ESTIMATION OF OPTIMUM CODING REDUNDANCY AND FREQUENCY DOMAIN ANALYSIS OF ATTACKS FOR YASS - A RANDOMIZED BLOCK BASED HIDING SCHEME

A. Sarkar, L. Nataraj, B. S. Manjunath and U. Madhow

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

ABSTRACT

Our recently introduced JPEG steganographic method called Yet Another Steganographic Scheme (YASS) can resist blind steganalysis by embedding data in the discrete cosine transform (DCT) domain in randomly chosen image blocks. To maximize the embedding rate for a given image and a specified attack channel, the redundancy factor used by the repeat-accumulate (RA) code based error correction framework in YASS is optimally chosen by the encoder. An efficient method is suggested for the decoder to accurately compute this redundancy factor. We also show experimentally which DCT coefficients are better suited for hiding and detection under various attacks. The effectiveness of YASS for robust steganography is demonstrated for certain attacks.

Index Terms— randomized hiding, steganography, steganalysis, repeat-accumulate code, redundancy factor

1. INTRODUCTION

Steganography is the art of secure communication and is aimed at resisting steganalysis, which is the art of detecting the existence of the secret communication. Many popular steganographic methods (such as F5[1], model-based steganography[2], OutGuess [3]) can now be detected using *blind* steganalysis schemes (such as [4, 5]). These detection schemes are blind in the sense that they are applicable to a variety of steganographic methods. The detection methods use feature vectors that incorporate higher-order joint statistics - the most successful method being the *self-calibration process* [4] that can reliably estimate the cover statistics from the available stego signal.

Our recently proposed steganographic method called *Yet Another Steganographic Scheme* (YASS) [6] resists the above mentioned blind steganalysis schemes, albeit with a relatively low hiding capacity. The importance of YASS for secure steganography was independently verified in [7]. In YASS, a grid of bigger blocks (of size B>8 where B is called the *big-block size*) is formed from which an 8×8 block is chosen randomly to hide data. The steganalyst computes the image features (self-calibration process) assuming a regular 8×8 grid and hence gets out-of-sync with the randomly chosen blocks used for hiding, resulting in stego features that do not directly capture the modifications done to the image after hiding. Since the embedding grid does not coincide with the JPEG grid, the JPEG compression after hiding introduces many errors - the embedded data can still be recovered using error correction codes of suitable redundancy. Due to the high rate of erasures in our data hiding framework [8], repeat-accumulate (RA) codes [9] are used.

An active adversary is considered here who, after intercepting the signal in a channel, may introduce some mild distortion, while maintaining the signal's perceptual transparency. For the active steganographic framework, the embedded data has to be perfectly decoded even after these attacks - therein lies the utility of the RA coding framework in YASS.

2. PROBLEM MOTIVATION AND CONTRIBUTIONS

Maximizing the Hiding Rate Using Optimum Coding Redundancy at the Encoder and the Decoder

With an increased redundancy factor (q) in the RA framework, the hiding rate decreases while the robustness to channel distortions increases. The hiding rate is maximized if the encoder uses the minimum q that guarantees zero bit error rate (BER) for a given image, *a known attack channel* and hiding parameters that ensure statistical security. This redundancy factor is referred to as q_{opt} in subsequent discussions. The decoder knows the embedding method and the error correction code (RA) used, but not the q used at the encoder. We present an efficient method by which the decoder can correctly estimate the q used by the encoder.

Frequency Domain Analysis of Hiding and Detection for Different Attacks

The channel distortions can be due to a variety of attacks, apart from just JPEG compression. The hiding rates computed for the different classes of attacks, with varying attack

This research is supported in part by a grant from ONR # N00014-05-1-0816.

levels, indicate the robustness of YASS to these attacks. We also study how the detection accuracy and the embedding rate vary with the choice of the frequency band used for hiding. Thus, the frequency domain analysis helps the steganographer and the steganalyst to decide on the best bands for hiding and detection, respectively, for a given attack.

3. FINDING THE OPTIMUM CODING REDUNDANCY FACTOR AT ENCODER AND ESTIMATING IT ACCURATELY AT DECODER

The serial concatenated turbo (RA) code based error correction is used in our data hiding setup - Fig. 1 shows the whole framework except for the iterative decoding part at the RA decoder. Let the total number of possible hiding locations in the image be ℓ. Using a redundancy factor of q, the maximum number of embeddable databits, denoted by N, equals $\lfloor \ell/q \rfloor$. The encoder repeats the N-bit data sequence u, as a whole, q-times (1), instead of repeating each bit q times. As is shown later, this makes it easier to compute q at the decoder using the auto-correlation (8) of the RA-encoded sequence.

Steps involved in mapping from u to y at the encoder

$$[r_{(i-1)N+1} r_{(i-1)N+2} \ldots r_{iN}] = [u_1 u_2 \ldots u_N],\ 1 \leq i \leq q \quad (1)$$

$$x = \pi(r),\ \text{where } \pi \text{ is the interleaver function} \quad (2)$$

$$y_1 = x_1,\ y_n = y_{n-1} \oplus x_n,\ 2 \leq n \leq Nq \quad (3)$$

After data embedding, we get a ternary sequence z of $\{0, 1, e\}$ based on what is actually embedded, where e denotes an erasure (Fig. 1). When a quantized discrete cosine transform (DCT) term in the image lies in the range [-0.5,0.5], an erasure occurs - this maintains perceptual transparency [8]. For DCT terms of higher magnitude, every DCT term is quantized to the nearest odd/even integer to embed 1/0, respectively. The ternary sequence obtained from the hiding locations in the noisy received image, decoded using the same principles used while embedding by the encoder, is called $\hat{y}$.

At the encoder side, the sender transmits a sequence u, embeds the RA-encoded sequence y in the image, subjects it to known attacks and finally obtains $\hat{y}$ from the image. Thus, by simulating the exact attack channel, the 2×3 transition probability matrix, $p(\hat{y}|y)$ can be computed. The capacity $\mathcal{C}$, for the channel that maps y to $\hat{y}$, is obtained by maximizing the mutual information $I(Y, \hat{Y})$ between the sequences y and $\hat{y}$ (4) - a discrete memoryless channel is assumed here.

$$\mathcal{C} = \max_{p(y)} I(Y, \hat{Y}) = \max_{p(y)} \sum_{y} \sum_{\hat{y}} p(y, \hat{y}) \log \left\{ \frac{p(y|\hat{y})}{p(y)} \right\} \quad (4)$$

From a capacity perspective, the minimum redundancy factor needed for perfect data recovery, *assuming an ideal channel code*, is $q_{min} = \lceil \frac{1}{\mathcal{C}} \rceil$. Thus, the minimum possible value of q_{opt} (q needed for perfect data recovery even after channel distortions) for the RA code is q_{min}. The sender simulates the decoder and attempts to recover the embedded databits by varying q. An upper limit (q_{max}) is set on the maximum redundancy factor to be used. Thus, the search for q_{opt}, needs to be done in the range $[q_{min}, q_{max}]$ - it will need at most $\log_2(q_{max} - q_{min})$ searches. *It is assumed here that the encoder knows the exact attack*, allowing it to compute q_{opt} precisely. In practice, the range of attacks may be known - the encoder can then design q_{opt} based on the worst-case attack.

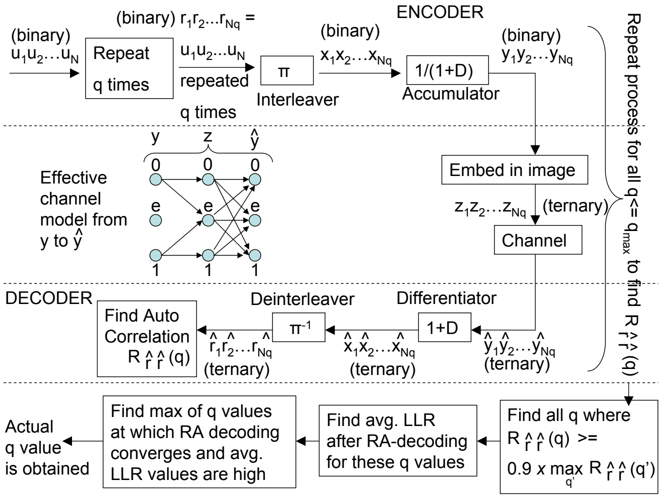

ENCODER
(binary) $u_1u_2\ldots u_N$
Repeat q times
(binary) $r_1r_2\ldots r_{Nq}$ = $u_1u_2\ldots u_N$ repeated q times
π
Interleaver
(binary) $x_1x_2\ldots x_{Nq}$
1/(1+D)
Accumulator
(binary) $y_1y_2\ldots y_{Nq}$
Effective channel model from y to $\hat{y}$
y
z
$\hat{y}$
0
e
1
Embed in image
$z_1z_2\ldots z_{Nq}$ (ternary)
Channel
Repeat process for all q<= q_{max} to find $R_{\hat{r}\hat{r}}(q)$
DECODER
Find Auto Correlation $R_{\hat{r}\hat{r}}(q)$
Deinterleaver
π^{-1}
$\hat{r}_1\hat{r}_2\ldots\hat{r}_{Nq}$ (ternary)
Differentiator
1+D
$\hat{x}_1\hat{x}_2\ldots\hat{x}_{Nq}$ (ternary)
$\hat{y}_1\hat{y}_2\ldots\hat{y}_{Nq}$ (ternary)
Find all q where $R_{\hat{r}\hat{r}}(q)$ >= 0.9 x $\max_{q'}$ $R_{\hat{r}\hat{r}}(q')$
Find avg. LLR after RA-decoding for these q values
Find max of q values at which RA decoding converges and avg. LLR values are high
Actual q value is obtained

Fig. 1. The data hiding system using RA-code based error correction, where q is efficiently estimated at the decoder

In (5) and also later in (7), it is assumed that the output of $\oplus$ is an erasure if any of the input bits is erased.

Steps involved in mapping from $\hat{y}$ to $\hat{r}$ at the decoder

$$\hat{x}_1 = \hat{y}_1,\ \hat{x}_n = \hat{y}_n \oplus \hat{y}_{n-1},\ 2 \leq n \leq Nq \quad (5)$$

$$\hat{r} = \pi^{-1}(\hat{x}),\ \text{where } \pi^{-1} \text{ is the deinterleaver function} \quad (6)$$

Since the decoder knows the hiding method and assuming that the image size is not altered by the attacks, it can compute ℓ - the total number of possible hiding locations. Let the actual q value used by the encoder be q_{act}. If the decoder assumes q=q', the number of databits equals $\lfloor \ell/q' \rfloor$. In an ideal case, the sequence $\hat{r}$ will be exactly equal to r, where r consists of the input message sequence u, repeated as a whole. Thus, if $\hat{r}$ is circularly shifted by the assumed input message length $\lfloor \ell/q' \rfloor$, the normalized correlation between the original and the shifted sequences $R_{\hat{r},\hat{r}}(q')$ (8) will be very high if q'=q_{act}. In (7), $b(q')$ is the sequence obtained after performing element-wise $\oplus$ between the original and shifted sequences, where shift $k = \lfloor \ell/q' \rfloor$. $R_{\hat{r},\hat{r}}(q')$ (8) is the fraction of 0's in $b(q')$ (matches in two corresponding bits after $\oplus$ result in 0's), without considering the erasures.

$$b(q') = (\{\hat{r}_1 \ldots \hat{r}_{kq'}\} \oplus \{\hat{r}_{kq'-k+1} \ldots \hat{r}_{kq'} \hat{r}_1 \ldots \hat{r}_{kq'-k}\}) \quad (7)$$

and shift $k = \lfloor \ell/q' \rfloor$ is the assumed number of databits

$$R_{\hat{r},\hat{r}}(q') = \frac{\text{Number of 0's in } b(q')}{\text{Number of 0's and 1's in } b(q')} \quad (8)$$

$$\mathcal{Q}_{top} = \left\{ q' : R_{\hat{r},\hat{r}}(q') >= 0.9 \times (\max_{q' \leq q_{max}} R_{\hat{r},\hat{r}}(q')) \right\} \quad (9)$$

The correlation is also high when the shift equals a multiple of the actual message length, i.e. q'=$q_{act}/m,\ m \in \mathbb{Z}^+$. Apart from the correlation peaks at q_{act} and its sub-multiples, other peaks may occur due to errors and erasures. In the experiments, the set of q values, $\mathcal{Q}_{top}$ (9), at which the correlation exceeds 90% of the maximum $R_{\hat{r},\hat{r}}$ value, are selected - the 90% cutoff was empirically determined. The turbo decoder is then run for these q values and the log-likelihood ratios (LLR) are computed for the extracted databits in each case. It is seen that due to a noisy channel, decoding may converge (two consecutive iterations produce the same output sequence) at values other than $q_{act}/m,\ m \in \mathbb{Z}^+$. However, the LLR value, averaged over the databits, is high only when perfect decoding occurs. It is seen that the maximum average LLR values occur only at q_{act} and its sub-multiples. Thus, the solution is to consider the maximum of these q values as q_{act}, as shown in Fig. 1. This method of estimating q for RA encoding is found to work even at high erasure rates ($\geq 95\%$).

Observations about the q-estimation method

- The use of auto-correlation based peaks reduces the search space for q while the average LLR-based measure, followed by taking the maximum, helps to identify the actual q.
- For our experiments, the search range for q was $[2, 50]$.
- Though the correlation in $\hat{r}$ is used for q-estimation, this correlation is not detectable by an adversary; $\hat{r}$ is obtained from $\hat{y}$ only after applying the deinterleaver (π^{-1}) - the key to generate π^{-1} is not known to an adversary.
- The q-estimation method is generic enough to be used for any hiding scheme which uses RA-q based error correction.

4. FREQUENCY DOMAIN ANALYSIS FOR VARYING ATTACKS AND ATTACK LEVELS

4.1. Capacity Estimation Along Individual Coefficients

For different attacks and attack levels, hiding is performed along individual bands and the respective hiding capacities are computed using (4). A *big-block size* [6] B=9 and a design quality factor QF_h of 70 are used for hiding. For capacity estimation, we use 500 images, from the MM270K database[1]. The average hiding rate is reported in terms of the bpnc (bits per non-zero coefficient) in Table 1.

In natural images, the lower frequencies are generally larger in magnitude than the mid frequencies and have more coefficients eligible for hiding. From Fig. 2(a), DCT coefficients $\{1, 6\}$ are nearest to the DC term, followed by $\{2, 7, 11\}$, and then $\{3, 8, 12, 15\}$. Hence, the hiding capacity is generally maximum for $\{6, 1\}$, followed by $\{2, 7, 11\}$, as seen from Fig. 2(b)-(h).

Since YASS is a JPEG steganographic method, the images, after the various attacks as shown in Fig. 2(b)-(h), are JPEG-compressed at an output quality factor of QF_o. To study the effect of just the JPEG-2000 based compression (Fig. 2(b)), a QF_o of 99 is used along with it to minimize the JPEG-based distortion. For JPEG2000, the ratio between the number of bits representing the compressed and original images is denoted by CR - higher CR denotes less severe compression.

For more severe JPEG and AWGN attacks (Fig. 2(c)-(d)), $\{6, 11\}$ have higher capacity, followed by $\{1, 7, 15, 2\}$. For more severe gamma variation ($|\gamma - 1| \geq 0.05$), the different bands have almost the same capacity with $\{6, 11\}$ doing marginally better than $\{7, 1, 2, 12\}$ (Fig. 2(h)). The individual hiding capacities are much lower for the averaging, median filtering and Gaussian blur based attacks, compared to the others (Fig. 2(e)-(g)). A mild attack using any of these methods can make secure hiding at practical rates impossible using YASS. Hence, for the hiding rate comparison in Table 1, these attacks are not considered.

In [6], it is shown that for QF_o=75, YASS-based hiding using B=9 and QF_h=70 is statistically undetectable. In Table 1, the effective attack consists of a (JPEG2000/AWGN/gamma variation) based attack, followed by JPEG compression at QF_o=75. It is seen that the reduction in bpnc over just the JPEG attack is about 10% for a JPEG2000 attack with CR=0.15, AWGN with SNR=45dB, and gamma variations with $|\gamma - 1| = 0.02$.

Table 1. Hiding rate comparison for various attacks - each attack is followed by JPEG compression using QF_o=75. The bpnc using only the JPEG attack at QF_o=75 is 0.1087.

JPEG2000		AWGN		Gamma:$\gamma < 1$		Gamma:$\gamma > 1$	
CR	bpnc	SNR(dB)	bpnc	γ	bpnc	γ	bpnc
0.10	0.0698	40	0.0741	0.95	0.0686	1.05	0.0675
0.15	0.0978	45	0.0958	0.98	0.0959	1.02	0.0942
0.20	0.1073	50	0.1042	0.99	0.1028	1.01	0.1027

4.2. Detection Results for Individual DCT Coefficients

We conduct the steganalysis experiments on 4500 JPEG images, from the MM270K database, compressed using QF=75. Half of the images are used for training and the other half for testing. We use a support vector machine (SVM) based classifier for steganalysis, where the SVM is trained using Pevny and Fridrich's 274-dimensional feature that merges Markov and DCT features [4]. The probability of classifying a test image correctly as cover or stego - the detection accuracy P_d ($P_d \approx 0.5$ implies undetectable hiding, and as the detectability improves, P_d increases towards 1) is obtained for different attacks and using different frequency bands. Both the cover and stego images suffer the same attacks and hiding occurs in the same band for the training and test sets. For hiding, B=9 and QF_h=50 are used - a lower design QF is used to magnify the difference in the detection accuracy, across the various bands. For AWGN, $P_d \approx 0.5$ for all the individual bands - hence, it is not included as an attack in Fig. 3.

In general, the bands that are able to hide more ($\{6, 1\}$) should also be better for detection. From Fig. 3, this holds true

[1] downloaded from http://www-2.cs.cmu.edu/yke/retrieval

for JPEG, JPEG2000, and gamma variation attacks. However, for averaging ($\{7,3\}$), unsharp masking($\{8,3\}$), median filtering ($\{3,7\}$) and Gaussian blur($\{3,15\}$) - some mid-frequencies are more detectable than the $\{6,1\}$ coefficients. In the future, we shall consider frequency-based noise models for the various attacks to explain these detection results.

5. CONCLUSION

In this paper, we have shown a method to maximize the hiding rate, by optimally choosing the RA-code redundancy factor at the encoder, followed by independently computing this parameter at the decoder, exploiting the structure of repeat-accumulate encoded sequences. The robustness of YASS to a variety of attacks has been studied and insight is gained into the proper choice of frequency bands, for these attacks, from both the hider and the detector's perspective. Future work shall focus on making YASS more effective against a wider variety of attacks through proper choice of the hiding bands.

References

[1] A. Westfeld, "High capacity despite better steganalysis (F5 - a steganographic algorithm)," in *4th Int. Workshop on Info. Hiding*, 2001, vol. 2137, pp. 289–302.

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

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

[4] T. Pevny and J. Fridrich, "Merging Markov and DCT features for multi-class JPEG steganalysis," in *Proc. of SPIE*, San Jose, CA, 2007, pp. 3–4.

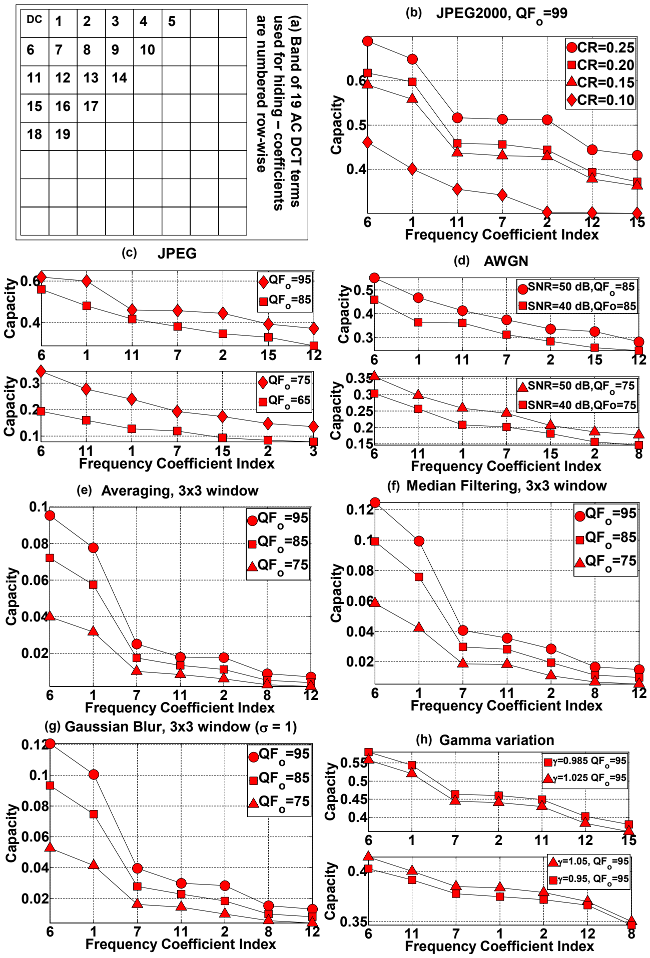
[5] Y. Q. Shi, C. Chen, and W. Chen, "A Markov process based approach to effective attacking JPEG steganography," in *8th Int. Workshop on Info. Hiding*, 2006, pp. 249–264.

[6] K. Solanki, A. Sarkar, and B. S. Manjunath, "YASS: yet another steganographic scheme that resists blind steganalysis," in *9th Int. Workshop on Info. Hiding*, Jun 2007.

[7] J. Kodovsky and J. Fridrich, "Influence of embedding strategies on security of steganographic methods in the JPEG domain," in *Proc. of SPIE*, San Jose, CA, Jan. 2008.

[8] 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.

[9] D. Divsalar, H. Jin, and R. J. McEliece, "Coding theorems for turbo-like codes," in *36th Allerton Conf. on Communications, Control, and Computing*, Sept. 1998, pp. 201–210.



DC	1	2	3	4	5		
6	7	8	9	10			
11	12	13	14				
15	16	17					
18	19						

Fig. 2. Capacity results for hiding in individual frequency bands for varying attacks and attack levels - the DCT coefficient indices (1-19) are explained in the 8×8 grid (Fig. 2(a)). After each attack, the images are JPEG compressed at QF_o.

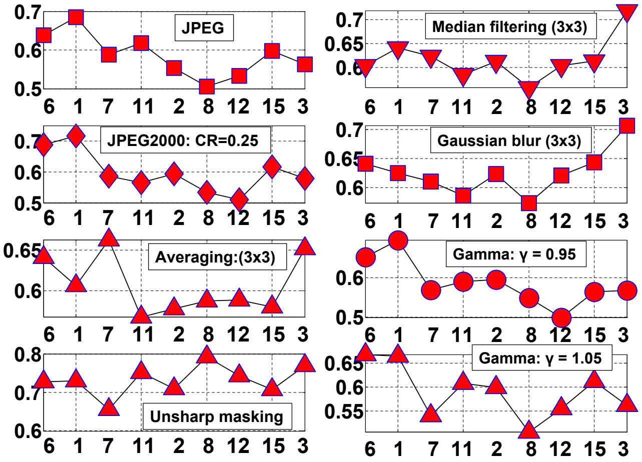


Fig. 3. The detection accuracy (P_d: y-axis) is plotted for individual frequency bands (x-axis) for various attacks - the images are JPEG compressed at QF_o=75 after an attack.