

A Bayesian Image Steganalysis Approach to Estimate the Embedded Secret Message

Aruna Ambalavanan
Stevens Institute of Technology
aambalav@stevens.edu

Rajarithnam Chandramouli
Stevens Institute of Technology
mouli@stevens.edu

ABSTRACT

Image steganalysis so far has dealt only with detection of a hidden message and estimation of some of its parameters (e.g., message length and secret key). To our knowledge, so far there is no steganalysis method that can estimate the hidden message itself. Our goal in this paper is to bridge this gap. We propose a steganalysis approach to estimate the hidden message based on a Bayesian framework, modeling the image as a Markov random field and exploiting the analogy between images and statistical mechanical systems. Message embedding in bit planes of an image is modelled as a binary symmetric channel. The theoretical framework is presented in detail. Experimental results are provided to support the theory.

1. INTRODUCTION

Image steganography deals with hiding a secret message(s) in a host image using a secret key such that its very presence is concealed (statistically). While the main goal of steganography is to hide the secret message the aim of steganalysis is to detect the same. According to [Trivedi and Chandramouli 2005], steganalysis could be passive or active. Passive steganalysis simply aims to identify the presence or absence of the secret message. Active steganalysis attempts to estimate the message length, secret key, message bits etc. Clearly, active steganalysis is a much more difficult task compared to its passive counterpart. Some recent work on active steganalysis include the following: [Trivedi and Chandramouli 2005, Chandramouli 2003, Holotyak et al. 2005, Fridrich et al. 2004].

Bit plane image data hiding in spatial domain is perhaps one of the most popular techniques. As noted in [Holotyak et al. 2005], several data hiding software available on the Internet free of charge take this approach. There are several reasons for the popularity of bit plane embedding including the following:

- simple encoding and decoding procedures
- perceptual transparency
- historical reasons.

In spatial domain bit plane embedding (e.g., [Eason and Kawaguchi 1998]), a bit plane in which the message will be embedded is first chosen. Then, based on a secret key, the message bits replace the host image bits. The decoder, using the secret key, extracts the message bits and decodes the message itself. There is obviously a trade-off between the error resilience (caused by an active warden attack) and perceptual transparency in the choice of a bit plane for message embedding. The least significant bit (LSB) embedding causes least perceptual distortion after embedding but at the same it is most sensitive to a distortion constrained active warden attack who could easily replace all the LSBs causing minimal distortion. Higher order bit plane embedded message is more error-resilient but also causes higher embedding induced distortion. By using several bit planes for embedding, the overall payload can be increased. Finally, we note that some $\pm K$ spatial embedding techniques can be modelled as embedding in higher bit planes.

The main question that we address about spatial domain, bit plane embedded image steganalysis is the following: *can a systematic mathematical approach to estimate the embedded message be developed?*. To develop such a systematic approach we use Bayesian analysis, statistical physics based image models [Besag 1986], maximum likelihood estimation, binary symmetric model and other ideas. We believe that this is the first steganalysis attempt to estimate the embedded message.

The paper is organized as follows. Section 2 describes the bit plane message embedding model. Section 3 describes the Bayes approach to the problem and gives an overview on the preliminaries of statistical physics. Section 4 gives a framework for the message estimation procedure and the hyper-parameter estimation algorithm. Section 5 gives details about the steganalysis experiments using the proposed approach. Conclusions are presented in Section 6.

2. EMBEDDING MODEL

Consider a binary image on the square lattice $\Omega = \{i = (x, y) | x = 1, 2, \dots, m; y = 1, 2, \dots, n\}$. The intensities at each pixel location i in the original and the stego images

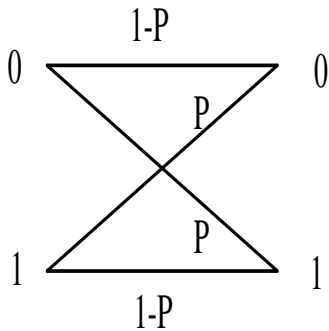


Figure 1: A binary symmetric channel model for bit plane embedding. P is the embedding rate.

are considered to be random variables denoted by F_i and G_i , respectively. The possible states that a pixel can take is ± 1 . The random fields of intensities for the original and the stego image are represented by $F = \{(f_{x,y})|(x,y) \in \Omega\}$ and $G = \{(g_{x,y})|(x,y) \in \Omega\}$.

We assume that the message is embedded in a particular bit plane across the image pixels. Then the stego image is obtained from the original image by changing the intensity of each pixel independently of the other pixels with a probability P . Note that P can be interpreted to be the message embedding rate. Then it is easy to see that the embedding process in a bit plane corresponds to a binary symmetric channel (BSC) as shown in Fig. 1. That is, each host image bit is flipped w.p. P during embedding. The BSC channel is completely specified by the transition probability P . Note that, using P , the length of the embedded message (L) can be computed as $L = P \times \text{image size}$.

3. BAYESIAN FRAMEWORK

To estimate the embedded message, it is enough to estimate the cover image. Bayes theorem relates the unknown host/cover image to the known stego image through a likelihood function. Therefore, using Bayes theorem, the conditional probability of the host image f given the stego image g is expressed as,

$$Pr\{F = f|G = g\} = \frac{Pr\{F = f\}Pr\{G = g|F = f\}}{\sum_z Pr\{F = z\}Pr\{G = g|F = z\}} \quad (1)$$

where the summation is taken over all possible configurations of the image. For example, in the binary bit plane $z_{(x,y)} = \{\pm 1\}$ where -1 stands for bit 0. Also in Eq. (1) we need a model for the prior $Pr(F = f)$ that we address next.

3.1 Markov Random Field

To compute the conditional probability in Eq. (1) knowledge of $Pr(F)$ is required. To obtain a good model for this prior probability of the host image, we exploit the dependency of a pixel on its neighborhood, thus modelling the image lattice as a Markov random field (MRF).

A locally dependent MRF in the spatial domain is a set of random variables representing the intensity of the pixel

(x, y) , dependent only on its neighbors. Even though several orders of dependence can be considered, we use only a first order dependence (neighborhood) model. In this model the set of neighboring pixels for a pixel (x, y) is denoted by,

$$C_{(x,y)} = \{(x \pm 1, y), (x, y \pm 1)\}$$

For $C_{(x,y)}$ to be a neighborhood system it is required that

$$(x, y) \notin C_{(x,y)}$$

and

$$(i, j) \in C_{(x,y)} \Leftrightarrow (x, y) \in C_{(i,j)}$$

For more information on MRF and neighborhood structure we refer to [Besag 1974].

It is proved in [Besag 1974] that the joint probability distribution of these random variables is completely specified by the local conditional probability distributions. It also proves that these local conditional distributions take the form of the energy function of the Ising model which is described later in this section. Then, using *Hammersley and Clifford theorem* [Besag 1986] the joint probability structure associated with these local conditional probability distributions are found to follow the Gibbs distribution. This theorem also implies that F is MRF with respect to C iff F is a Gibbs random field on C . Therefore it allows us to assume that the global joint probability density function of the host image follows the Gibbs distribution.

3.2 Gibbs Distribution and Ising Model

Ising model is the best known and the simplest lattice structure which has been extensively used in studying the order-disorder transitions of a physical system. In [Geman and Geman 1984], an analogy is made between images and lattice structured physical system. Pixel gray levels are viewed as states of a particle in a physical system. The correlation between the neighboring pixels correspond to the interaction between particles.

The Ising model considers an idealized system of say N interacting particles arranged on a planar grid. Each particle can have one of the two states ± 1 . The state of the i th particle is labelled as f_i . In addition to the interaction amongst the neighboring particles there is an external field acting on the system affecting the states of the particle. As Ising model was originally developed to explain the phase transitions in magnetic systems the external field was taken to be a magnetic field B . The external magnetic field corresponds to the distortion induced by message embedding in the image.

The criterion on which the Ising model was developed was that each particle can interact only with its neighbors and the contribution of each particle to the energy of the system depends only on its orientation with respect to its neighbors and the external field. The configuration of the whole system is specified by the states of the individual particles. The total energy of an Ising model is given by,

$$E = -J \sum_{\langle ij \rangle} f_i f_j - B \sum_i f_i \quad (2)$$

where J is the interaction parameter and $\langle ij \rangle$ denotes neighboring particles. This simple model tries to capture

the original states of the particles of the system, whose state has been changed by the alteration in the behavior of the neighboring particles. This concept has been previously applied to image processing [Geman and Geman 1984, Besag 1986].

Let us consider an image lattice of size $N = m \times n$. Let us associate the state of the pixel to its intensity given by $F = f_1, f_2 \dots f_N$. For such an Ising model the energy function $U(F)$ of a N-tuple state F is given by

$$U(F) = -\beta \sum_{\Omega} f_{x,y} - \alpha \left(\sum_{\Omega} f_{x,y} f_{x+1,y} + \sum f_{x,y} f_{x,y+1} \right) \quad (3)$$

where α and β denote the correlation between the neighboring pixels and the effect due to embedding in a pixel, respectively. From (3) it is easy to see that the lowest energy configuration occurs when the neighbors have the same value. According to *Boltzmann's law* the probability that a state F occurs is given by

$$\phi(F) = \frac{e^{(-1/T)U(F)}}{\sum_z e^{((-1/T)U(z))}} \quad (4)$$

which is the Gibbs distribution. Let us denote

$$Z = \sum_z e^{((-1/T)U(z))} \quad (5)$$

where $\sum_z \equiv \prod_{(x,y) \in \Omega} \sum_{z_{(x,y)} = \pm 1}$ and Z is a normalization constant known as *partition function* and T is the temperature. For a non-physical system such as an image we can set T arbitrarily and throughout the rest of the paper we take $T = 1$.

Note that the energy function $U(F)$ in the context of an image lattice is regarded as a function of clique potentials [Geman and Geman 1984] and the Gibbs distribution gives the global probability density function in terms of the local conditional distributions.

3.3 Gibbs Free Energy and Kullback-Leibler Divergence

In this section we will explain the notion of *free energy* and its relevance to Kullback-Leibler divergence. As above consider a system of N particles each of which can be in one of a discrete number of states, where the state of the i th particle is given by f_i . Then the Helmholtz free energy of such a system is defined as [Reichl 1998]

$$F_{Helmholtz} = -\ln Z. \quad (6)$$

This is an important measure because if we can calculate the dependence of this energy on quantities like external force then we can compute the response of the system to a change in the external force. There are several approximation techniques to compute $F_{Helmholtz}$. One popular approximation technique is the *variational technique* [Yedidia 2000]. In this technique we introduce a trial probability, say $\rho(f)$, and the corresponding *variational energy* or *Gibbs free energy* is given by

$$F(\rho) = E(\phi) - S(\phi) \quad (7)$$

where $E(\phi)$ is the *variational average energy* and it is given by:

$$E(\phi) = \sum_{f \in N} \rho(f) U(f) \quad (8)$$

and $S(\phi)$ represents the *variational entropy* given by,

$$S(\phi) = - \sum_{f \in N} \rho(f) \ln(\rho(f)). \quad (9)$$

It follows from the above definitions that

$$F(\rho) = F_{Helmholtz} + D(\rho || \phi)$$

where

$$D(\rho || \phi) = \sum_{f \in F} \rho(f) \ln \frac{\rho(f)}{\phi(f)}$$

Following a theorem in information theory [T.M.Cover and J.A.Thomas 1991] this divergence is always non-negative which implies that *variational free energy* is greater than the *Helmholtz free energy* and the divergence is zero only when the trial probability distribution equals the true probability distribution. Therefore minimizing *Gibbs free energy* implies returning to the original state of the system.

4. STEGANALYSIS ESTIMATION FRAMEWORK

The a priori probability distribution $Pr\{F = f\}$ and the conditional distribution $Pr\{G = g | F = f\}$ are modelled with hyperparameters α and β respectively. If the *a priori* probability distribution that the original image is f is assumed to be,

$$\begin{aligned} Pr(F = f | \alpha) &= \frac{\exp(-1/2\alpha \sum_{(x,y) \in \Omega} (f_{x,y} - f_{x,y+1})^2 + (f_{x,y} - f_{x+1,y})^2)}{\sum_z \exp(-1/2\alpha \sum_{(x,y) \in \Omega} (z_{x,y} - z_{x,y+1})^2 + (z_{x,y} - z_{x+1,y})^2)} \end{aligned} \quad (10)$$

then the conditional distribution $Pr\{G = g | F = f\}$ is given by

$$Pr\{G = g | F = f, \beta\} = \frac{\exp(-1/2\beta \sum_{(x,y) \in \Omega} (f_{x,y} - g_{x,y})^2)}{(1 + \exp(-2\beta))^{|\Omega|}}. \quad (11)$$

Here α and β are the hyperparameters of the prior and the embedding process. By substituting eqns (10) and (11) in (1) the *a posteriori* probability distribution written as a spin-1/2 Ising model is,

$$Pr\{F = f | G = g, \alpha, \beta\} = \frac{\exp(-U(f|g, \alpha, \beta))}{\sum_z \exp(-U(z|g, \alpha, \beta))} \quad (12)$$

where

$$\begin{aligned} U(f|g, \alpha, \beta) &= -\beta \sum_{(x,y) \in \Omega} g_{x,y} f_{x,y} - \alpha \sum_{(x,y) \in \Omega} (f_{(x,y)} f_{(x+1,y)} + (f_{(x,y)} f_{(x,y+1)})) \end{aligned} \quad (13)$$

we denote

$$Pr\{G = g | \alpha, \beta\} = \sum_z Pr\{G = g | F = z, \beta\} Pr\{F = z | \alpha\} \quad (14)$$

as *evidence*. Evidence in terms of Ising model is expressed as

$$Pr\{G = g|\alpha, \beta\} = \frac{\sum_z (\exp(-U(z|g, \alpha, \beta)))}{\sum_z (\exp(-U(z|g, \alpha, \beta = 0))) + (1 + \exp(2\beta))^\Omega} \quad (15)$$

The steganalysis problem is then to estimate these hyperparameters that would maximize this evidence or in other words minimize the Gibbs free energy.

4.1 Hyperparameter Estimation

From (15) the negative logarithm of evidence is given by

$$\begin{aligned} -\ln(Pr(G = g|\alpha, \beta)) &= \ln\left[\sum_z \exp(-U(z|g, \alpha, \beta))\right] \\ &+ \ln\left[\sum_z \exp(-U(z|g, \alpha, \beta = 0))\right] + \Omega \ln(2 \cosh(\beta)) \end{aligned} \quad (16)$$

Therefore the maximum likelihood estimate of the hyperparameters are given by

$$(\hat{\alpha}, \hat{\beta}) = \arg \max_{(\alpha, \beta)} Pr\{G = g|\alpha, \beta\} \quad (17)$$

Then it can be shown that the conditions for extremum at $\hat{\alpha}, \hat{\beta}$ can be reduced to the following equations:

$$\sum_{(x,y) \in \Omega} \sum_z g_{(x,y)} z_{(x,y)} Pr\{F = z|G = g, \hat{\alpha}, \hat{\beta}\} = \tanh(\hat{\beta}) \quad (18)$$

$$\begin{aligned} &\sum_{(x,y) \in \Omega} \sum_z z_{(x,y)} z_{(x+1,y)} + z_{(x,y)} z_{(x,y+1)} Pr\{F = z|G = g, \alpha, \beta\} \\ &= \sum_{(x,y) \in \Omega} \sum_z z_{(x,y)} z_{(x+1,y)} + z_{(x,y)} z_{(x,y+1)} Pr\{F = z|\hat{\alpha}\} \end{aligned} \quad (19)$$

In the next section we show that these two equations can be further simplified.

4.2 Bethe Approximation and Gibbs Free Energy Minimization

Minimizing Gibbs free energy is computationally intractable and there are several approximation techniques to handle this problem. Popular among them are mean field technique [Yedidia 2000] and Bethe approximation [R.Kikuchi 1951].

Mean field approximation is a function of only one body marginal distribution while Bethe approximation considers both one and two body marginal distributions. Therefore the host image estimates obtained from Bethe approximation are better than the the mean field technique.

Following along the lines of Bethe approximation we express the joint probability distribution $P(f)$ in terms of marginal

distributions involving local neighborhoods:

$$\begin{aligned} P(f) &\simeq \left(\prod_{(x,y) \in \Omega} P_{(x,y)}(f_{(x,y)}) \right) \\ &\times \prod_{(x,y) \in \Omega} \frac{P_{(x,y)}^{(x+1,y)}(f_{(x,y)}, f_{(x+1,y)})}{P_{(x,y)}(f_{(x,y)}) P_{(x+1,y)}(f_{(x+1,y)})} \\ &\times \prod_{(x,y) \in \Omega} \frac{P_{(x,y)}^{(x,y+1)}(f_{(x,y)}, f_{(x,y+1)})}{P_{(x,y)}(f_{(x,y)}) P_{(x,y+1)}(f_{(x,y+1)})} \end{aligned} \quad (20)$$

Let us define the entropy $S(\rho)$ associated with the Gibbs free energy $F(\rho)$ to be

$$S(\rho) = - \sum_z \rho(z|g) \ln \rho(z|g) \quad (21)$$

Expressing the entropy defined above in terms of Bethe approximation we get,

$$\begin{aligned} S(\rho) &\simeq \sum_{(x,y) \in \Omega} (-S(\rho_{(x,y)})) \\ &+ \sum_{(x,y) \in \Omega} (S(\rho_{(x,y)}^{(x+1,y)}) - S(\rho_{(x,y)}) - S(\rho_{(x+1,y)})) \\ &+ \sum_{(x,y) \in \Omega} (S(\rho_{(x,y)}^{(x,y+1)}) - S(\rho_{(x,y)}) - S(\rho_{(x,y+1)})) \end{aligned}$$

which implies

$$\begin{aligned} S[\rho] &= -3 \sum_{(x,y) \in \Omega} S[\rho_{(x,y)}] + \sum_{(x,y) \in \Omega} S[\rho_{(x,y)}^{(x+1,y)}] \\ &+ \sum_{(x,y) \in \Omega} S[\rho_{(x,y)}^{(x,y+1)}] \end{aligned} \quad (22)$$

Substituting (22) in the expression for Gibbs free energy $F(\rho) = \sum_z \rho(z|g) (\ln(\rho(z|g)) U(z|g))$ we get an approximate expression for the Free energy in terms of the local marginal probability distributions called the Bethe Free Energy (F_{Bethe}). By taking the first variations of the Bethe free energy with respect to the marginal distributions we can express the marginal distributions in terms of Lagrange multipliers. The Lagrange multipliers can be formulated as the influence of the pixels within a clique (termed as *message* in belief propagation) and it is given by,

$$\begin{aligned} &\lambda_{(x,y \pm 1)}^{(x,y)} \\ &= \tanh^{-1}(\tanh(\alpha) \tanh(\beta g_{(x,y)} + \sum_{(x',y') \in c_{(x,y)} \setminus \{(x,y \pm 1)\}} \lambda_{(x',y')}^{(x',y')})) \end{aligned} \quad (23)$$

$$\begin{aligned} &\lambda_{(x \pm 1,y)}^{(x,y)} \\ &= \tanh^{-1}(\tanh(\alpha) \tanh(\beta g_{(x,y)} + \sum_{(x',y') \in c_{(x,y)} \setminus \{(x \pm 1,y)\}} \lambda_{(x',y')}^{(x',y')})) \end{aligned} \quad (24)$$

The one body and two body marginal distributions obtained from the variational approach of [Tanaka 2002] are given below:

$$\rho_{(x,y)}(\xi) = \frac{\exp(-U_{x,y}(\xi))}{\sum_{\zeta=\pm 1} \exp(-U_{(x,y)}(\zeta))} \quad (25)$$

$$\rho_{(x,y)}^{(x+1,y)}(\xi, \xi') = \frac{\exp(-U_{(x,y)}^{(x+1,y)}(\xi, \xi'))}{\sum_{\zeta=\pm 1} \sum_{\zeta'=\pm 1} \exp(-U_{(x,y)}^{(x+1,y)}(\zeta, \zeta'))} \quad (26)$$

$$\rho_{(x,y)}^{(x,y+1)}(\xi, \xi') = \frac{\exp(-U_{(x,y)}^{(x,y+1)}(\xi, \xi'))}{\sum_{\zeta=\pm 1} \sum_{\zeta'=\pm 1} \exp(-U_{(x,y)}^{(x,y+1)}(\zeta, \zeta'))} \quad (27)$$

where

$$U_{(x,y)}(\xi) \equiv -(\beta g_{(x,y)} + \sum_{(x',y') \in C_{(x,y)}} \lambda_{(x,y)}^{(x',y')}(g, \alpha, \beta)) \xi \quad (28)$$

$$U_{(x,y)}^{(x+1,y)}(\xi, \xi') = -\alpha \xi \xi' + U_{(x,y)}(\xi) + U_{(x+1,y)}(\xi') \quad (29)$$

$$U_{(x,y)}^{(x,y+1)}(\xi, \xi') = -\alpha \xi \xi' + U_{(x,y)}(\xi) + U_{(x,y+1)}(\xi') \quad (30)$$

Using these marginal probability distributions we can express (18) and (19) as

$$\sum_{(x,y) \in \Omega} \sum_{\zeta=\pm 1} g_{(x,y)} \zeta \rho_{(x,y)}(\zeta | g, \alpha, \beta) = \tanh(\beta) \quad (31)$$

$$\begin{aligned} & \sum_{(x,y) \in \Omega} \sum_{\zeta=\pm 1} \sum_{\zeta'=\pm 1} \zeta \zeta' \rho_{(x,y)}^{(x+1,y)}(\zeta, \zeta' | g, \alpha, \beta) \\ & + \rho_{(x,y)}^{(x,y+1)}(\zeta, \zeta' | g, \alpha, \beta) \\ & = \sum_{(x,y) \in \Omega} \sum_{\zeta=\pm 1} \sum_{\zeta'=\pm 1} \zeta \zeta' \rho_{(x,y)}^{(x+1,y)}(\zeta, \zeta' | g, \alpha, \beta = 0) \\ & + \rho_{(x,y)}^{(x,y+1)}(\zeta, \zeta' | g, \alpha, \beta = 0) \end{aligned} \quad (32)$$

where $\zeta, \zeta' = \pm 1$. Now let us define

$$m_{x,y}^{x',y'} = \tanh(\beta g_{x,y} + \sum_{(x',y') \in c_{x,y}} \lambda_{x,y}^{x',y'}).$$

Then using (15), (18) can be rewritten as

$$\beta = \tanh^{-1} \left(\sum_{(x,y) \in \Omega} g_{x,y} m_{x,y}^{x',y'} \right). \quad (33)$$

and (19) becomes

$$\begin{aligned} & \frac{\exp(2\alpha) \cosh(6\lambda(\alpha)) - 1}{\exp(2\alpha) \cosh(6\lambda(\alpha)) + 1} \\ & = \frac{1}{\Omega} \sum_{(x,y) \in \Omega} \tanh^{-1} [\exp\{2\alpha - 2 \tanh^{-1}(\tanh(\beta g_{x,y} \\ & + \sum_{(x',y') \in c_{x,y} \setminus (x+1,y)} \lambda_{x,y}^{(x',y')})\})] \\ & \times \tanh(\beta g_{x+1,y} + \sum_{(x',y') \in c_{x+1,y} \setminus (x,y)} \lambda_{x+1,y}^{(x',y')}) \\ & + \tanh^{-1} [\exp\{2\alpha - 2 \tanh^{-1}(\tanh(\beta g_{x,y} \\ & + \sum_{(x',y') \in c_{x,y} \setminus (x,y+1)} \lambda_{x,y}^{(x',y')})\})] \\ & \times \tanh(\beta g_{x,y+1} + \sum_{(x',y') \in c_{x,y+1} \setminus (x,y)} \lambda_{x,y+1}^{(x',y')}) \end{aligned} \quad (34)$$

where $\lambda(\alpha)$ is the effective field in the absence of external field (secret message).

4.3 Belief Propagation

Belief propagation is an extremely powerful approach used in probabilistic inference problems, especially in artificial intelligence. In a general inference problem it is required to estimate the state of the hidden variables given the states of the rest of the variables in the system.

In image steganalysis this inference problem corresponds to estimating the intensity of a host pixel given the intensities of the stego neighbors. The estimated intensity is termed as the *belief* we have about the intensity of the host pixel given its neighbors. The term *message* in the belief propagation algorithm corresponds to the effective field λ in Bethe approximation. Having estimated the state of the pixel, this belief about the estimated pixel is propagated throughout the lattice and it is used in estimating the intensities of other pixels.

It is proved in [J.S.Yeddidia and Y.Weiss 2004] that the fixed points in the belief propagation algorithm corresponds to the stationary points of Bethe free energy. This fact is exploited in designing the steganalysis algorithm that follows.

4.4 Steganalysis Algorithm

1. Initialize $t = 0$ and $\lambda(0) = 1$

2. Update $t = t + 1$

3. Compute

$$\begin{aligned} & \lambda_{(x \pm 1, y)}^{(x, y)}(t + 1) \\ & = \tanh^{-1}(\tanh(\alpha) \tanh(\beta g_{(x, y)} + \sum_{(x', y') \in c_{(x, y)}^{(1)}} \lambda_{(x, y)}^{(x', y')}(t))) \end{aligned}$$

where $c_{(x, y)}^{(1)} \in c_{(x, y)} \setminus (x \pm 1, y)$

$$\begin{aligned} & \lambda_{(x, y \pm 1)}^{(x, y)}(t + 1) \\ & = \tanh^{-1}(\tanh(\alpha) \tanh(\beta g_{(x, y)} + \sum_{(x', y') \in c_{(x, y)}^2} \lambda_{(x, y)}^{(x', y')}(t))) \end{aligned}$$

where $c_{(x,y)}^{(2)} \in c_{(x,y)\setminus(x,y\pm 1)}$

4. If

$$(\lambda_{(x,y\pm 1)}^{(x,y)}(t+1) - \lambda_{(x,y\pm 1)}^{(x,y)}(t)) + (\lambda_{(x\pm 1,y)}^{(x,y)}(t+1) - \lambda_{(x\pm 1,y)}^{(x,y)}(t))$$

averaged over all pixels is less than a threshold value go to step 5 else go to step 2 after updating α and β .

5. For a fixed value of α and β calculate the one body and two-body marginal probability distributions given by (25), (26) and (27)
6. With these distributions calculate the evidence.
7. With the evidence search for α and β that satisfies (33) and (34) deterministically.
8. Generate an estimate of the host (\hat{F}) using the estimated hyperparameters.
9. Estimate the secret message by comparing the \hat{F} and G .

5. EXPERIMENTAL RESULTS

For experiments, 8 bit, 256-gray level Lena host image (Fig. 2) of size 154×103 was used. In this paper we consider higher bit plane embedding and not the first bit plane (LSB). This is because the correlation between the neighboring pixels is negligible in LSB plane. Since the proposed algorithm exploits neighborhood correlation it is not directly applicable in this bit plane. We are currently exploring methods to extend the approach for LSB steganalysis by exploiting inter bit plane correlations. We however note that embedding in higher bit planes is similar to $\pm K$ embedding, where K is an integer. And, the proposed method may also be applied to estimate messages embedded using this technique.

Fig. 3 shows the fifth, fourth and third bit planes of the host image. The presence of a structure in these bit planes can be noted in these figures. Stego images were obtained by embedding messages of different lengths in these bit planes. The embedding probability (P) was computed as the ratio of the message size to the host image size. Fig. 4 shows the stego versions of the host image when $P = 0.1$. Even though 5th bit plane embedding causes significant distortions, we have included it here only for the sake of comparison. The proposed estimation algorithm was run on these stego images and Fig. 5 shows the estimated host images. The performance of the steganalysis algorithm is judged based on the *false alarm* defined as the probability that a bit is estimate to be a message bit given that it is actually not and *detection* probability defined as the probability that a message bit is estimated given that it is indeed a message bit. Note that probability of miss is equal to $1 - \text{Pr}(\text{detection})$.

Fig. 6 shows the detection probability for various probabilities of embedding. Fig. 7 shows the corresponding false alarm probabilities. We see from these figures that the proposed algorithm works well for smaller embedding probabilities. Note that a small embedding probability does not necessarily mean small message size. We observe that the probability of false alarm increases with P while the detection probability decreases. The reason for this is the type of statistical structure exploited by the estimation algorithm.

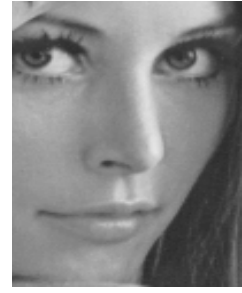


Figure 2: 8-bit, 256-gray level Lena host image.

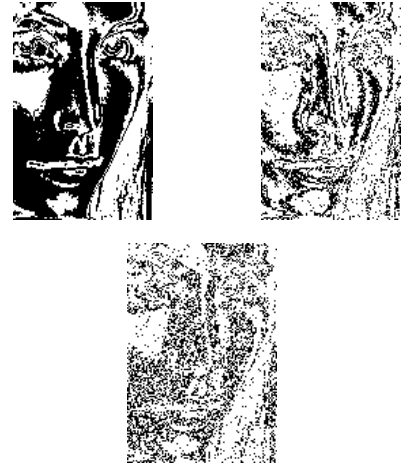


Figure 3: 5th, 4th, 3rd bit planes of Lena image.

A higher value of P means that the binary symmetric channel that models the embedding process is more noisy. This in turn makes the estimation process less reliable for the steganalysis estimator. Also, higher value of P randomizes the statistical structure thus making it difficult to estimate the host image. We can also observe that the performance degrades drastically when the embedding probability is 0.25. Embedding with a probability of 0.25 generally implies that for every host pixel that is not used for embedding there are 4 pixels used for embedding. There is a high probability that these four embedded pixels are the neighbors of the pixel to be estimated. Therefore the steganalysis algorithm which estimates the pixel intensity by looking at its neighbors fails to give a correct estimate as all the four neighbors are affected.

6. CONCLUSIONS

A binary symmetric channel model for bit plane message embedding is used to design a image steganalysis approach. The role of Bayesian statistics and statistical mechanics are explored. The goal of the proposed steganalysis algorithm is the estimation of the embedded message. Bit plane image data hiding is considered for experimental evaluation of the theoretical analysis. It is observed that the steganalysis approach performs well for lower message embedding rates. Note that a small message embedding rate could translate into a large message size if the host image size is large. The method does not work for LSB embedding due

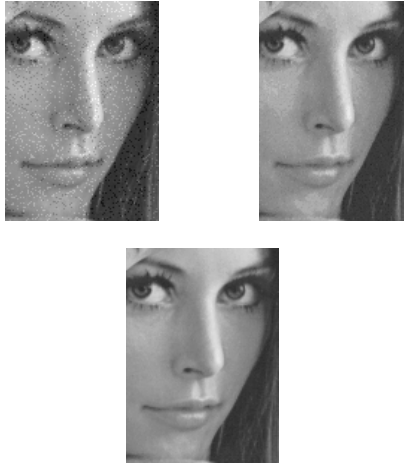


Figure 4: Lena with secret message embedded in 5th, 4th, and 3rd bit planes.



Figure 5: Estimated 5th, 4th, 3rd bit planes of original Lena.

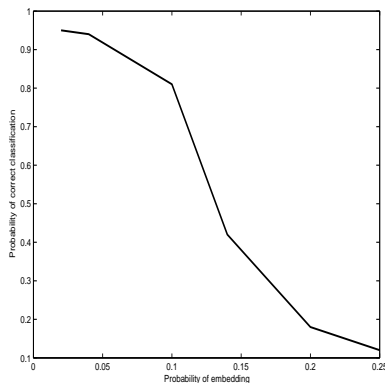


Figure 6: Probability of correct estimation of message bit.

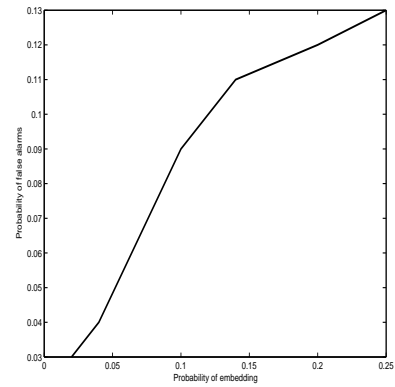


Figure 7: Probability of false alarm.

to the lack of statistical structure in this bit plane. Part of our future work is to extend this technique to LSB secret message extraction by exploiting inter-bit plane statistical structure.

Acknowledgement

This work was supported by a grant from the U.S. Air Force Research Laboratory, Rome, NY.

7. REFERENCES

- BESAG, J. 1974. Spatial interaction and statistical analysis of lattice systems. *Journal of the Royal Statistical Society B* 36, 192–225.
- BESAG, J. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B* 48, 259–302.
- CHANDRAMOULI, R. 2003. A mathematical framework for active steganalysis. *ACM Multimedia Systems* 9, 3 (Sept.), 303–311.
- EASON, R. AND KAWAGUCHI, E. 1998. The principle and applications of bpcs-steganography. *SPIE International Symposium on Voice, Video, and Data Communications: Multimedia Systems and Applications*, 464–473.
- FRIDRICH, J., GOLJAN, M., SOUKAL, D., AND HOLOTYAK, T. 2004. Forensic steganalysis: Determining the stego key in spatial domain steganography. *Proc. SPIE EI*.
- GEMAN, S. AND GEMAN, D. 1984. Stochastic relaxation, gibbs distributions and bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 721–741.
- HOLOTYAK, T., FRIDRICH, J., AND SOUKAL, D. 2005. Stochastic approach to secret message length estimation in $\pm k$ embedding steganography. *Proc. SPIE EI*.
- J.S.YEDDIDIA, W. AND Y.WEISS. 2004. Constructing free energy approximations and generalized belief propagation algorithms. *Mitsubishi Electric Research Labs Report TR2004-040*.

- REICHL, L. 1998. *A modern course in statistical physics* (2 ed.). J. Wiley and Sons.
- R.KIKUCHI. 1951. A theory of cooperative phenomena. *Physical Review* 81, 988–1003.
- TANAKA, K. 2002. Statistical mechanical approach to image processing. *Journal of the Physics A:Math.Gen.*, R81–R150.
- T.M.COVER AND J.A.THOMAS. 1991. Spatial interaction and statistical analysis of lattice systems.
- TRIVEDI, S. AND CHANDRAMOULI, R. 2005. Secret key estimation in sequential steganography. *Supplement on Secure Media, IEEE Trans. on Signal Processing* 53, 2 (Feb.), 746–757.
- YEDIDIA, J. 2000. An idiosyncratic journey beyond mean field theory. Technical report, Mitsubishi Electric Research Laboratory.