host symbols are set aside for restoration, while the rest are used for embedding. A secret key, shared between the encoder and the decoder, determines the embedding and compensation locations.

While statistical restoration for reducing detectability is a general concept, we focus in this paper on quantization index modulation (QIM) based data hiding. QIM allows the embedding of large volumes of data in a manner that is resistant to a variety of attacks: see [10, 11] for an information-theoretic treatment, and [12] for a constructive coding framework for image-adaptive QIM-based hiding. However, standard QIM is relatively easy to detect for the steganalyst using simple statistical criteria. In this paper, we propose a framework that preserves the desirable properties of QIM (high-volume embedding, robustness to attacks), while significantly reducing its detectability.

We first design a system that matches the first-order statistics while hiding. The scheme uses dithered QIM, based on our existing scheme [12] with a reduced rate, saving some host discrete cosine transform (DCT) coefficients for modifying the histogram back to the original. If enough host coefficients remain, then the original histogram can be matched exactly by the new histogram, ensuring zero K-L divergence. While our approach is similar in spirit to prior histogram-compensation methods such as OutGuess [3], and HPDM [4], there are some significant differences. First, unlike [3, 4], dithering is used, so that the histogram matches the original source, not the quantized (compressed) version of the original. The stego image can therefore be advertised as any uncompressed format, (e.g. TIFF, BMP, RAW) or subsequently compressed at any quality factor and will continue to match the source exactly. We believe that the ability to exactly match the original *unquantized* source histogram is an important contribution of the present work. An additional advantage of the proposed scheme is that, because it employs QIM, it is robust against benign distortion-

This research is supported in part by a grant from ONR # N00014-01-1-0380.

constrained attacks (e.g., recompression, and additive noise), unlike HPDM and OutGuess, which employ fragile LSB hiding.

Guillon et al [7] suggest transforming the source to get a uniform PMF source. The message is hidden in this with the quantization hiding scheme, which is known not to change the PMF of uniform sources. Therefore, the PMF after transforming back is also the same as the original. This method, however, is not likely to be robust, and also, there is no way to control the distortion induced by the embedding process. Another interesting approach is that of Wang and Moulin's [8], who propose a reduced rate variant of standard QIM, called the stochastic QIM, which can be made to have zero K-L divergence. However, because of the stochastic nature of the hiding process, the method is likely to yield high error rates when embedding large volumes of data.

Sallee's model-based steganography [13] provides an interesting and different perspective in the design of steganographic systems, with the hider ensuring that the stego signal conforms to a given model. However, in the absence of a perfect model for the host, nothing stops the steganalyzer from selecting a *better* model by spending more computational power. This is indeed practically shown in [14], where Sallee's Cauchy-model based JPEG steganography is broken by using only the first order statistics. Our approach is very difficult to detect in this manner, since the stego marginals are simply restored to conform to the host's empirical density, rather than invoking a statistical model for the host's marginals.

For any statistical restoration technique, the steganalyst can always go one step further, and use higher order statistics than those that have been compensated for, typically at the cost of higher computational complexity. Thus, hiding techniques that compensate marginals are easily detected using the cover memory. For example, a few approaches (Fridrich et al [5], and Wang and Moulin [6]) detect block-DCT embedding by modeling the increase in *blockiness* of the image due to the block-DCT hiding. We use our framework of statistical restoration to design a method that defeats this type of block-based steganalysis. In this case, the statistic to be restored is the difference of adjacent pixels values within the blocks and on the block boundaries.

We use supervised learning on a set of over 1000 natural images to evaluate the performance of our schemes. We find that statistical restoration severely affects the steganalysis performance of both DCT-histogram and blockiness methods. We achieve very low K-L divergence between original and cover distributions at fairly high embedding rates. The image could also survive JPEG compression or recompression without compromising the undetectability.

2. STATISTICAL RESTORATION

Let us consider an image to be a particular realization of an underlying stochastic process. If this 'magic' stochastic process is known to a steganalyzer, all he or she needs to do is to use this model to decide whether the image is *natural* or not. Obviously, it would be impossible to characterize such a random process, and hence, certain simplified statistics are considered for steganalysis in practice. This is what generates the room for the data-hider. The advantage with the data-hider is that he or she is 'informed' of the cover image, and hence its statistics. Thus, he or she can be assured of perfectly secure communication simply by sending a composite image 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. Note that we are considering the simplified statistics under scrutiny, and not the complete underlying random process.

In order to make sure the restoration process does not interfere with decoding, we allocate certain coefficients for embedding and use the rest for restoration. By separating the hiding and compensation locations, we make sure that the robustness properties of the employed embedding algorithm remain intact. This is unlike previous compensation approaches that use entropy codecs [4, 13], and hence, are fragile against attacks.

The restoration process reduces the size of the message that can be hidden, which is the cost of increasing the security. We can characterize this cost by studying the amount of data that can be hidden in an idealized data source with a given probability mass function (PMF). Let λ be the ratio of host symbols used for hiding, so $1-\lambda$ is the ratio remaining to match the cover PMF. If $P_X(n)$ is the cover PMF, $P_S(n)$ the standard (uncompensated) stego PMF, $P'_C(n)$ and $P_C(n)$ the PMF of compensating host symbols before and after compensation respectively, and $P_Y(n)$ the PMF of the final output, our goal can be summarized as:

$$P_Y(n) = \lambda P_S(n) + (1-\lambda)P_C(n) = P_X(n)\ \forall n \qquad (1)$$

Typically P_S can be derived directly from P_X. The amount of data that can be hidden is proportional to the number of samples that can be hidden in. So to maximize the amount of data we send, we seek to maximize λ for a given cover PMF subject to the constraint in (1), and the constraints imposed on the compensating PMF, namely $\sum P_C(n) = 1$ and $P_C(n) \geq 0\ \forall n$. Substituting $P_C = \frac{P_X(n)-\lambda P_S(n)}{1-\lambda}$ from (1), the first constraint is true for any λ. For the second constraint we find $\lambda \leq \frac{P_X(n)}{P_S(n)}\ \forall n$. This gives us an upper limit on the percentage of samples we can use for hiding, or equivalently, the rate at which we can secretly embed. Since the data-hider must choose a fixed percentage of symbols beforehand, λ can not be a function of n, and hence a worst-case λ is chosen: $\lambda = \min_n \frac{P_X(n)}{P_S(n)}$. We now address the next obvious question of how to actually perform the restoration. A strategy to modify the compensation host symbols with a minimum mean squared error (MMSE) criteria is discussed below.

2.1. Restoration with MMSE Criteria

The distribution of the compensation host symbols $P'_C(n)$ must be modified to a target distribution: $P_C(n) = \frac{P_X(n)-\lambda P_S(n)}{(1-\lambda)}$. This would not be as straightforward as saying that if the embedding process modifies a host symbol from A to B, find another host symbol (in the compensation stream) with value B and modify it to A. If for example the hiding process itself modifies another host symbol from B to A, the above change would not be required. It would be very inefficient if such an approach is followed. Another situation could be when $P(B) < P(A)$ so that one would soon run out of symbols with value B to compensate for data embedding. To efficiently use our distortion budget, we must modify the compensation stream to achieve a target distribution $P_C(n)$ with a MMSE criteria.

This problem of histogram modification with MMSE criteria was first considered by Mese and Vaidyanathan [15], who propose solving an integer linear programming problem to obtain a mapping matrix. Tzschoppe et al [16] show that a simpler solution exists, which does not require solving a linear programming problem. They prove a theorem essentially showing that to achieve a

MMSE mapping, all the bins of the target histogram must be filled in an increasing order by mapping the input data with values in increasing order. This means that first the bin $n = 1$ of the target histogram must be filled with $P_C(1)$ smallest compensation host symbols. The bin $n = 2$ will be filled next with the $P_C(2)$ smallest remaining symbols, and so on. We note that the mapping would be similar even if the process is started from the last bin and filled in a decreasing order.

In the actual implementation, the above algorithm is slightly modified to ensure that the high probability regions are compensated before the low probability tail. Instead of starting the compensation from the first index (i.e., the lowest value), we separate the positive and negative sections of the histogram and perform their restorations independently. For the histograms centered around zero, which is the case for both the practical scenarios considered in this paper, this procedure compensates the high probability regions first.

2.2. Rate vs. Security

Here we study the tradeoff between embedding rate and security. Let us revisit the conditions on the embedding rate λ derived above. If we apply the constraint $\lambda = \min_n \frac{P_X(n)}{P_S(n)}$ to typical PMFs, we run into erratic behavior in the low-probability tails. The ratio $\frac{P_X(n)}{P_S(n)}$ can vary widely here, from infinitesimally small to huge. e.g. $P_X(\text{event } A) = 1 \times 10^{-9}, P_S(A) = 1 \times 10^{-6}, \lambda = 0.001$; only a tenth of a percent of the samples can be used. Since this happens only in the low probability regions in general, the effect of PMF differences in these regions on the net divergence is small. So to avoid this problem we can relax exact equality constraint and ignore a small region of low probability. That is, we do not require compensation in a small, low probability region of the PMF. So now λ is chosen as the minimum $\frac{P_X(n)}{P_S(n)}$ over the high-probability compensated region.

In addition to the divergence introduced due to the ignored region, since (1) is not true for all n, P_C must be normalized to satisfy the unity sum constraint, adding a small change across the PMF. Though the net effect is to introduce a small amount of divergence, λ and the corresponding hiding rate can only increase.

The tradeoff between the desired security from detection and the hiding rate can be studied by finding the rate corresponding to several different sizes of ignored (uncompensated) regions. We also note that simply embedding in fewer coefficients also reduces the detectability. However, in Figure 1 we see that a large decrease in divergence can be made with a small drop in rate using restoration, which is not possible by merely embedding less. This is true for both Laplacian and Gaussian PMFs over a range of variances.

3. PRACTICAL SCHEMES

In this section, we describe two practical schemes based on the idea of statistical restoration.

3.1. Restoring Marginal Statistics

Several steganalysis approaches [1, 17] detect the JPEG steganography techniques by hypothesis testing on the marginal distribution of the DCT coefficients. We here propose a method that restores the histogram of the DCT coefficients so as to evade this type of steganalysis.

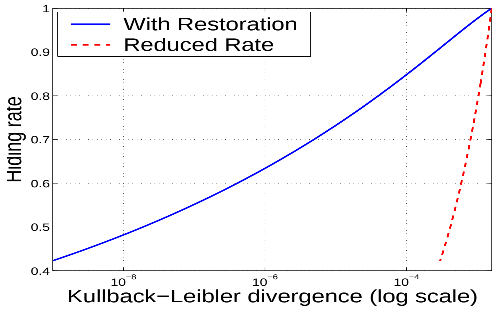

Fig. 1. Rate, security tradeoff for Gaussian cover. As expected, compensating is a more efficient means of increasing security than simply reducing the rate.

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, a fixed percentage of eligible coefficients (about 25-40%) are set aside for hiding and the rest are used for compensation. Data is embedded into the coefficients designated for hiding using dithered quantization. Finally, the compensation coefficients are modified using the algorithm described in Section 2.1 so that the stego image histogram closely matches that of the original cover.

The use of dithering in our design allows us to avoid *gaps* in the histogram so that, after restoration, we can match the *unquantized* source histogram. This way, we neither lose the embedded data nor compromise the undetectability even after the image is compressed or recompressed by the data-hider or an adversary.

The tradeoff between rate and security (Section 2.2) implies that the source histogram cannot be matched exactly if we want to communicate at a reasonable rate. Also, in practice, we must work with a limited number of compensation coefficients. Hence, depending on the chosen rate of embedding, we cannot perfectly match a part of the source histogram towards the low probability tail region. Therefore, we would expect a smart detector to perform better than just a random guess, and this partly explains the better-than-random performance of our supervised learning tests.

3.2. Defeating Block-Based Steganalysis

We now turn our attention to steganalysis schemes that use the cover memory to detect the hidden data. In particular, we focus on techniques that bank on the increase in the blockiness due to block-DCT embedding [5, 6]. It can be seen that these methods basically use a function or a subset of a two-dimensional histogram. For example, Wang and Moulin [6] use one-dimensional histograms of value differences of two populations: one within the blocks, and another along the block boundaries. We note that the value difference histogram can be derived by summing along the diagonals of the two-dimensional histogram. This way the most relevant information is kept while reducing the complexity (of a two-dimensional histogram). Here we propose a method that restores the pixel value differences within the blocks as well as along the block boundaries, so as to survive the steganalysis technique proposed in [6].

A subset of 8×8 blocks are used for data embedding and the rest are set aside for restoring the pixel difference histograms. In the blocks designated for data embedding, data is hidden in a low

Table 1. Performance of uncompensated vs. compensated methods for over 1000 images in supervised learning tests. It is seen that restoration can severely affect the steganalysis performance.

	Dithered QIM		Adaptive dithered QIM		Blockiness based scheme	
	Un-comp.	comp.	Un-comp.	comp.	Un-comp.	comp.
P(m)	0.075	0.525	0.701	0.796	0.043	0.259
P(fa)	0.177	0.000	0.000	0.074	0.000	0.007
P(m)+ P(fa)	0.252	**0.525**	0.701	**0.870**	0.043	**0.266**

frequency band comprised of 21 DCT coefficients. Next, the pixel values of the compensation blocks are modified (with MMSE criteria, as described in Section 2.1) so that the difference histograms are very close to the original. Note that the two histograms (within the blocks and along the block boundaries) are restored separately to match their respective originals.

4. RESULTS

As a first step in examining the efficacy of statistical restoration, we compare the divergence between cover and stego for standard hiding and for hiding with compensation at the same rate. Embedding at a rate of $\lambda = 0.35$ in a Gaussian cover, the divergence for statistically restored dithered-QIM hiding is 1.3×10^{-3}, roughly a five-fold improvement over the standard QIM which yields a divergence of 5.9×10^{-3}. Similar improvement is also seen for a set of real image statistics, wherein, the average divergence for standard hiding is 6.5×10^{-3}, which reduces to 2.1×10^{-3} for compensated embedding. Although detection is still possible, restoration greatly increases the error probabilities of an ideal detector. For example, a steganalyst would require more than three times as many samples to achieve the same detection rates with standard hiding in images as with hiding with restoration.

We next use a supervised learning machine on a set of over 1000 natural images to discriminate between the cover and the stego images (as in [17]). The machine is trained on the statistics of hundreds of examples of distinct stego and cover images, and is then tested on its ability to correctly classify a different, unknown set of cover and stego images. Three embedding methods were tested: dithered QIM, adaptive dithered QIM (of [12]), and blockiness based scheme (of Section 3.2). For each of these schemes, we trained and tested two machines on the same sets of images and at the same rate; one with compensation, one without. Table 1 lists the probability of false alarm, P(fa), and the probability of missed detection, P(m), for each of these configurations. It can be seen that for the dithered QIM hiding, the detector has twice the sum of errors while detecting restored hiding as compared to standard hiding. For the blockiness compensation scheme, the sum of errors is six times greater for restored hiding than for standard hiding. Note that, a λ of 0.35 is used in all the cases, which translates to hiding roughly 30100 bits in a 512×512 image.

5. CONCLUSION

We have demonstrated how statistical restoration can be employed for robust and secure communication. Our experiments indicate that the detectability of our statistically compensated QIM schemes is lower than the standard QIM. We show that we can significantly lower the detection rates for block-based steganalysis as well. However, the detection is still better than a random guess probably because we ignore certain low probability region for compensation. Possible modifications to improve the performance of these schemes are currently under investigation. We are also investigating whether and how we can employ the statistical restoration framework to other hiding schemes.

References

[1] N. Provos and P. Honeyman, “Detecting steganographic content on the internet,” in *ISOC NDSS’02*, San Diego, CA, 2002.

[2] O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran, and B.S. Manjunath, “Detection of hiding in the least significant bit,” *IEEE Transactions on Signal Processing, Supplement on Secure Media I*, vol. 52, no. 10, pp. 3046–3058, Oct 2004.

[3] N. Provos, “Defending against statistical steganalysis,” in *10th USENIX Security Symp.*, Washington DC, USA, 2001.

[4] J. J. Eggers, R. Bauml, and B. Girod, “A communications approach to image steganography,” in *Proceedings of SPIE*, San Jose, CA, 2002.

[5] J. Fridrich, M. Goljan, and D. Hogea, “Steganalysis of JPEG images: Breaking the F5 algorithm,” in *LNCS: 5th Int’l Workshop on Info. Hiding*, 2002, vol. 2578, pp. 310–323.

[6] Y. Wang and P. Moulin, “Steganalysis of block-DCT image steganography,” in *IEEE workshop on Statistical Signal Processing*, St Louis, MO, USA, Sept. 2003.

[7] P. Guillon, T. Furon, and P. Duhamel, “Applied public-key steganography,” in *Proceedings of SPIE*, San Jose, CA, 2002.

[8] Y. Wang and P. Moulin, “Steganalysis of block-structured stegotext,” in *Proceedings of SPIE*, San Jose, CA, 2004.

[9] C. Cachin, “An information theoretic model for steganography,” *LNCS: 2nd Int’l Workshop on Info. Hiding*, vol. 1525, pp. 306–318, 1998.

[10] B. Chen and G. W. Wornell, “Quantization index modulation: A class of provably good methods for digital watermarking and information embedding,” *IEEE Trans. on Info. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[11] P. Moulin and J. A. O’Sullivan, “Information-theoretic analysis of information hiding,” *IEEE Trans on Info. Theory*, vol. 49, no. 3, pp. 563–593, June 2003.

[12] 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.

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

[14] 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.

[15] M. Mese and P.P. Vaidyanathan, “Optimal histogram modification with MSE metric,” in *Proc. ICASSP*, Salt Lake City, Utah, USA, May 2001.

[16] R. Tzschoppe, R. Bauml, and J.J. Eggers, “Histogram modifications with minimum MSE distortion,” tech. rep., Telecom. Lab., Univ. of Erlangen-Nuremberg, Dec 2001.

[17] K. Sullivan, Z. Bi, U. Madhow, S. Chandrasekaran, and B.S. Manjunath, “Steganalysis of quantization index modulation data hiding,” in *Proc. ICIP*, Singapore, Oct. 2004.