A STEGANOGRAPHY SCHEMA AND ERROR-CORRECTING CODES

M. B. OULD MEDENI, EL MAMOUN SOUIDI

Laboratoire de Mathématique informatique et Application, Faculty of Sciences, B.P 1014 , Rabat MOROCCO

E-mail: sbaimedeni @yahoo.co.in, souidi@fsr.ac.ma

ABSTRACT

The idea of "Matrix encoding" was introduced in steganography by Crandall in 1998 [6]. The implementation was then proposed by Westfeld with steganography algorithm F5 [1]. The objective is to transmit a message within an image, but with the constraint of minimizing the number of changed coefficients of this image. In this paper, a new construction of steganography protocol is considered, which is an extension of the error correcting code and steganography construction. The proposed method consists of use the Majority logic decoding introduced in [5], for embedding the message in the cover image, the extraction function is always based on syndrome coding. An asymptoticly tight bound on the performance of embedding schemes is given.

Keywords: *Steganography, Error correcting code, Majority logic decoding average distortion, matrix encoding, embeding efficient.*

1. INTRODUCTION

The goal of digital steganography is to modify a digital object (cover) to encode and conceal a sequence of bits (message) to facilitate covert communication. The goal of steganalysis is to detect (and possibly prevent) such communication. Often, the cover media correspond to graphics files. Graphics files are the typical choice because of their ubiquitous presence in digital society, but any medium that contains a substantial amount of perceptually insignificant data can be used.

An interesting steganographic method is known as matrix encoding, introduced by Crandall [6]. Matrix encoding requires the sender and the recipient to agree in advance on a parity check matrix H, and the secret message is then extracted by the recipient as the syndrome (with respect to H) of the received cover object. This method was made popular by Westfeld [1], who incorporated a specific implementation using Hamming codes in his F5 algorithm, which can embed t bits of message in $2^t - 1$ cover symbols by changing, at most, one of them.

There are two parameters which help to evaluate the performance of a steganographic method over a cover message of N symbols : the average distortion D= $\frac{R_a}{N}$, where R_a is the expected number of changes over uniformly distributed messages ; and the embedding rate E = $\frac{t}{N}$, which is the amount of bits that can be hidden in a cover message [4]. In general, for the same embedding rate a method is better when the average distortion is smaller.

Furthermore, we will also assume that a discrete source produces a sequence x = $(x_1, \dots, x_N)$, where, N is the block length and each $x_i \in F_2$ (F_2 is the Galois field {0,1}). The message $s = (s_1, s_2, \dots, s_M)$, where M is the message length, we want to hide into a host sequence x produces a composite sequence y = f(x, s), where $y = (y_1, y_2, \dots, y_n)$, and each $y_i \in F_2$. The composite sequence y is obtained from distorting x, and the distortion will be assumed to be a squared-error distortion. In these conditions, if information is only carried by the least significant bit (LSB) of each x_i, the appropriate solution comes from using binary Hamming codes [1].

In this work we propose a steganography method based on a [n, k, t] codes BCH. While the size of each cover block is $n = 2^m - 1$, m = 3, 4, 5…, and the rate of capacity is m×t. Our method uses t = 2. Because with t = 2 BCH is semi perfect.

in this paper we find embedding efficiency=(rate of capacity/number of changes), better than Hamming code used in [1].

The rest of this paper is organized as follows. Section 2 introduces the relationship between error-correcting codes and steganography systems. We present our approach in section 3. Section 4 is devoted to the bound on the performance of embedding schemes. Conclusions and future work are presented in section 5.

2. STEGANOGRAPHY AND ERROR CORRECTING CODE

R. Crandall introduced in [6] the matrix encoding idea to improve the embedding efficiency for steganography. The algorithme F5 proposed by Westfeld [1] is to reduce modification of the quantized DCT coefficients. Since F5, steganographers take the reduction of embedding capacity sincerely and coding theory into consideration. Basically the matrix encoding technique in F5 modifies at most 1 coefficient among n nonzero coefficients to hide k bits. For example, the matrix encoding method modifies at most one coefficients among seven coefficients to hide three bits like a [7, 3] Hamming code. Thus, distortion of image is reduced at the cost of sacrificing the embedding capacity. Now, not all coefficients have to be modified by using [n, k, 1] code where $n = 2^k - 1$. Modified matrix encoding (MME) [7] uses [n, k, 2] code where one more coefficients may be changed in each group compared with the matrix encoding. The concept of the matrix encoding technique is « the less number of modification to the DCT coefficients, the less amount of distortion in the image »[8]. Matrix encoding using linear codes (syndrome coding) is a general approach to improving embedding efficiency of steganographic schemes. The covering radius of the code corresponds to the maximal number of embedding changes needed to embed any message. Steganographers, however, are more interested in the average number of embedding changes rather than the worst case. In fact, the concept of embedding efficiency- the average number of bits embedded per embedding change-has been frequently used in steganography to compare and evaluate performance of steganographic schemes.

Example (LSB EMBEDDING)

- Cover object = { 3, 6, 5, 0} = {011; 110; 101; 000}
- Secret message bits : (0, 0, 1, 1)
- Stego-support={010, 110, 101, 001}
- Embedding efficiency = $\frac{1}{\frac{1}{2}}$=2

2.1 Error Correcting Code in Steganography

An important kind of steganographic protocols can be defined from coding theory. Error-correcting codes are commonly used for detecting and correcting errors, or erasures, in data transmission. An explicit description of the relationship between error-correcting codes and steganographic systems was presented by Menuera and Zhang, Li in [2,9] and shows that there is a corresponding relation between the maximum length embeddable (MLE) codes and perfect error correcting codes. The most used codes in steganography are linear. The existence of a parity check matrix helps on designing good steganographic protocols.

Let C be a linear [n, n-t] code over the finite field F_q, equivalently, a linear subspace of $F_q{}^n$, of dimension is k =n-t. The covering radius δ of the code C is defined as δ =$\max_{v \text{ in } F_q{}^n}$d(v,C), where d(v,C) means the minimum Hamming distance from the vector v to the code C. The support of a vector $v = (v_1, v_2, \dots, v_n)$ in $F_q{}^n$ is defined to be supp(v) = {i \ v_i≠0}.

Let $F_q{}^n$ and H be a parity check matrix of C. The syndrome of any v in $F_q{}^n$ is the vector r(v) = $H \times v^T$, where v^T means the vector v as a column vector. A coset C + v is the set of all vectors in $F_q{}^n$ with the same syndrome. A vector $I_r(v)$ of the minimum weight in C + v will be called leader of the coset, it is not necessarily unique. The above syndrome map r : $F_q{}^n \rightarrow F_q{}^t$ such that r(v) = $H \times v^T$, is called the retrieval map of a [n, t, δ] steganographic protocol, which will be called linear to emphasize that the retrieval map r is a linear map. The embedding algorithm to compute e(s, v) for a linear steganographic protocol works in the following way [10] :

COSET ALGORITHM

- Compute u := r(v) - s,

- define e(s; v) := v - $I_r(u)$, where $I_r(u)$ is a leader of the coset C + u of all the vectors in $F_q{}^n$ with the same syndrome u. So, r($I_r(u)$) = u.

SINGLE-ERROR CORRECTING CODES

We now give an example of a protocol steganography constructed from a linear single-error-correcting code. This was also discussed, for example, in [1]. Start from the matrix

$$H=\begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

whose entries are elements of F_2. Extracting Scheme is defined

$F : F_2^{\,7} \rightarrow F_2^{\,3}$ $\quad F(x_1, x_2, \ldots, x_7) = (y_1, y_2, y_3)$

$y_1 = x_1 + x_4 + x_5 + x_7$,

$y_2 = x_2 + x_4 + x_5 + x_7$

$y_3 = x_3 + x_5 + x_6 + x_7$

This function can be described in terms of matrix H. In fact, y_i, is the dot product of x and the i-th row of H. We claim that F is an extracting function of the (7, 3 ,1) protocol steganography.

Embedding Scheme for example, F(0,0,1,1,0,1,0) =(1,0,0). Assume y = (1,1,1). We claim that it is possible to replace x = (0,0,1,1,0,1,0) by x_0 such that F(x_0) =(1,1,1) and d(x, x_0) = 1. In fact, we claim more : the coordinate where x has to be changed is uniquely determined. In our case, this is coordinate number 6, so x_0 = (0,0,1,1,0,0,0),

Here is the general embedding rule : form F(x) + y, (in the example this is 011). Find the column of H which has these entries (in our example, this is the sixth column). This marks the coordinate where x needs to be changed to embed payload y. This procedure indicates how H and F were constructed and how this can be generalized : the columns of H are simply all nonzero 3-tuples in some order. In general, we start from our choice of n and write a matrix H whose columns consist of all nonzero n-tuples. Then H has N = $2^m - 1$ columns. The extracting function,

$$F : F_2^{\,N} \rightarrow F_2^{\,m}$$

is defined by the way of the dot products with the rows of H. Finally it is clear that embedding efficiency= 3.

2.2 MAJORITY LOGIC DECODER

Majority logic decoding algorithm was introduced in [5] and is briefly explained below. Let H be a parity check matrix of a [n, k] linear code C. Majority logic decoding implements a voting scheme among a set of check sums orthogonal on a bit or subset of error bits. The majority logic decoding rule is defined for a set of, J check sums orthogonal on error bit e_j as follows :

Let the estimate $\hat{e}_j$ of error bit e_j be the value assumed by the majority of the J check sums. In the case of a tie, let $\hat{e}_j$ = 0. The majority logic decoding rule guarantees a correct estimate of e_j as long as there are no more than $[\frac{J}{2}]$ errors among the error bits being checked. It is clear that for a block code with minimum distance d_{min}, majority logic decoding will be optimal when J = d_{min} -1. In such a case, the code is said to be completely orthogonalizable [5]. The J orthogonal check sums provide reliability information. In general, the greater the number of check sums which agree, the higher the reliability of the estimate. The lack of an extensive majority among the J orthogonal check sums can be used to generate a retransmission request.

3. NEW PROTOCOL STEGANOGRAPHY

The proposed approach works by dividing the cover into blocks of equal size. A block of binary data, e.g., LSB values of cover data, $\{v_0, v_1, \ldots, v_{n-1}\}$ over F_2 can be represented by a polynomial of X over $F_2^{\,m}$ such as v(X) = v_0 + v_1X + ... + $v_{n-1}X^{n-1}$. Embedding message m into the cover data v produces the stego data r which is represented as r(X) = $r_0 + r_1X + \ldots + r_{n-1}X^{n-1}$. The relation between m and r can be expressed as a matrix form as follows :

$$m = r \times H^T \qquad (1)$$

Decoder also uses Equation (1) to extract message from the stego data. By hiding message, some of the cover data bits are flipped from 0 to 1 and vice versa. Let e(X) be the flip pattern that represents which bit positions are flipped [8]. As a result, stego data is modified according to the flip pattern as follows :

$$r(X) = v(X) + e(X) \qquad (2)$$

From Equations (1) and (2), we get

$$m - v \times H^T \quad e \times H^T = \qquad (3)$$

The left-hand side of Equation (3) is called syndrome S. In other words, the syndrome is expressed as follows:

$$S = m - v \times H^T \qquad (4)$$

or, equivalently

$$S = e \times H^T \qquad (5)$$

From the steganographic point of view, our objective is to find a minimal number of flips of e(X) satisfying Equation (5) in order to decrease distortion. This is the syndrome coding. Data hiding by error-correcting code solves Equation (5) based on the vector e. The solution shows the proper positions of the elements in vector v(X) to be modified in order to hide message m to vector v(X). The stego vector r(X) is calculated according to the Equation (2). The hidden message can be recovered from stego vector r(X) using Equation (1).

3.1 PROPOSED EMBEDDING SCHEME

Inputs : message m, block of image C, distortion maximum allowed t.

Output : stego-image

- Step(1) : compute S by the equation 4.

- Step(2) : if S = 0 ; supp(e) = $\emptyset$, e(X) = 0, then the message is already hiding else go to Step(3).

- Step(3) : S = (S_0 , S_1 ,… , S_{n-k-1}) ; It seeks to construct a set P systems of equations parity , Such that each system L $\in$ P, there is an i $\in$ supp(e) ; such that L is orthogonal to e_i

-Step(4) : solve all the systems belonging to P and find e_i = 0 or 1 ; for all i $\in$ supp(e(X))

-Step(5) : if w(e(X)) $\leq$ t ; go to step(6), else go to step(3).

-Step(6) : compute the stego-support r(X) by Equation 2, you spin the algorithm for every image blocks

EXTRACTING SCHEME

The message is retrieved from the stego-support, with the function of syndrome :

$$m = r \times H^T$$

3.2 EXPERIMENTAL RESULTS

A generalization of perfect codes is the following : a t error- correcting code is said to be quasi-perfect if its covring radius is δ = t+1 (or equivalently, if the spheres of radius t+1 around the codewords contains all vectors of $F_q{}^N$). For example, all double-error-correcting BCH codes are quasi-perfect (see [11], Chapter 9, Section 8). We can use these codes to construct steganographic protocols. For every integer m > 2, the binary two error correcting BCH code C_m has parameters [$2^m - 1$, 2^m −2m- 1, 5] [11]. Its covering radius is δ = 3. Let S_m be the protocol obtained from C_m, by taking a parity check matrix as described above. It is a [$2^m - 1$, 2m, 3] protocol. The following tables collects the parameters of S_m and the corresponding version of F5 [1] (obtained from the Hamming code)

S_m					
m	n	k	δ	$\frac{k}{n}$	$\frac{k}{\delta}$
3	7	6	3	0,857	2
4	15	8	3	0,533	2,66
5	31	10	3	0,322	3,33
6	63	12	3	0,190	4
7	127	14	3	0,110	4,66

F5					
m	n	k	δ	$\frac{k}{n}$	$\frac{k}{\delta}$
3	7	3	1	0,428	3
4	15	4	1	0,266	4
5	31	5	1	0,161	5
6	63	6	1	0,095	6
7	127	7	1	0,055	7

$\frac{k}{n}$; $\frac{k}{\delta}$ measure, respectively, the embeding rate and embeding efficient.

4. ASYMPTOTICLY TIGHT BOUND ON THE PERFORMANCE OF EMBEDDING SCHEMES

Since for any given coverword v only $\sum_{i=0}^{t} C_n^i$ different (stego) words can be obtained by changing at most t coordinates of v, then we have the following proposition [3] :

5. PROPOSITION

For any embedding scheme of distorting t, using binary words of length n as coverwords, the number of different messages M(n, t) that can be embedded is bounded by

$$M(n,t) \leq \sum_{i=0}^{t} C_n^i \quad (6)$$

proposed h(n, t) = logM(n, t). The right hand side of (7) is upper bounded by

$$2^{nH_2\left(\frac{t}{n}\right)}$$

for 2t < n, where

$$H_2(x) = -x\log x - (1 - x)\log(1- x)$$

is the entropy function [3]. Therefor, we have

$$h(n, t) = nH_2\left(\frac{t}{n}\right)$$

Note this is a good approximation of (1) when t grows linearly with n. For other important case t fixed and n growing to infinity it follows

$$h(n,t) = t \log n - \log(t!); \text{ t fixed:} \quad (7)$$

Fortunately there are known constructions of steganography methode very close to the Hamming bound.

6. CONCLUSIONS

We have seen that there exists a close relation between steganographic protocols and error correcting codes (see subsection 2.2). Construction, parameters and properties of both are similar, and the ones can be deduced from the others. Since error-correcting codes are rather well known, this relation can be used to construct good steganographic protocols and study their properties. In this paper, we have presented a new method for steganography, based on error correcting code. This methode uses a class of decoding for error correcting code "majority logic decoding". This technique has representation that makes them efficient to work with. Future work will consider doing the following modifications to the proposed method :

- Investigating the proposed method on color images.
- modifying the proposed approach to embed image inside another image.
- Preserve the secret message even if we do some transformations on the image like rotation, scaling compression.
- Relate the encryption process with steganography in which we encrypt the message before embedding it inside the image in order to increase the security of the proposed method.

REFRENCES:

[1] A. Westfeld (eds.), F5 steganographic algorithm, In I.S Moskowitz éditeur : proc Information Hiding 4th International Workshop, IHW 2001.

[2] C. Munuera, Steganography and error correcting codes, Signal Process. 87 (2007) pp. 1528- 1533

[3] F. Galand and G. Kabatiansky, Information Hiding by Coverings, in proceedings of IEEE Information Theory Workshop 2003, IEEE, pp.151 154.

[4] Helena Rifa, Josep Rifa and Lorena Ronquillo, Perfect Z2Z4-linear codes in Steganography, available online at, http://arxiv.org/pdf/1002.0026

[5] J. L. Massey, Threshold Decoding. Cambridge, MA: M.I.T. Press, 1963

[6] R. Crandall Some notes on steganography, available at http://os.inf.tu-dresden.de/~westfeld/crandall.pdf, 1998.

[7] Y. Kim, Z. Duric, D. Richards: Modified matrix encoding technique for minimal distortion steganography, In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 314-327 (2007).

[8] Zhang. R, Sanchev, H. J. Kim Fast BCH Syndrome Coding for Steganography, In: Katzenbeisser, S. and Sadeghi, A.-R (Ed.) Information Hiding 2009, IH'2009, LNCS 5806, pp. 48-58, 2009, Springer-Verlag Berlin Heidelberg 2009.

[9] Zhang. W, S. Li, A coding problem in steganography, Des. Codes Cryptogr, 46(2008) pp 67-81

[10] H. Rifà-Pous, J. Rifà Product perfect codes and steganography, Digital Signal Processing 19 (2009) 764-769.

[11] F.J. Mac Williams, N. Sloane, The Theory of Error Correcting Codes, North-Holland, Amsterdam, 1977.