A NEW INTERPRETATION OF DATA HIDING CAPACITY

Çağatay Candan and Nikil Jayant

Multimedia Communications Laboratory
Georgia Institute of Technology, Atlanta, GA 30332
Contact:{candan, jayant}@ece.gatech.edu

ABSTRACT

We present a new definition of data hiding capacity which complements the established theory in the field and produces practical estimates under many attacks. We discuss the relation between the proposed definition and the current theoretical work on data hiding capacity. The definition proposed is applied to still images to estimate the hiding capacity of a particular image under attacks such as JPEG compression and additive noise.

1. INTRODUCTION

Data hiding is the art of embedding application-oriented information, such as captions and copyright notice, in a host signal without causing a perceptible distortion to the host. Even though the definition of data hiding emphasizes imperceptibility, the main challenge is to combine this property with robustness to adversarial attacks on the hidden signal.

Data hiding can be modeled as the communication between two parties under the cover of a host signal. The transmitter embeds the application-oriented information in the host signal; this composite signal propagates through a channel which models the attacks of the hostile parties. At the decoder, the received signal is processed to extract the host data and the hidden information. With this modeling, we can use some information theoretical results to bring a theoretical foundation to the data hiding problem. But an important difference between the two theories is that the attack channel in data hiding is in general carefully designed to remove the hidden data, but in communication theory the attack or noise is almost always due to uncontrolled or accidental behaviour of the communication system.

If we list the main challenges of data hiding, we can include the following: the lack of complete understanding of the perception mechanisms, the difficulty of attack modeling, the difficulty of a metric definition to measure the success of a particular method and the lack of a founding theory for data hiding like the Shannon's theory of communication [1].

In this paper, we present a new definition for the capacity of data hiding systems which complements the theory set forth in [2, 3, 4]. Furthermore this definition gives practical estimates on capacity in many cases. The main result of the paper is stated as follows:

Conjecture: The amount of information that can be imperceptibly hidden in a media carrier is the difference between the bit rate used in the compression of it and the perceptual entropy of the signal.

A discussion of the perceptual methods in multimedia signal processing can be found in [5, 6]; applications of these concepts in the data hiding context have been given in [7, 8].

The theoretical approach to data hiding has been initiated by the recognition of the analogy between data hiding and the communication channel whose state information is only known by the transmitter [3]. This analogy has been extended by modeling the data hiding process as a game between the information hider and the attacker [4]. Some of the earlier work on the definition of the capacity has also recognized the game-theoretic approach and resulted in simple, but nevertheless elegant results [2].

Very recently, the duality between source coding with side information at the receiver and data hiding (source coding with side information at the transmitter) has been explored [9, 10]. A mathematically proven method based on these ideas has been proposed in [3].

In this paper, we give a discussion of the conjecture and establish the relation between the conjecture and the theoretical results given by Moulin *et. al.* and Chen [4, 3]. After the exploration of this relation, we give some upper-bounds on the data hiding capacity for the digital images under different attacks. The paper is concluded with a discussion of the capacity-achieving conditions and the results of the computer experiments on images.

2. DATA HIDING

Before the discussion of the capacity problem, we will give a model for a general data hiding system. In Figure 1, the host signal and the hidden information are represented with $\tilde{x}$ and w respectively. These two signals may or may not be independent depending on the application. The first block of the transmitter is the perceptual source coder [5]. The output of this block represents the perceptually relevant components of the signals as represented by x_p and w_p. The next block in the transmitter is the data hiding block which combines x_p and w_p in a fashion that signal x is perceptually indifferent from $\tilde{x}$ and the hidden signal w is robustly protected from the attacks. Therefore, this block serves two purposes, imperceptible data hiding and attack compensation (channel coding). Attack on the composite signal which can be deterministic (compression) or random (additive noise) is represented by the channel. Finally, the channel output y is processed at the

0-7803-7041-4/01/$10.00 ©2001 IEEE

decoder to estimate x and w. Depending on the intended application, the host signal may or may not be available at the decoder. The latter case, which is known as blind data hiding, may bring more difficulty to the decoder due to the influence of the host signal acting as an additional noise source.

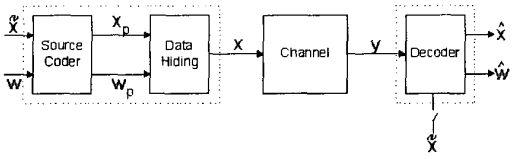

Fig. 1. A system model for data hiding

Data hiding problem with this model can be stated as the maximization of the rate of the signal w (R_w), under the maximum distortion constraint on x, while keeping the probability of extraction error of the hidden information ($P(\hat{w} \neq w)$) at an arbitrarily small value.

We present an example to clarify the details of the model. Let's assume that $\tilde{x}$ and w are text messages, that is $\tilde{x}[n]$ is the nth letter of a novel and w is the secret text message to be inserted in the novel. The distortion constraint on x allows us to change at most 1 letter out of 100 letters of the original text. A more frequent insertion (deliberate typo) will render the text to be useless. Under these circumstances, the first block of the encoder compresses messages to their essentials without any loss: that is, the redundancy of the language in its structure such as grammar, punctuation is removed. For example in English the letter q is always followed with the letter u. Therefore it is possible to remove all of the u letters coming after q's. Similarly all of the vowels in the novel can be replaced with dashes and an experienced reader should be able guess all the vowels. An important point is that after the perceptual compression, the information rate of text messages is reduced from $log_2(27)$ bits/letter (26 alphabet letters and space character) to R_x and R_w which is strictly less than $\log_2(27)$ bits/symbol. Data hiding block then constructs the composite message x from the perceptual messages x_p and w_p in a way that there is no ambiguity of message extraction at the decoder. The composite message passes through a proof-reader (attack channel) and reaches the hands of the intended party.

A important detail is that the rate of the signal x has to be at least $R_x + R_w$,b ecause both signals x_p and w_p have to be combined together in an invertible fashion (the hidden information should be separable from the composite signal at the decoder). Therefore, if the alphabet of x_p has 2^{R_x} symbols and w_p has an alphabet size of 2^{R_w}, the composite alphabet has to have at least $2^{R_x+R_w}$ symbols, so that the composite signal can be partitioned into two components in a unique way.

Another point regarding the system is that the capacity of the attack channel should be at least $R = R_x+R_w$. Otherwise it is not possible to have a reliable communication between the input and output of the channel.

We present an interpretation of the capacity conjecture based on this model in the next section.

2.1. The Conjecture

The equation for the capacity of an arbitrary channel is given as $C = \max_{p(w)}\{I(w;y)\}$ by Claude Shannon in 1948 [1]. In this equation w and y denote the channel input and output respectively. The channel is defined through an input-output map which can be probabilistic or deterministic (with probability distribution consisting of only 1's and 0's).

We first give the discussion for the *non-blind case*. The data hiding capacity C_h in this case can be written as follows:

$$\begin{aligned}
C_h &\overset{(a)}{=} \max_{p(w)}\{I(w;y|\tilde{x})\} \\
&\overset{(b)}{=} \max_{p(w)}\{H(w|\tilde{x}) - H(w|\tilde{x},y)\} \\
&\overset{(c)}{\leq} H(w_*|\tilde{x}) \\
&\overset{(d)}{=} H(w_*|\tilde{x}) + H(x_p) - H(x_p) \\
&\overset{(e)}{=} H(w_*|\tilde{x},x_p) + H(x_p) - H(x_p) \\
&\overset{(f)}{\leq} H(w_*|\,x_p) + H(x_p) - H(x_p) \\
&\overset{(g)}{=} H(w_*,x_p) - H(x_p) \\
&\overset{(h)}{\leq} (R_x + R_w) - H(x_p) \\
&\overset{(i)}{\leq} R - H(x_p) \\
&\overset{(j)}{\leq} C - H(x_p) \qquad (1)
\end{aligned}$$

Line (a) is the definition of the capacity for the non-blind case. Line (b) is the definition of the mutual information. The maximizing distribution is inserted in line (c) and the inequality is due to non-negativeness of entropy. In line (d), we introduce the variable x_p. Line (e) is valid since x_p is a function of $\tilde{x}$. In line (f) we use the rule that conditioning reduces entropy. Line (g) is the definition of the joint entropy. Line (h) follows from the Slepian-Wolf theorem (joint source coding [11, Theorem14.4.1]). Line (i) follows from the requirement of unique separation of the host data and the hidden data. Line (j) is due to the assumption of reliable communication.

We see from the chain of inequalities that data hiding capacity for the non-blind case is upper bounded by the difference of the capacity of the attack channel and perceptual entropy of the host signal, as conjectured.

The *blind case* is more difficult to analyze, but recent studies have established important steps in this direction. In [4], data hiding operation has been defined as a game between the hider and the attacker. If the optimum strategy for both players is exercised, the capacity of the data hiding game is given by $C = \max_{p(x,u|\tilde{x})} \min_{p(y|x)}(I(U;Y) - I(U;\tilde{X}))$. The composite signal x is constrained to be below a distortion limit. The attacker also has a maximum distortion limit which prohibits the use of excessive distortion on x. In this theory the variable u is represented as the auxiliary variable, or as a dummy variable, over which maximization is accomplished. We believe that the signal u has an important role

in the data hiding context. We propose to interpret the signal u as the signal x_p which represents the perceptually coded version of the signal $\tilde{x}$ according to our model.

Assuming that we have fixed the attack channel ($p(y|x)$), the capacity in this case can be written as follows:

$$\begin{aligned} C_h &\overset{(a)}{=} \max_{p(x,u|\tilde{x})} (I(U;Y) - I(U;\tilde{X})) \\ &\overset{(b)}{\leq} \max_{p(x,u|\tilde{x})} I(U;Y) - \min_{p(u|\tilde{x})} I(U;\tilde{X}) \\ &\overset{(c)}{=} \max_{p(x,x_p|\tilde{x})} I(X_p;Y) - \min_{p(x_p|\tilde{x})} I(X_p;\tilde{X}) \\ &\overset{(d)}{=} C - H(x_p) \end{aligned} \quad (2)$$

Line (a) is the definition of the capacity for a fixed attack channel. In line (b), we upper bound the equality in (a) by maximizing the two terms of (a) independently. In line (c), we make the analogy of identifying u with x_p. The second term of the line (d) can be recognized as the minimum rate that is necessary to construct signal x_p from $\tilde{x}$, which is the entropy of the signal x_p (perceptual source coding). The first term of line (d) is the definition of the capacity of the attack channel (channel coding) for the signal x_p.

The proposed analogy can be viewed as follows: the host signal is first coded to the signal u ($\tilde{x} \rightarrow u$) and then the signal u is coded once more to the signal x ($u \rightarrow x$). The final signal is transmitted through the attack channel. For data hiding applications, attacker has to watch the perceptual quality of the resultant signal after the attack. Because of this, attack tools can be pictured as tools operating on the perceptual components of the host signal or they can be visualized as the operators working in the perceptual domain. With this visualization, the first coding operation, from $\tilde{x}$ to u, can be thought as the perceptual source coding (projection operation of the signal $\tilde{x}$ to the domain of the attack tools). The second coding operation, from u to x, is the channel coding for a particular attack tool (transformation of the signal x_p to the signal x whose components lie in the range space of that particular attack). With this analogy, the maximum value of the expression $I(U;Y) = I(X_p;Y)$ represents the maximum rate of reliable communication of the perceptual components of the host signal. We emphasize that if the attacker could apply arbitrary attacks, the analogy proposed would not be valid, since there would not be a common domain for the attacks.

2.2. Capacity Achieving Conditions

We list the capacity achieving conditions for the two cases of data hiding. We start with the non-blind case. The requirements can be listed as follows: 1. There is a small probability of error at the decoder (ignored term in line (c) is bounded by the Fano's inequality [11, Lemma 8.9.1]) 2. The hidden information should depend only on x_p i.e. signals $\{x, x_p, w_*\}$ should form a Markov chain of $x \rightarrow x_p \rightarrow w_*$ (from line f). 3. The perceptual source coder should be perfect (line h). 4. The data hiding operation should be invertible (line i). 5. Hidden data should be embedded at the maximum rate allowed by the attack channel which is $R_w = C - R_x$ (line j).

All the requirements, other than the second one, emphasize the ideal operating conditions for the data hiding system. The second requirement says that to maximize the capacity, the hidden data should be in relation with the perceptual components of the host signal, but not with the host signal itself.

For the blind case, line (b) implies that the data hiding throughput C_h is maximized when the $p(u|x)$ appearing in both terms of line (b) are the same (The probability distribution $p(x_*, u_*|\tilde{x}) = p(x_*|u_*\,\tilde{x})p(u_*|\tilde{x})$ maximizes the first term and at the same time the distribution $p(u_*|\tilde{x})$ minimizes the second term). If the analogy between u and x_p is applicable, we expect this relation to be satisfied (perceptual coding is independent of the channel).

3. CAPACITY ESTIMATES FOR SOME PRACTICAL CASES

In this section we present the capacity estimates of data hiding systems under some practical attacks. We experiment with the 512x512 Lena image whose pixels are represented with 256 gray levels. To determine the perceptual entropy of the Lena image, we used Watson's human visual system model and assumed that pixels below the just noticeable distortion (JND) threshold do not contribute anything perceptually [6].

Noiseless Channel: If the attack channel is noiseless (no-attack condition), the capacity is given by $C(D) = 8 - R(D)$ bits / pixel, where $R(D)$ is the perceptual rate distortion function [6]. The top panel of the Figure 2 shows the percentage of the pixels available for data hiding ($C(D)/(512 \times 512)$) as the allowable distortion due to the embedding increases. The Lena images at different allowable distortion levels (zero distortion, JND, 3 $\times$ JND, 20 $\times$ JND) are shown in Figure 3.

JPEG Compression Channel: JPEG compression operation is inserted in the attack channel. The bottom panel of Figure 2 shows a relationship similar to that in the top panel. It is clear that JPEG compression takes most of the redundancy, but the left over redundancy is enough to insert hidden data at approximately 1310 pixels of the Lena image ,which corresponds to 0.5% of total number of pixels, without any perceptual distortion. It is clear that under perfect compression, there would be no room left for data hiding.

In Figure 3, we show the original Lena image and some distorted versions of it to see the extent of hiding distortion to achieve a particular capacity value. The distorted images are the quantized versions of the Lena image at the multiples of the JND threshold. As expected hiding capacity increases as the allowed distortion on the host image increases.

Additive Noise Channel: Assume an attack of a binary symmetric memoryless channel with the transition probability of ϵ functioning independently on each transmitted bit of Lena image. The capacity of this case is given as $C(D) = 8(1 - H(\epsilon)) - R(D)$ bits/pixel when this value is greater than zero; otherwise zero. According to the vision model adopted, the Lena image can be compressed at 0.52 bits/pixel without a perceptible distortion. Therefore for the $\epsilon = 0.25$ it is possible to encode almost 1 bit of hidden data per pixel without a perceptual quality loss.

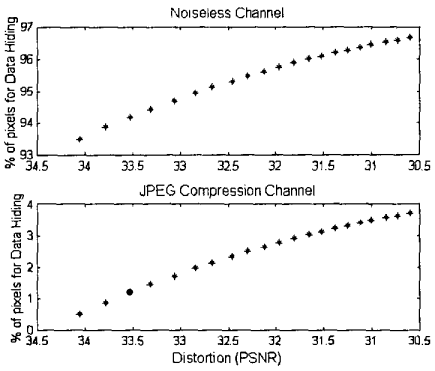

Fig. 2. The top panel shows the data hiding capacity of the Lena image as the distortion due to the embedding increases gradually. The lowest distortion value of the figure corresponds to the JND distortion which is 34.05 dB in terms of the PSNR metric. The hiding capacity is given by the percentage of the total number of transform coefficients which can be modified for the data hiding purposes. The bottom panel gives the capacity estimate of the same image under the JPEG compression attack.

Image rotation, flipping and other invertible operations: According to the Shannon's information theory, an invertible attack on a signal does not reduce the entropy of the signal. Therefore any invertible attack such as image flipping or rotation does not pose a threat to the capacity. But in practice, undoing the effects of these operations (especially sequential combination of these operations) can be computationally very intensive and decoding of the hidden data may be close to impossible.

4. CONCLUSION

In this paper, we have presented a simple definition for the capacity of data hiding systems based on perceptual models. This definition is not only in accord with the theory established so far, but also gives us practical capacity bounds which can be useful for the benchmarking of data hiding systems.

The information theoretical results presented in this paper strengthen our belief that efficient and robust data hiding methods can result from the joint study of hiding and source coding methods. We believe that successful combination of source coding, channel coding and techniques from cryptography will provide maximal resistance to attacks on data hiding systems.

5. REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *BELLSYS*, vol. 27, pp. 379–423, 1948.

[2] S. Servetto, C. Podilchuk, and K. Ramchandran, "Capacity issues in digital image watermarking," *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, vol. 1, pp. 445–449, 1998.

[3] B. Chen, *Design and Analysis of Digital Watermarking, Information Embedding and Data Hiding Systems*. PhD thesis, Department of EE, MIT,20 00.

Fig. 3. This figure shows the relation between the distortion due to data hiding and the hiding capacity under the JPEG compression attack with the default quantization matrix. The top left image is the original. The top right image is distorted up to the JND threshold (the hiding capacity is 1310 pixels). The bottom left image is perceptibly distorted (up to 3 times the JND threshold and its capacity is 9670 pixels). The bottom right image is severely distorted (up to 20 times JND threshold and its capacity is 15000 pixels).

[4] P. Moulin and J.A.O'Sullivan, "Information theoretic analysis of information hiding," *IEEE Trans. Information Theory*, submitted 2000.

[5] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1422, 1993.

[6] A. Watson, "Perceptual optimization of dct color quantization matrices," *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, vol. 1, pp. 100–104, 1994.

[7] C. Podilchuk and W. Zeng, "Image-adaptive watermarking using visual models," *IEEE Journal on Selected Areas in Comm.*, vol. 16, pp. 525–539, 1998.

[8] R. Wolfgang, C. Podilchuk, and E. Delp, "Perceptual watermarks for digital images and video," *Proc. IEEE*, vol. 87, pp. 1108–1126, 1999.

[9] J. Chou, S. Pradhan, and K. Ramchandran, "On the duality between distributed source coding and data hiding," *Asilomar Conf. on Signals, Systems & Computers*, pp. 1503–1507, 1999.

[10] R. Barron, B. Chen, and G. Wornell, "The duality between information embedding and source coding with side information and its implications/applications," *IEEE Trans. Information Theory*,sub mitted 2000.

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