ON THE CHOICE OF TRANSFORMS FOR DATA HIDING IN COMPRESSED VIDEO

Mahalingam Ramkumar, Ali N. Akansu and A. Aydın Alatan

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 compressed image sequences. We show how addition of the message signal in a suitable transform domain rather than the spatial domain can significantly increase the data hiding capacity. We compare the data hiding capacities achievable with different block transforms and show that the choice of the transform should depend on the robustness needed. While it is better to choose transforms with good energy compaction property (like DCT, Wavelet etc.) when the robustness required is low, transforms with poorer energy compaction property (like Hadamard or Hartley transform) are preferable choices for higher robustness requirements.

1. INTRODUCTION

Data hiding or Steganography, is a rapidly growing field with potential applications for copyright protection (watermarking), hiding executables (*e.g.*, 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 video sequences. In Ref. [1] we obtained estimates of the data hiding capacity for compressed still images for JPEG and SPIHT compression schemes. In this paper we extend our work to video sequences employing MPEG-2 compression at different bit rates.

2. PROBLEM STATEMENT

Let I_k be an original frame of the video sequence, to which a message signal S_k (a representation for a few bits of information) is added, such that

$$\hat{\mathrm{I}}_k = \mathrm{I}_k + \mathrm{S}_k, \tag{2.1}$$

the modified frame, is *visually indistinguishable* from I_k. The frames $\hat{\mathrm{I}}_k$, $k = 1 \cdots N_f$ would then be subjected to lossy compression (MPEG). Let $\tilde{\mathrm{I}}_k = \mathcal{C}(\hat{\mathrm{I}}_k)$, where $\mathcal{C}(.)$ denotes the compression / decompression operation. The buried bits in image I_k are to be extracted from $\tilde{\mathrm{I}}_k$. Under this scenario we would like to know the maximum number of bits that can be buried and recovered from an image frame 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. S is the message to be transmitted through the channel which has two sources of noise; I, the noise due to the original frame, and P, the noise due to processing (compression / decompression). $\tilde{\mathrm{S}}$ is the "corrupted" message.

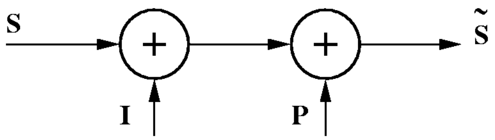

Figure 1: The Data Hiding Channel.

It is relevant to point out here, that data hiding schemes can be broadly classified into two categories. The first category is called *cover image escrow* hiding techniques, where, the original image (or frame) is needed for extracting the hidden information, (for example, see Ref. [2]). In the second category, we have the oblivious detection techniques [3], or techniques for which the original frame is not required for extraction of the hidden message. For the schemes in the first category, there is only one source of noise - due to processing, as the image noise can be subtracted before extracting the hidden information. We can expect such schemes therefore to have higher capacity than the oblivious detection schemes. However the schemes in the first category are of limited use. For most data hiding applications, the receiver does not have access to the original image or frame.

3. CAPACITY OF THE DATA HIDING CHANNEL

In Figure 1, the two (independent) noise sources in the channel $I \sim [f_I(i),\, \sigma_i^2]$, and $P \sim [f_P(p),\, \sigma_p^2]$, can be substituted with a single Gaussian noise source as follows. We first obtain the differential entropy,

$$\mathcal{H} = h(\mathrm{I}) = -\int f_I(i) \log_2(f_I(i)) \mathrm{d}i \text{ bits}, \tag{3.2}$$

and then obtain the variance σ_{ig}^2 of the *equivalent Gaussian noise*, which has the same entropy as I. It is well known that the Gaussian distribution has the lowest variance for a given entropy. For the purpose of calculating the capacity of the channel, we can now replace $I \sim [f_I(i),\, \sigma_i^2]$ with $I_g \sim \mathcal{N}[0, \sigma_{\text{ig}}^2]$. As the processing noise is usually a result of many independent operations, we shall call upon the Central Limit Theorem [4] and assume Gaussian distribution for the processing noise P. The two independent noise sources in the channel can now be substituted by a single Gaussian noise source of variance $\sigma_{\text{ig}}^2 + \sigma_p^2$.

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

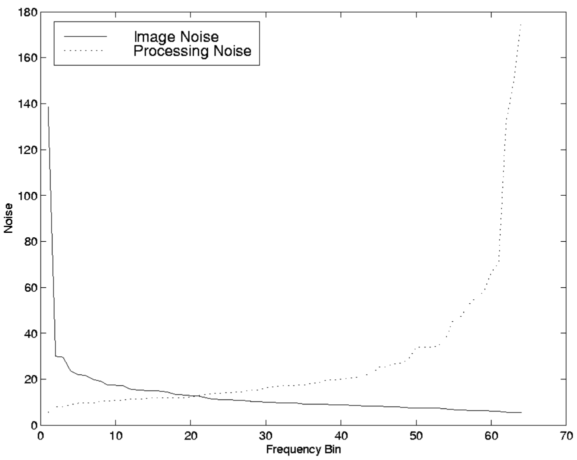

Figure 2: Typical Frequency Distribution of Image and Processing Noise

If σ_s^2 is the energy of the message signal S, then the channel capacity is given by [1, 5]

$$\mathrm{C}_h = \frac{1}{2}\log_2(1 + \frac{\sigma_s^2}{\sigma_{\mathrm{ig}}^2 + \sigma_p^2})\ \text{bits}. \tag{3.3}$$

Typically, the image noise I, (image pixels) is uniformly distributed. The (differential) entropy, $h(I)$, and the equivalent Gaussian variance σ_{ig}^2 of a uniformly distributed random variable I with variance σ_i^2, are given by [5]

$$h(I) = \frac{1}{2}\log_2(12\sigma_i^2)\ \text{bits, and}\ \sigma_{\mathrm{ig}}^2 = \frac{12}{2\pi \mathrm{e}}\sigma_i^2. \tag{3.4}$$

In order to be more explicit, let us derive the capacity of the data hiding channel quantitatively. We would expect the variance of I, the pixel values to be given by $\sigma_i^2 = \frac{255^2}{12}$ (or $\sigma_i = 73.6$). However, statistics obtained from many frames show that $\sigma_i = 55$. Therefore we assume that I has a uniform distribution with $\sigma_i = 55$. From Eq.(3.4) $\sigma_{\mathrm{ig}} \approx 55(\frac{12}{2\pi e})^{0.5} = 46.1$. If we allow a degradation of the image frame after the addition of the message to a PSNR of 42 dB, then $\sigma_s^2 = 4$. Furthermore, if the image frame goes through MPEG compression (say 50 fold compression), then it is measured for test sequences that $\sigma_p \approx 6.7$. This would yield a C_h value of 0.0013 bits/pixel. Even if the processing noise is increased to say $\sigma_p \approx 20$, (the resulting frames would be barely recognizable) C_h would still be 0.0011 bits/pixel.

4. DECOMPOSITION OF THE DATA HIDING CHANNEL

4.1. Need for a Decomposition

Figure 2 shows the typical distribution of image and processing noise in the channel, as a function of frequency. At low frequencies the image noise is high and the processing noise is low, while at high frequencies the image noise is low and processing noise is high. At midband frequencies however, we strike a compromise. Obviously, we could make better utilization of the data hiding channel if we decompose the channel into multiple sub-channels. It is also clear that if the processing noise is negligible, a decomposition with good energy compaction property or high Transform Coding Gain (GTC) [6], would concentrate the image noise in a small number of sub-channels, leaving a large number of sub-channels with very little image noise. However if the processing noise is high (low quality MPEG), the high frequency sub-channels would be affected drastically, leaving very few useful sub-channels. On the other hand, a transform with inferior energy compaction property may still have midband frequencies relatively unaffected by compression. Figure 3 shows the position of different transforms in the "scale" of energy compaction. At the left end we have the Identity transform which has no energy compaction (GTC= 1). At the right extreme we have the the best energy compacting transform, or KLT [6]. In this paper we obtain the achievable capacities for 2 block transforms at various positions in the scale- DCT and Hartley transform, and Identity transform.

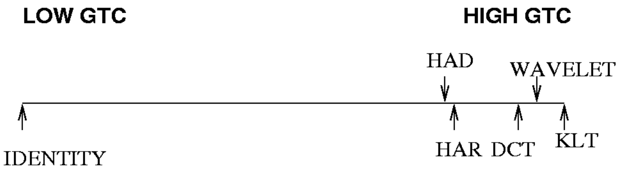

Figure 3: The Energy Compaction Scale

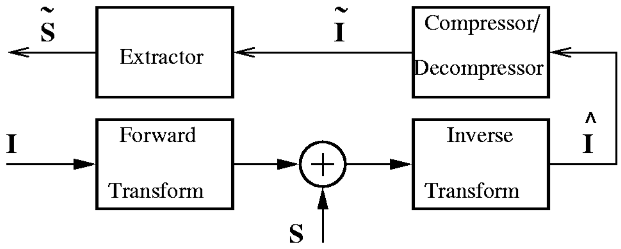

Figure 4: Schematic of Data Hiding / Retrieval

4.2. Capacity of Multiple Channels

Figure 4 displays the block diagram of a typical data hiding scheme. The Forward and Inverse Transform blocks decompose the single channel of Figure 1 into multiple sub-channels of Figure 5. The decomposition of a frame into its L sub-bands results in L parallel sub-channels with two noise sources in each sub-channel. Let σ_{ij}^2, $j = 1 \cdots L$, be the variances of the coefficients for each sub-band (or the variances of the image noise in each sub-channel) of

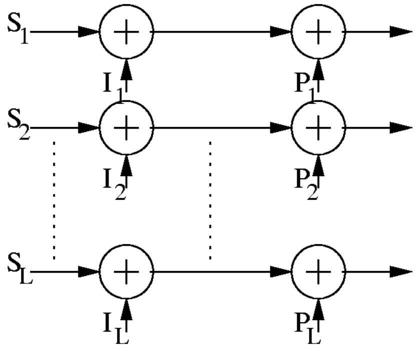

Figure 5: Decomposition of the Data Hiding Channel

the decomposition. Similarly, let their corresponding equivalent Gaussian variances be $\sigma^2_{\mathrm{ig}_j}$. If $\sigma^2_{p_j}$ is the variance of the processing noise (Gaussian) in the j^{th} sub-channel, then, the combined total capacity of the L parallel sub-channels is given by

$$\mathrm{C}_h = \frac{MN}{2L} \sum_{j=1}^{L} \log_2(1 + \frac{v_j^2}{\sigma^2_{\mathrm{ig}_j} + \sigma^2_{p_j}}) \text{ bits} \tag{4.5}$$

for a frame of size MN pixels. In the above equation, v_j is the *visual threshold* of band j. Inm other words, v_j^2 is the maximum message signal energy permitted in band j. The term *visual threshold* however, is highly subjective. To derive a reasonable model we argue that JPEG, at a reasonably good quality factor (like 75) is well tuned visually in distributing the quantization errors amongst the bands, at least with respect to preserving the visual fidelity of the compressed image. Let i_{j_k} be the coefficients of some original images, and $\tilde{i}_{j_k}$ the coefficients of the same images that have gone through JPEG compression and decompression. Let $\sigma^2_{q_j}$ be the variance of the quantization error, $\mathrm{e}_{q_j} = \tilde{\mathrm{i}}_j - \mathrm{i}_j$, for sub-band j. If quantization error (due to JPEG) of variance $\sigma^2_{q_j}$ in sub-band j, results in an image that is visually satisfactory, we can argue that addition of message signal with energy $\sigma^2_{q_j}$ in sub-band j, would still render the image $\hat{\mathrm{I}}$ with an acceptable visual quality. However, in order to maintain the PSNR of $\hat{\mathrm{I}}_k$ in the range of 40-50 dB (so that the $\hat{\mathrm{I}_k}$ is visually indistinguishable from I_k), we choose the sub-band visual thresholds as

$$v_j^2 = K_2 \sigma^2_{q_j} \tag{4.6}$$

where $K_2 < 1$.

4.3. Modeling Channel Noise

In order to model the channel noise (The noise sources $\mathrm{I}_j, j = 1 \cdots L$ and $\mathrm{P}_j, j = 1 \cdots L$ in Figure 5), we obtain their statistics from 90 frames of monochrome video sequences, and their MPEG-2 compressed versions.

The original frames (or the image noise 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 $\sigma^2_{i_j}$. Having obtained the variances of the image noise in each sub-channel, the next step is to obtain their equivalent Gaussian variance. 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 $\sigma^2_{\mathrm{ig}_j}$ are obtained as

$$\sigma^2_{\mathrm{ig}_j} = \frac{2^{2\mathcal{H}_j}}{2\pi \mathrm{e}}, \quad \mathcal{H}_j = -\sum_{i=1}^{n} \frac{g(i)}{p\Delta x} \log_2(\frac{g(i)}{p\Delta x})\Delta x. \tag{4.7}$$

The image noise in sub-channel j can then be substituted by Gaussian noise of variance $\sigma^2_{\mathrm{ig}_j}$.

Let the noise due to compression in each sub-channel be $\sigma^2_{p_j}$, $j = 1 \cdots L$. As in Section 2, we are justified in assuming a Gaussian distribution for the processing noise for each sub-channel. The variance of the equivalent additive Gaussian noise is estimated as follows. We obtain $\frac{MNn_f}{L}$ samples of each band from N_f frames. Let i_{j_k}, $k = 1, \ldots, \frac{MNn_f}{L}$, be the coefficients of the band j of the decomposition of the original frames. Let $\tilde{i}_{j_k}$, $k = 1, \ldots, \frac{MNn_f}{L}$ be the corresponding coefficients of the reconstructed frames. We obtain the equivalent additive noise in each sub-channel as noise uncorrelated with i_j, that would cause the same reduction of correlation between i_j and $\tilde{\mathrm{i}}_j$. We define the intra-band correlation as

$$\frac{\langle \mathrm{i}_j, \tilde{\mathrm{i}}_j \rangle}{|\mathrm{i}_j||\tilde{\mathrm{i}}_j|} = \frac{\langle \mathrm{i}_j, (\mathrm{i}_j + \mathrm{n}_j) \rangle}{|\mathrm{i}_j||\mathrm{i}_j + \mathrm{n}_j|} = \rho_j. \tag{4.8}$$

where n_j is a vector of Gaussian (zero mean) random variables which is uncorrelated with i_j. Then $\sigma^2_{n_j} = |\mathrm{n}_j|^2$ is the variance of the *equivalent additive noise due to compression.* Or $\sigma_{p_j} = \sigma_{n_j}$. As $\langle \mathrm{i}_j, \mathrm{n}_j \rangle = 0$, Eq.(4.8) can be simplified to obtain

$$\sigma^2_{p_j} = |\mathrm{n}_j|^2 = (\frac{1}{\rho_j^2} - 1)|\mathrm{i}_j|^2 \tag{4.9}$$

Note that for the purpose of calculating the processing noise, we make the following two assumptions:

- The message signal added to some sub-channel is statistically similar the the sub-channel coefficients themselves. So, the message signal is treated in the same way as the image coefficients by the compressor.
- The entropy of the image sub-channel coefficients is significantly larger than the entropy of the added message signal. This is to make sure that the compressors performance is not affected much by the addition of the signature.

5. RESULTS AND CONCLUSIONS

We calculate the data hiding capacity of 90 frames for 2 monochrome image sequences, viz. Table Tennis and Football. The capacities, calculated for 3 transforms - DCT, Hartley (HAR) and Identity (ID) for 5 different processing noise scenarios are displayed in Figure 6 for I Frames (top row) and P/B Frames (bottom row). The processing noise scenarios 1-5 are respectively MPEG-2 compression (30 frames/ sec, 15 frames in GOP and I/P frame distance of 3). at compression ratios 1 (lossless compression), 10, 25, 50 and 100 respectively. For our simulations we have chosen $K_2 = 0.3$ in Eq. 4.6. From the plots in Figure 6, we can see that the bitrates for all decompositions fall with increased processing noise, as expected (though the fall is barely noticeable for Identity transform). Figure 7 shows the image and processing noise in the 64 sub-channels of the DCT and the Hartley decomposition (for Processing Scenario 4 or compression ratio of 50). While the high frequency DCT sub-channels suffer very high processing noise, the high frequency Hartley transform sub-channels are not affected to the same extent. As expected, DCT performs better than Hartley transform for low processing noise scenarios (Scenarios 1 and 2). However, the performance of DCT falls drastically if the processing noise is increased. Thus Hartley transform is a better decomposition to use for higher processing noise scenarios (3,4 and 5). For even higher processing noise scenarios, it might turn out that Identity transform is more suitable. However, for the processing noise scenarios the message signal is *usually expected to survive*, Hartley transform (or some other transform with similar data compaction property), would probably be the best decomposition to use. Finally, there was not much difference between the performance of I, P, and B frames, though the "processing" they undergo is different. P and B Frames yielded slightly (10 %) higher capacities than I Frames (which is not surprising as typically the average PSNR of P/B Frames where about 2 dB better than I Frames). The

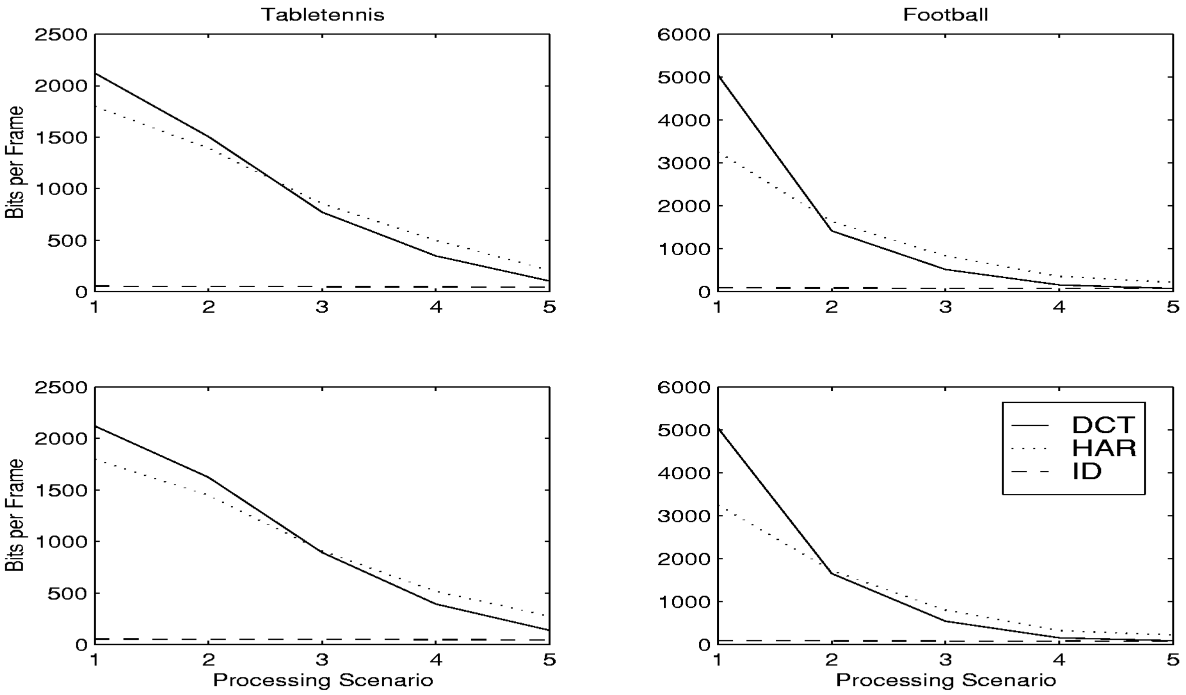

Figure 6: Channel Capacities of Different Decompositions. Top Row: I Frames; Bottom Row: P/B Frames. The Processing Scenarios 1-5 correspond to lossless compression, and compression ratios of 10, 25, 50 and 100 respectively.

difference between P and B Frames were however negligible. So they have been grouped together in Figure 6.

Finally, it should be noted that there may be other factors apart from the Transform Coding Gain which could affect the data hiding capacity. For instance a transform with lower GTC than Hartley transform may yield better capacity at lower processing noise scenarios than Hartley transform. However, what is clear is that, in general, high GTC transforms are not preferable choices for data hiding.

6. REFERENCES

[1] M.Ramkumar, A.N. Akansu, "Information Theoretic Bounds for Data Hiding in Compressed Images", to be presented in the 1998 Workshop on Multimedia Signal Processing (MMSP-98), Los Angeles, CA, USA, December 7-9 1998.

[2] I.J. Cox, J. Kilian, F.T. Leighton, and T.G. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia", IEEE Transactions on Image Processing, **6** (12) pp 1673-1687, 1997.

[3] W. Zeng, B. Liu, "On Resolving Rightful Ownerships of Digital Images by Invisible Watermarks", Proceedings of ICASSP, ICIP-97, vol **1**, pp 552-555.

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

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

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

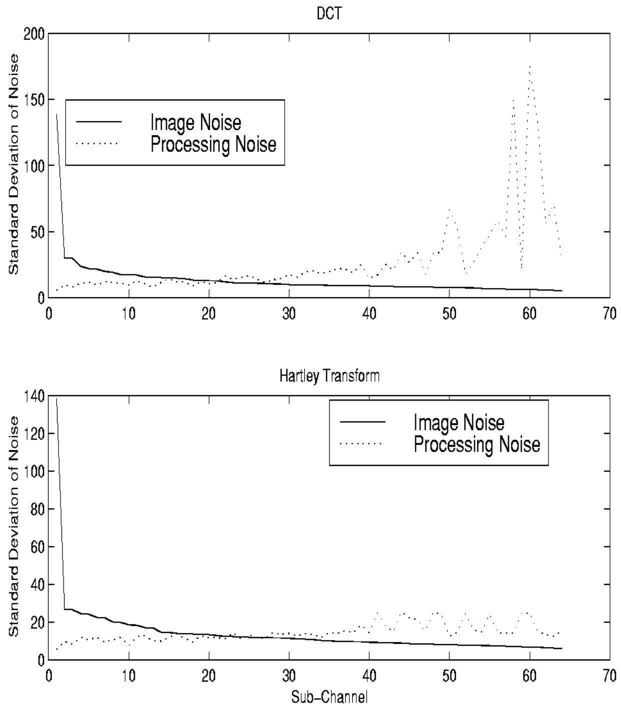

Figure 7: Image and Processing Noise for Various Sub-Channels of DCT and Hartley Decomposition