HIGH-VOLUME DATA HIDING IN IMAGES: INTRODUCING PERCEPTUAL CRITERIA INTO QUANTIZATION BASED EMBEDDING

K. Solanki, N. Jacobsen, S. Chandrasekaran, U. Madhow, and B. S. Manjunath

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

ABSTRACT

Information-theoretic analyses for data hiding prescribe embedding the hidden data in the choice of quantizer for the host data. In this paper, we consider a suboptimal implementation of this prescription, with a view to hiding high volumes of data in images with low perceptual degradation. Our two main findings are as follows:
(a) In order to limit perceptual distortion while hiding large amounts of data, the hiding scheme must use perceptual criteria in addition to information-theoretic guidelines.
(b) By focusing on "benign" JPEG compression attacks, we are able to attain very high volumes of embedded data, comparable to information-theoretic capacity estimates for the more malicious Additive White Gaussian Noise (AWGN) attack channel, using relatively simple embedding techniques.

1. INTRODUCTION

Data hiding is the process by which a message signal, or signature, is covertly embedded within a host data set to form a composite signal. A well-accepted application of data hiding is for watermarking, which requires embedding a relatively short string of data in the host data (e.g., for identifying the source/copyright for the host), in a manner that is robust to a variety of attacks aimed at destroying the watermark while preserving the usability of the host. In contrast, our objective here is to obtain techniques for embedding high volumes of data in images, in a manner that causes minimal perceptual degradation, and is robust to "benign" JPEG compression attacks. The latter would occur naturally, for example, when an image with embedded data is being transmitted over a link with limited capacity.

Information-theoretic prescriptions for data hiding typically focus on hiding in independent and identically distributed (i.i.d) Gaussian host samples. Roughly speaking, these guidelines translate to hiding the data by means of the choice of quantizer for the host data, as first observed by Costa [1], based on results of Gel'fand and Pinsker [2] on coding with side information (with the host data playing the role of side information). Moulin et. al. [3] have subsequently built on these results to provide a game-theoretic analysis of data hiding, with the hider and attacker as adversaries, and have provided parallel Gaussian models for images that facilitate application of information-theoretic analyses. Chen and Wornell [4] present a variety of practical approaches based on similar ideas. A scalar quantization based data hiding scheme, together with turbo coding to protect the hidden data, is considered in [5], while a trellis coded vector quantization scheme is considered by Ramchandran et. al. [6].

In this paper, we consider scalar quantization based hiding schemes (as in [4]) in the Discrete Cosine Transform (DCT) domain. According to information-theoretic prescriptions for parallel Gaussian models of images in the DCT domain [3], data should be hidden in the low and mid frequency ranges of the host image, which have larger energies. However, we find that, in order to hide high volumes of data with low perceptual distortion, it is essential to add local perceptual criteria regarding which host coefficients to hide data in. Two different schemes for making such decisions are provided. By optimizing for JPEG attacks, even without coding, it is possible to attain practical hiding capacities that are comparable to the capacity estimates given in [3] for AWGN attacks.

This research was supported in part by a grant from ONR # N0014-01-1-0380.

2. QUANTIZATION BASED DATA HIDING

2.1. Embedding data in choice of quantizer

As in [4], signature data symbols can be embedded in the host medium through the choice of scalar quantizer. For example, consider a uniform quantizer of step size Δ, used on the host's coefficients in some transform domain. Let odd reconstruction points represent a signature data bit '1'. Likewise, even multiples of Δ are used to embed '0'. Thus, depending on the bit value to be embedded, one of two uniform quantizers of step size 2Δ is chosen. Moreover, the quantizers can be pseudo-randomly dithered, where the chosen quantizers are shifted by a pseudo-random sequence available only to encoder and decoder. As such, the embedding scheme is not readily decipherable to a third party observer, without explicit knowledge of the dither sequence.

Decoding is performed by quantizing the received coefficient to the nearest reconstruction point of all quantizers. An even reconstruction point indicates that a '0' has been hidden. Likewise, if a reconstruction point lies on an odd quantizer, a '1' has been hidden.

2.2. JPEG attacks

JPEG compression is a common and practical attack model for data embedded in digital images, and arises naturally when compression is used to fit the modified image into communication links or memories of varying capacities. JPEG operates on 8×8 blocks of DCT coefficients. The $(i,j)^{th}$ coefficient, x_{ij}, of each block is quantized uniformly with step size Δ_{ij}, taken from an 8×8 JPEG

quantization matrix. The JPEG quantization matrix is determined by the level of desired compression, or quality factor (QF). Quality factors range from 0 to 100, 100 corresponding to no compression, and 75 being a typical amount of compression.

It is well understood that high frequency distortion in images is less perceptible than its low frequency equivalent. Accordingly, JPEG uses finer quantizers for low frequency coefficients. Depending on the quality factor, most mid to high coefficients are quantized so coarsely that their reconstruction value is zero. The quantized coefficients are subsequently run-length and entropy encoded.

For a DCT domain scalar quantization embedding scheme to survive such an attack, the spacing between a '0' and '1' quantizer must be at least Δ_{ij}. This can be guaranteed by adopting the JPEG quantization matrix and using odd multiples of Δ_{ij} to embed a '1' and even multiples of Δ_{ij} to embed a '0', when hiding in coefficient x_{ij}. Thus, one can tune a DCT coefficient quantization embedding scheme to guarantee survival of the signature data for a given amount of JPEG compression. In fact, data hidden in this fashion will be robust to all JPEG compression attacks lesser than or equal to that of the design quality factor.

2.3. Performance penalty under AWGN attack

While our scalar quantization based scheme is well matched to JPEG compression attacks, it does incur a substantial penalty for the worst-case AWGN attack. We quantify this in the context of an i.i.d. Gaussian host as follows. Letting D_1 and D_2 denote the mean squared embedding induced distortion and mean squared attack distortion, the hiding capacity with AWGN attack is given by [1, 7] $C = \frac{1}{2}\log(1 + \frac{D_1}{D_2})$, in the small D_1, D_2 regime that typical data hiding systems operate. We compare this "vector capacity" (termed thus because the optimal strategy involves vector quantization of the host) to that of a scalar quantizer embedding scheme with hard decision decoding. Letting Δ denote the distance between a '0' and '1' quantizer, the variance of the quantization error is approximately $D_1 = \frac{(2\Delta)^2}{12}$. The probability of bit error is given by

$$p_e = 2Q(\frac{\Delta}{2\sqrt{D_2}}) = 2Q(\frac{\sqrt{3}}{2}\sqrt{\frac{D_1}{D_2}}) \quad (1)$$

where Q denotes the complementary cumulative distribution function of a standard Gaussian random variable. The capacity of a binary symmetric channel with transition probability p_e is given by [8] $C_{bsc} = 1 - H(p_e)$, where $H(p) = -p\log p - (1-p)\log(1-p)$.

As with the vector capacity, the scalar capacity is solely a function of $\frac{D_1}{D_2}$, the distortion to noise ratio (DNR). Figure 1 shows roughly a $5dB$ loss due to the suboptimal encoding strategy used here, a gap that can be closed using soft decisions and vector quantization.

3. PERCEPTUAL EMBEDDING CRITERIA: TWO APPROACHES

In the previous section we described how coefficients are quantized to embed information bits. Next we decide which coefficients should be used for embedding. This will have a significant effect on the perceptual quality of the embedded image. We have devised two such approaches – (i) entropy thresholding and (ii) selectively embedding in coefficients. Both use some criterion to decide which coefficients should be used to embed so that the perceptual quality of the host image is preserved. Thus, the amount of data hidden is adapted to the characteristics of the host.

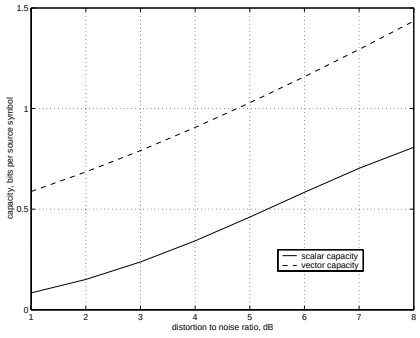

Fig. 1. Gap between scalar and vector quantizer data hiding systems.

3.1. Entropy thresholding

For a quantizer hiding scheme based on JPEG, one observes less distortion when embedding in low frequency DCT coefficients because of JPEG's finer quantization in this range. However, JPEG uses predictive encoding for the DC coefficients, x_{00}, of successive blocks, so the additive uniform noise model does not apply. Furthermore, distortion induced in these coefficients would not be localized to their 8×8 block. We therefore do not use these to embed data.

Next, we computed the 2-norm entropy of each 8×8 block as follows,

$$E = \sum_{i,j} \|x_{ij}\|^2, \quad (i,j) \neq (0,0) \quad (2)$$

Only those coefficient blocks whose entropy exceeds a predetermined threshold are used to hide data. Likewise, the decoder checks the entropy of each 8×8 block to decide if data has been hidden. The threshold entropy is determined by the desired embedding rate (or allowable distortion) for a particular image.

In general, compression will decrease the entropy of the coefficient blocks. Therefore, it is necessary to check that the entropy of each block used to embed information, compressed to the design quality factor, still exceeds the threshold entropy. If a particular block fails this test, we keep it as such, and embed the same data in the next block that passes the test

3.2. Selectively embedding in coefficients

The above thresholding scheme uses an entropy criterion to determine when to embed in a DCT block. We can take this idea one step further. Instead of embedding in a fixed number of coefficients in qualifying blocks, we now decide to embed information on a coefficient by coefficient basis. In this manner, we embed precisely in those coefficients that cause minimal perceptual distortion.

Coefficients that are not quantized to zero by the design JPEG quantizer are used to embed information. Quantization embedding is performed as usual when embedding a '1'. If a '0' is to be embedded, the coefficients are quantized to even reconstruction values. However, if this results in quantization to zero, we leave

it as such, and the same '0' is embedded in the next coefficient satisfying the non-zero criterion. The decoder simply disregards all coefficients that quantize to zero. Otherwise, decoding is performed as usual.

Selecting non-zero coefficients for embedding minimizes the perceptual distortion incurred. Many image coefficients are very close to zero once divided by the JPEG quantization matrix, and would be quantized to zero upon JPEG compression. To embed '1' in such coefficients introduces a large amount of distortion relative to the original coefficient size, a factor which seems to be perceptually important. This is avoided by choosing not to use zeros for embedding.

4. RESULTS

4.1. Entropy thresholding

The entropy thresholding scheme was implemented to withstand JPEG compression at various quality factors. A "just noticeable" criterion for embedding induced distortion was used to determine the entropy threshold and number of low frequency coefficients used per qualifying block with a 512×512 Lena test image. Table 1 shows the compression (bits per pixel), number of embedded bits, distortion to noise ratio (DNR), and embedding rate results at each QF. Note, our D_1 and D_2 denote the mean squared embedding induced distortion and mean squared attack distortion, respectively. Figure 2 shows the compressed, hidden Lena image at 50 and 75 quality factors. The performance of this scheme is severely degraded at QF=25. Note that decoding of the embedded data is perfect for all JPEG attacks lesser than or equal to the design QF.

By confining attention to JPEG attacks, we are able to achieve large embedding rates without employing any error correction coding. Our empirical results cannot be compared directly with the capacity estimates in [3], since the latter were derived assuming that both the hider and the attacker use an optimal strategy (forming a so-called saddlepoint solution for the data hiding game considered there), whereas we use a suboptimal hiding strategy optimized for a suboptimal (JPEG compression) attack. In principle, therefore, our capacity can be smaller or larger than the estimates in [3]. As it happens, for $D1/D2 \approx 0.5$,[1] we are able to embed at a rate of 0.132 bits per pixel in Lena against a JPEG attack, which is significantly larger than the corresponding saddlepoint capacity estimate of 0.04 bits per pixel in [3] (which corresponds to optimal embedding for a worst-case AWGN attack).

As expected, when the JPEG attack is replaced by an AWGN attack inducing the same distortion, the performance of our schemes deteriorates. Figure 3 plots, for both the entropy thresholding and coefficient based data hiding schemes, the bit error rate (BER) versus DNR for an AWGN attack. The uncoded BER is significant, which shows that, while our uncoded, hard-decision based, systems are ideally matched to JPEG attacks, error correction coding must be employed in order to handle other additive attacks. In such a setting, we anticipate that achieving capacity will require use of more sophisticated vector quantization based schemes for embedding, as well as powerful codes with soft decision decoding.

[1] Here D_1 denotes the average mean squared error induced by the hider, and D_2 the average mean squared error induced by the attacker. The notation differs from that in [3], where D_2 denotes the sum of the mean squared errors induced by both the hider and attacker.

(a) 0.66 bpp (QF=50) (b) 1.04 bpp (QF=75)

Fig. 2. Entropy thresholding scheme

QF	compression (bpp)	# bits	DNR (dB)	rate (bpp)
25	0.42	4,970	2.9	0.019
50	0.66	15,344	3.8	0.059
75	1.04	34,460	6.5	0.132

Table 1. Performance of thresholding scheme.

4.2. Selectively embedding in coefficients

The coefficient based data hiding scheme was implemented to withstand JPEG compression at different quality factors. Table 2 has the size of the JPEG attacked composite image in bpp, total number of embedded bits, DNR, and embedding rate at each QF for the Lena test image. Figure 4 shows the original and compressed composite Lena images for the various quality factors. Decoding is perfect for all JPEG attacks lesser than or equal to the design QF.

The coefficient scheme operates in the high DNR regime because it does not use an entropy criterion to discriminate between 8×8 DCT blocks and actually does most of JPEG's work, thus allocating minimal power to the attack channel. Figure 3 shows the BER for an AWGN attack with D_1 as given in the QF=75 implementation. Again, channel codes would be used to deal with the degraded performance under an AWGN attack, with a corresponding loss in rate.

5. CONCLUSIONS

The key contribution of this paper is the use of perceptual criteria for embedding in images, in conjunction with the quantization based embedding prescribed by information theory in the context of simple host models. Both of our embedding methods are highly optimized for JPEG compression attacks, which enables them to offer excellent performance without error correction coding. In particular, the capacity obtained without coding under JPEG attack using our schemes is comparable to the capacity estimates

QF	compression (bpp)	# bits	DNR (dB)	rate (bpp)
25	0.38	11,045	26.2	0.042
50	0.60	18,730	22.5	0.071
75	0.94	29,871	19.0	0.114

Table 2. Performance of coefficient scheme.

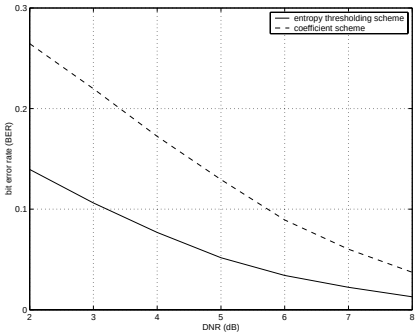

Fig. 3. BER for AWGN attack

under AWGN attacks in [3], attaining which in general would presumably require complex embedding and decoding schemes.

It is relatively straightforward to add standard error correction (including the use of soft decisions and iterative decoding [5]) to make our methods robust to non-JPEG attacks, such as the AWGN attack or wavelet compression. Another extension is the use of vector quantization techniques such as those in [6]. Our use of perceptual criteria to determine the embedding locations (rather than only statistical criteria which may specify which frequency bands to embed in), however, result in a new requirement in terms of error correction coding. Since the decoder must now decide which blocks or coefficients the data is hidden in, our methods are potentially vulnerable to insertion and deletion errors. Thus, an important direction for future work is the incorporation of insertion and deletion codes [9], in addition to standard error correction into our hiding schemes.

Our coefficient embedding scheme demonstrates superior performance at high JPEG compression. This scheme minimizes the power allocated to the JPEG attack channel, thereby operating in a high DNR regime. Another key advantage of this scheme is that it adapts the amount of data embedded to the characteristics of the host image (i.e the rate of this scheme is limited by the number of non-zero scaled DCT coefficients). However, for certain highly textured images, in which we were able to embed at very high rates ($> 35,000$ bits per image), the decoder suffered from insertion and deletion errors of about 0.2%.

The entropy thresholding scheme allows the data hider greater flexibility between embedding rate and induced distortion. On average, compression will lower the entropy of the embedded image. In a few cases, the entropy of a coefficient block is actually increased, causing the decoder to look for data in unused blocks. We have observed less than 1% inserted/deleted blocks in the test images. However, this problem is exacerbated when the attack quality factor is mismatched to that of the encoder.

While hiding based on the JPEG quantization matrix was a convenient choice for illustrating our ideas, the performance of our hiding schemes can be improved by using a more powerful compression mechanism. For example, hiding data in the wavelet domain promises to be a robust scheme that would survive a wavelet compression attack by design, and likely survive JPEG compression with small error rates. Thus, applying our results to this domain is an avenue for future work.

(a) Original Lena (b) 0.94 bpp (QF=75)

(c) 0.59 bpp (QF=50) (d) 0.38 bpp (QF=25)

Fig. 4. Coefficient based scheme

6. REFERENCES

[1] M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[2] S. I. Gel'Fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, Jan. 1979.

[3] P. Moulin and M. K. Mihcak, "The data-hiding capacity of image sources," Preprint, June 2001.

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

[5] M. Kesal, M. K. Mihcak, R. Koetter, and P. Moulin, "Iteratively decodable codes for watermarking applications," in *Proc. 2nd Int. Symp. on Turbo Codes and Related Topics*, Sept. 2000.

[6] J. Chou, S. S. Pradhan, and K. Ramchandran, "A robust optimization solution to the data hiding problem using distributed source coding principles," in *Proceedings of Conference on Information Sciences and Systems (CISS)*, Mar. 2000.

[7] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," Preprint, Jan. 2001.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[9] M. C. Davey and D. J. C. Mackay, "Reliable communication over channels with insertions, deletions, and substitutions," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 687–698, Feb. 2001.