INFORMATION THEORETIC BOUNDS FOR DATA HIDING IN COMPRESSED IMAGES

Mahalingam Ramkumar and Ali N. Akansu *
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
New Jersey Center for Multimedia Research
University Heights, Newark, NJ, 07102

Abstract - We present an information-theoretic approach to obtain an estimate of the number of bits that can be hidden in still images, or, the *capacity of the data-hiding channel*. We show how addition of the message signal in a suitable transform domain rather than the spatial domain can significantly increase the channel capacity. We compare the capacities achievable with different decompositions like DCT, DFT, Hadamard, and subband transforms.

INTRODUCTION

Data hiding or Steganography, is a rapidly growing field with potential applications for copyright protection (watermarking), hiding executables for access control of digital multimedia data, embedded captioning, secret communications, etc. It is therefore of significant interest to have a theoretical estimate of the number of bits that can be hidden in multimedia data. In this paper we provide an information-theoretic approach to estimate the number of bits that can be hidden in still images.

Let $\boldsymbol{I}$ be the original (cover) image, to which a message $\boldsymbol{S}$ (a representation for a few bits of information) is added, such that $\hat{\boldsymbol{I}} = \boldsymbol{I} + \boldsymbol{S}$, the modified image, is *visually indistinguishable* from $\boldsymbol{I}$. The image $\hat{\boldsymbol{I}}$ may typically be subjected to lossy compression, like JPEG; $\tilde{\boldsymbol{I}} = \mathcal{C}(\hat{\boldsymbol{I}})$, where $\mathcal{C}(.)$ denotes the compression / decompression operation. The buried bits in image $\boldsymbol{I}$ are to be extracted from $\tilde{\boldsymbol{I}}$. Under this scenario we would like to know the maximum number of bits that can be buried and recovered from the image with an arbitrarily low probability of error, or in other words, the *capacity of the data-hiding channel*. A block diagram of the data hiding channel is shown in Figure 1(a). $\boldsymbol{S}$ is the message to be transmitted through the channel which has two sources of noise: $\boldsymbol{I}$, the noise due to the cover image, and $\boldsymbol{P}$, the noise due to processing (compression / decompression). $\tilde{\boldsymbol{S}}$ is the "corrupted" message. Note that the receiver does not have access to the cover image.

*This work was partly supported by Panasonic Technologies Inc., Princeton, NJ

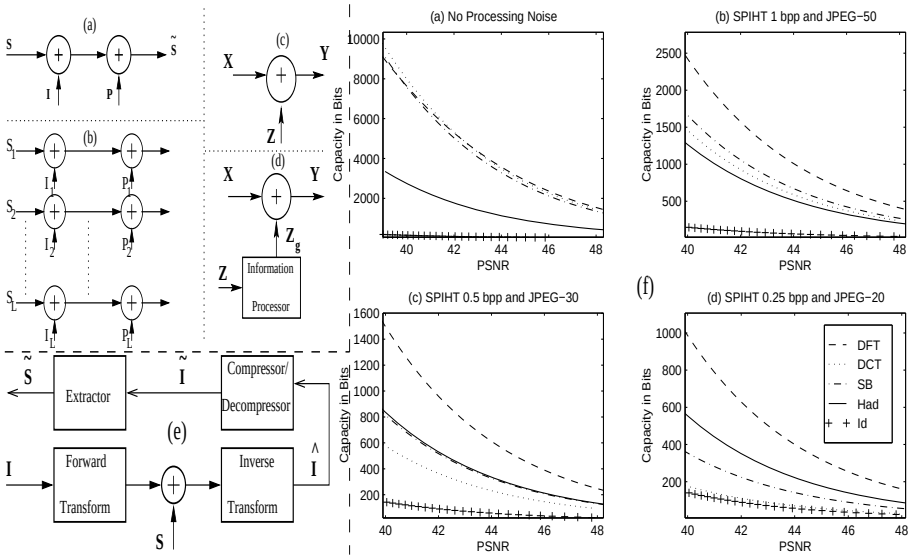

Figure 1: (a) The Data Hiding Channel. (b) Decomposition of the Data-Hiding Channel into L Parallel Channels. (c) A Simple Additive Noise Channel. (d) Channel of (c) Modified to Obtain Equivalent Additive Gaussian Noise. (e) Schematic Block Diagram of Data Hiding / Retrieval (f) Channel Capacities of 64 Band Decompositions for various processing noise scenarios.

Figure 1(e) is a block diagram of a typical data-hiding scheme. In this paper, we use the schematic of Figure 1 (e) to determine the capacity of the data hiding channel. The first attempt in obtaining an information theoretic view-point of the capacity of the data-hiding channel was reported in [1]. However, the attempt was limited in scope, in that it was implicitly assumed that the message is added in the spatial domain. We show how the capacity of the data-hiding channel can be improved by a suitable choice of transform.

CAPACITY OF THE DATA HIDING CHANNEL

Capacity of Additive Noise Channels

Before we consider the data-hiding channel of Figure 1(a), we shall consider a simpler channel shown in Figure 1(c). Here, $\boldsymbol{X} \sim [f_X(x),\, \sigma_x^2]$ is the message to be transmitted, $\boldsymbol{Z} \sim [f_Z(z),\, \sigma_z^2]$ is the additive noise in the channel, and $\boldsymbol{Y} \sim [f_Y(y),\, \sigma_y^2]$ is the received signal at the output of the channel. We shall also assume that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, implying that $\sigma_y^2 = \sigma_x^2 + \sigma_z^2$. The channel capacity is given by [2]

$$\boldsymbol{C} = \max_{f_X(x)} \mathcal{I}(\boldsymbol{X}, \boldsymbol{Y}) = \max_{f_X(x)} h(\boldsymbol{Y}) - h(\boldsymbol{Y}|\boldsymbol{X}) = \max_{f_X(x)} h(\boldsymbol{Y}) - h(\boldsymbol{Z}) \text{ bits.} \quad (1)$$

where $\mathcal{I}(\boldsymbol{X}, \boldsymbol{Y})$, is the *mutual information* between $\boldsymbol{X}$ and $\boldsymbol{Y}$. For a given statistics $f_Z(z)$ and σ_x^2, one should maximize the entropy of $\boldsymbol{Y}$, $h(\boldsymbol{Y}) =$

$-\int f_Y(y)\log_2(f_Y(y))\mathrm{d}y$ (bits), by choosing a suitable distribution $f_X(x)$ of the message $\boldsymbol{X}$. For a given σ_y^2, the maximum value of $h(\boldsymbol{Y}) = \frac{1}{2}\log_2(2\pi\mathrm{e}\sigma_y^2)$ bits is achieved when $\boldsymbol{Y}$ has a normal distribution. For instance, the maximum value of $h(\boldsymbol{Y})$ is achievable if both $f_Z(z)$ and $f_X(x)$ are normally distributed. However, for an arbitrary distribution $f_Z(z)$, and a fixed σ_x^2, it is not immediately obvious what the maximum achievable value of $h(\boldsymbol{Y})$ is. In order to find that, we pass $\boldsymbol{Z}$ through an ideal *information processor* (Figure 1 (d)), which does not alter the amount of information in $\boldsymbol{Z}$, but changes its statistics, to a Gaussian distributed $\boldsymbol{Z}_g$. As the output of the processor has the same entropy as the input, the variance of the output, σ_{zg}^2, can be obtained by solving $h(\boldsymbol{Z}_g) = h(\boldsymbol{Z}) = \frac{1}{2}\log_2(2\pi\mathrm{e}\sigma_{\mathrm{zg}}^2)$ bits. For the purpose of calculating the channel capacity, we can replace $f_Z(z)$ by $N[0, \sigma_{\mathrm{zg}}^2]$;

$$\boldsymbol{C} = \max_{f_X(x)} h(\boldsymbol{Y}) - h(\boldsymbol{Z}_g) = \frac{1}{2}\log_2(1 + \frac{\sigma_x^2}{\sigma_{\mathrm{zg}}^2}) \text{ bits.} \tag{2}$$

Going back to Figure 1(a), as the processing noise is usually a result of many independent operations, we call upon the Central Limit Theorem [3], and assume a Gaussian distribution for the processing noise $\boldsymbol{P}$. The two noise sources in the channel (I of variance σ_i^2 per pixel and P of variance σ_p^2 per pixel), can be substituted with a single Gaussian noise source of variance $\sigma_{\mathrm{ig}}^2 + \sigma_p^2$, where σ_{ig}^2 is the equivalent Gaussian variance of the noise due to the cover image. If σ_s^2 is the energy of the message signal (per pixel), the capacity of the data-hiding channel can be expressed as

$$\boldsymbol{C}_h = \frac{1}{2}\log_2(1 + \frac{\sigma_s^2}{\sigma_{\mathrm{ig}}^2 + \sigma_p^2}) \text{ bits per pixel.} \tag{3}$$

Decomposition into Multiple Channels

In Figure 1 (b) the channel of Figure 1 (a) is decomposed into multiple channels. The decomposition is performed by the Forward and Inverse Transform blocks of Figure 1 (e). The decomposition of the image into L bands results in L parallel channels with two noise sources in each channel. Let $\sigma_{i_j}^2$, $j = 1 \cdots L$ be the variances of the coefficients of each band (or the variances of the image noise in each channel) of the decomposition. Let their corresponding equivalent Gaussian variances be $\sigma_{\mathrm{ig}_j}^2$. If $\sigma_{p_j}^2$ is the variance of the processing noise in the j^{th} channel, then, the total capacity of the L parallel channels is given by

$$\boldsymbol{C}_h = \frac{MN}{2L}\sum_{j=1}^{L}\log_2(1 + \frac{v_j^2}{\sigma_{\mathrm{ig}_j}^2 + \sigma_{p_j}^2}) \text{ bits} \tag{4}$$

for an image of MN pixels. In the above equation, v_j is the *visual threshold* of band j. In other words, v_j^2 is the maximum message signal energy permitted in band j.

Modeling Channel Noise

The cover image (or the image noise $\boldsymbol{I}$) is decomposed into L bands using an orthonormal transform. Let $f_{I_j}(i_j)$ be the distribution of the j^{th} band

with variance $\sigma_{i_j}^2$. Having obtained the variances of the image noise in each channel, the next step is to obtain their equivalent Gaussian variances. This is achieved by plotting a histogram of the coefficients for each band, and calculating the entropy. If Δx is the width of the n bins of the histogram $g(m)$, $m = 1 \cdots n$, and p is the total number of coefficients in the band, the entropy $\mathcal{H}_j$ and the equivalent Gaussian variance $\sigma_{\mathrm{ig}_j}^2$ are obtained as

$$\mathcal{H}_j = -\sum_{i=1}^{n} \frac{g(i)}{p\Delta x} \log_2(\frac{g(i)}{p\Delta x})\Delta x \qquad \sigma_{\mathrm{ig}_j}^2 = \frac{2^{2\mathcal{H}_j}}{2\pi \mathrm{e}}. \tag{5}$$

The image noise in channel (band) j can now be substituted by Gaussian noise of variance $\sigma_{\mathrm{ig}_j}^2$.

Let the noise due to compression in each channel be $\sigma_{p_j}^2$, $j = 1 \cdots L$. As in the Section 2.1, we assume Gaussian distribution for the processing noise in each channel. We obtain $\frac{MNn_i}{L}$ samples of each band from n_i test images of size $M \times N$. Let i_{j_k}, $k = 1, \ldots, \frac{MNn_i}{L}$, be the coefficients of the band j of the decomposition of the original images. Let $\tilde{i}_{j_k}$, $k = 1, \ldots, \frac{MNn_i}{L}$ be the corresponding coefficients of the images subjected to some lossy compression scheme. We obtain the equivalent additive noise in each channel as noise uncorrelated with $\boldsymbol{i}_j$, that would cause the same *reduction in correlation* between $\boldsymbol{i}_j$ and $\tilde{\boldsymbol{i}}_j$. We define the intra-band correlation as

$$\frac{\langle \boldsymbol{i}_j, \tilde{\boldsymbol{i}}_j \rangle}{|\boldsymbol{i}_j||\tilde{\boldsymbol{i}}_j|} = \frac{\langle \boldsymbol{i}_j, (\boldsymbol{i}_j + \boldsymbol{n}_j) \rangle}{|\boldsymbol{i}_j||\boldsymbol{i}_j + \boldsymbol{n}_j|} = \rho_j, \tag{6}$$

where $\boldsymbol{n}_j$ is a vector of Gaussian (zero mean) random variables uncorrelated with $\boldsymbol{i}_j$. Then $\sigma_{n_j}^2 = |\boldsymbol{n}_j|^2$ is the variance of the *equivalent additive noise due to compression*. Or $\sigma_{p_j} = \sigma_{n_j}$. As $\langle \boldsymbol{i}_j, \boldsymbol{n}_j \rangle = 0$, Eq. (6) can be simplified to obtain

$$\sigma_{p_j}^2 = |\boldsymbol{n}_j|^2 = (\frac{1}{\rho_j^2} - 1)|\boldsymbol{i}_j|^2 \tag{7}$$

We obtain the coefficient statistics σ_{i_j} for various decompositions (4×4 to 32×32 size DCT, DFT, Hadamard and 16 to 1024 band uniform subband (wavelet) decomposition using 8-tap Daubechies filter), and σ_{p_j} for JPEG (quality factors 20-75) and SPIHT (bit rates 0.25 to 1 bpp) compression schemes. The $n_i = 10$ test images of size 256×256 included Lena, Baboon, Barbara, Goldhill, Airplane, Peppers and Boats.

Visual Threshold

The *visual threshold* v_j in Eq. (4) however, is highly subjective. As the amount of message signal energy permitted in any band is determined by the visual threshold, different models for visual thresholds would yield different estimates of achievable capacity. To derive the model, we argue that JPEG, at a reasonably good quality factor (like 75) is optimal in distributing the quantization errors amongst the bands, at least with respect to preserving

visual fidelity of the compressed image. Let i_{j_k} be the coefficients of some decomposition of the original images, and $\tilde{i_{j_k}}$ the coefficients of the same decomposition of images that have undergone JPEG-75 compression and decompression. Let $\sigma^2_{q_j}$ be the variance of the quantization error, $\boldsymbol{e}_{q_j} = \tilde{\boldsymbol{i}}_j - \boldsymbol{i}_j$, for each band j. If quantization error (due to JPEG-75) of variance $\sigma^2_{q_j}$ in band j of the decomposition, results in an image that is visually satisfactory, we can argue that addition of message signal of energy $\sigma^2_{q_j}$ in band j, would still render the image $\hat{\boldsymbol{I}}$ of acceptable visual quality. However to maintain the PSNR of $\hat{\boldsymbol{I}}$ between 40-50 dB (so that the $\hat{\boldsymbol{I}}$ is visually indistinguishable from $\boldsymbol{I}$), we choose the visual thresholds as $v_j^2 = K_2\sigma^2_{q_j}$. where $K_2 < 1$. (The average PSNR of JPEG-75 images is only about 35 dB. So a choice of $K_2 = 1$ would yield images $\hat{\boldsymbol{I}}$ of PSNR 35 dB, which might not be acceptable.) Our simulations show that σ_{q_j}s are independent of j. Or in other words, $\sigma_{q_j} = K \forall j$.

Channel Capacity vs Choice of Transform

For the no-processing noise case (or if the processing noise is very low), if we assume that the all channels have the same pdf (such that $K\sigma_{i_j} = K_1\sigma_{\mathrm{ig}_j}$), the channel capacity is given by

$$C_h = \frac{MN}{2L}\sum_{j=1}^{L}\log_2\left(1+\frac{K}{\sigma^2_{i_j}}\right) \approx \frac{MN}{2L}\log_2\left(1+\sum_{j=1}^{L}\frac{K}{\sigma^2_{i_j}}\right). \tag{8}$$

It is obvious that the *minimum* channel capacity is obtained when $\sigma_{i_j} = \sigma \forall j$, or when *no decomposition* is employed. A transform with good energy compaction or higher Transform Coding Gain (GTC) [4] would result in more *imbalance* of the coefficient variances, resulting in increased channel capacity. So DCT and subband transforms would be good decompositions for low processing noise scenarios. However, we should expect that the *reduction in capacity* with *increase in processing noise* to be lower for transforms like Hadamard and DFT, which are unsuitable for compression. While JPEG at low quality is certain to remove almost all the high frequency components of DCT coefficients, it will not affect the high frequency DFT and Hadamard coefficients to the same extent. Thus decompositions unsuitable for compression would in general be more immune to processing noise than decompositions with high GTC.

RESULTS AND CONCLUSIONS

The channel capacities of different 64 Band decompositions (for 256×256 images, or for 65536 pixels), DFT, DCT, subband (SB), Hadamard (Had) and Identity (Id) transformations, are shown in Figure 1 (f) for various processing noise scenarios. For example if the processing noise is from "JPEG-25 and SPIHT 0.5 bpp" it implies we consider the worst of the two cases for each band. This is to ensure that the message survives JPEG-25 or SPIHT at 0.5 bpp. For the DFT decomposition, we use the magnitude DFT coefficients.

Note that this causes a reduction in the number of available channels from L to $L/2+2$, as only $L/2+2$ magnitude coefficients are unique (the magnitudes of $L/2-2$ complex and 4 real coefficients). In addition, this also reduces the message energy available to each channel by a factor of (approximately) half - only half the message signal energy distributed among the $L-4$ complex coefficients is available for detection. But by sacrificing some channels, (or by reducing the degrees of freedom), we obtain smaller noise variances in each channel.

From the plots in Figure 1(f), we see that the bit-rates for all decompositions fall with increased processing noise, as expected. All transformations perform much better than no-transformation or Identity (Id) transformation. DCT and subband decompositions are better than Hadamard for detection of the message signal *when there is no processing noise*, and Hadamard turns out to be better than DCT and Subband transforms when processing noise is high, as Hadamard Transform is more resistant to processing noise. However, it is surprising that the magnitude DFT decomposition is as good or better than DCT and Subband transforms *even when there is no processing noise*. In this case a reduction in the entropy of the image noise is achieved by ignoring the phase of the DFT coefficients. The reduction in entropy is precisely the "information content" in DFT phase. The signal power available for detection is also "divided" between magnitude and phase. So only half the signal power is available for detection. Yet magnitude DFT decomposition performs better than other decompositions because *DFT phase contains disproportionately more information than DFT magnitude!* In addition, DFT is also robust to processing noise as it is not a high GTC transform. It has also been observed that the channel capacities increase with increase in the number of bands (L) of the decomposition. However the increase in capacity is marginal when processing noise is high. Also, for all L (16, 64, 256 and 1024), it has been found that magnitude DFT decomposition performs better than the other decompositions.

References

[1] J. R. Smith and B. O. Comiskey, "Modulation and Information Hiding in Images", Workshop on Information Hiding, University of Cambridge, UK, 30 May - 01 Jun. 1996.

[2] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, Second Edition, John-Wiley and Sons Inc, 1991.

[3] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd Edition, McGraw Hill Inc. 1991.

[4] A. N. Akansu, R. A. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands and Wavelets*, Academic Press Inc. 1992.