CURRENT TRENDS IN STEGANALYSIS: A CRITICAL SURVEY

R. Chandramouli and K.P. Subbalakshmi

Department of ECE
Stevens Institute of Technology

ABSTRACT

This paper presents a critical analysis of some of the current steganalysis methodologies. The pros and cons of these methods are discussed from statistical and usability perspectives. It is concluded that no single strategy works best. Depending on the amount of statistical information available at hand, a proper choice has to be made.

1. INTRODUCTION

While steganography deals with techniques for hiding information (such as watermarking), the goal of steganalysis is to detect and/or estimate potentially hidden information from observed data with little or no knowledge about the steganography algorithm and/or its parameters. It is fair to say that steganalysis is both an art and a science. The art of steganalysis plays a major role in the selection of features or characteristics a typical stego message might exhibit while the science helps in reliably testing the selected features for the presence of hidden information. While it is possible to design a reasonably good steganalysis technique for a specific steganographic algorithm, the long term goal is to develop a steganalysis framework that can work effectively at least for a class of steganography methods, if not for all. Current trend in steganalysis seems to suggest two extreme approaches: (a) little or no statistical assumptions about the image under investigation. Statistics are learnt using a large database of training images and (b) a parametric model is assumed for the image and its statistics are computed for steganalysis detection.

In this paper we discuss image steganalysis though many of the techniques are applicable to other data types as well. Several approaches have been proposed to solve the steganalysis problem and we broadly classify them into the following groups:

- **Supervised learning based steganalysis [1, 2, 3]:** Supervised learning based steganalysis techniques employ two phase strategies: (a) training phase and (b) testing phase. In the training phase, examples of the type $\{(d_i, t_i)\}$ where d_i denotes a stego image feature(s) and t_i denotes whether a secret message is embedded or not, are provided to a statistical classifier. The classifier "learns" the best classification rule using these examples. In the testing phase unknown images are given as input to the trained classifier to decide whether a secret message is present or not. There are some steganalysis methods (e.g., [3]) that do not directly use this type of classical learning by example rather training data is used to compute a regression model for a set of selected features. This model is then used for steganalysis.

- **Blind identification based steganalysis [4]:** Blind identification methods pose the steganalysis problem as a system identification problem. Some statistical properties such the independence of host and secret message etc. are exploited. The embedding algorithm is represented as a channel and the goal is to invert this channel to identify the hidden message.

- **Parametric statistical steganalysis [5, 6, 7, 8]:** These approaches tend to assume a certain parametric statistical model for the cover image, stego image and the hidden message. Steganalysis is formulated as a hypothesis testing problem, namely, H_0 :no message (null hypothesis) and H_1 :message present (alternate hypothesis). A statistical detection algorithm is then designed to test between the two hypotheses.

- **Hybrid techniques**: Hybrid techniques overlap more than one of the above approaches.

The type and amount of information needed for successful steganalysis is a critical issue. The following two information types for steganalysis have been identified in [4]:

- **Spatial diversity information based steganalysis**: Steganalysis methods can look for information in the spatial domain that repeats itself in various forms in

R. Chandramouli was supported by an U.S. AFRL grant. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposed notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory or the U.S. Government.

different spatial locations (e.g., different blocks within an image or, in different images). We call this spatial diversity based steganalysis.

- **Temporal diversity information based steganalysis**: Steganography information that appears repeatedly over time can also aid steganalysis. Such techniques are called temporal diversity information based steganalysis, e.g., video steganalysis.

Clearly, it is important to choose a proper steganalysis domain, appropriate features, statistical models and parameters, detector design, user inputs such as detection error probability etc. We discuss later some of the popular choices of current steganalysis algorithms in this regard.

The paper is organized as follows. The pros and cons of supervised learning based steganalysis are presented in Section 2, blind identification based steganalysis is discussed in Section 3 and parametric techniques are presented in Section 4. Concluding remarks are given in Section 5.

2. SUPERVISED LEARNING BASED STEGANALYSIS

Supervised learning methods construct a classifier to differentiate between stego and non-stego images using training examples. Some image features are first extracted and given as training inputs to a learning machine. These examples include both stego as well as non-stego messages. The learning classifier iteratively updates its classification rule based on its prediction and the ground truth. Upon convergence the final stego classifier is obtained.

Some factors in favour of this class of steganalysis algorithms are the following.

- Learning based steganalysis has been observed to perform quite well using features such as wavelet coefficient statistics, image quality metrics etc.

- By training the classifier for a specific embedding algorithm a reasonably accurate detection can be achieved. Since the classifier is given multiple examples there is no need to assume prior statistical models for the images. The classifier learns a model by averaging over multiple examples.

- Universal steganalysis detectors can be constructed using learning techniques.

- Non-stationarity of images do not pose a major problem due to the averaging process.

- Since machine learning has been an active research area for several years, there is a well developed theory and general methodology.

- Several freely available software packages on the Internet could be directly used to train a steganalysis detector.

This type of steganalysis detectors are limited by several factors such as the following.

- A separate classifier has to be trained for each embedding algorithm. This could be time consuming and sometimes impractical.

- Choice of proper features to train the classifier upon is a critical step. If the selected features are not appropriate for the specific embedding algorithm then the detector may completely fail. There is no systematic rule for feature selection. It is mostly a heuristic, trial and error method.

- Some classifiers have several parameters that have to be chosen by the steganalyst. For example, what type of kernels to choose, learning rate, linear or non-linear classifier, how many iterations to run the training phase before terminating it, how large a training set to choose, what type of training set to choose, etc. This could be a daunting task. Again, there is no straightforward manner in which these parameters could be chosen. It is also mostly a trail and error process.

- Any training based method suffers from the classical bias versus variance trade-off. That is, the classifier can be trained very well to given very high accuracy for the training images but may loose the generalization capability to perform on test images.

- False alarm and miss probabilities are not controllable by the steganalyst. That is, the steganalyst cannot achieve a desired false alarm and miss probability.

- It is extremely difficult or even impossible to identify portions of the image where a message is hidden, message extraction etc. The ultimate goal of learning steganalyzers is to arrive at a binary decision—presence or absence of a secret message.

3. BLIND IDENTIFICATION BASED STEGANALYSIS

Let $\mathbf{z}(k)$ denote a random stego message vector observed by the steganalyst, $\mathbf{A}$ be a representation of the embedding algorithm in matrix form (e.g., embedding message strength matrix, etc.), and $\mathbf{r}$ is the vector with the cover message and the secret message as its components. The steganalyst is now faced with the problem of inferring $\mathbf{A}^{-1}$ from $\mathbf{z}(k)$. This can be viewed as a blind system identification problem as shown in Figure 1. If $\mathbf{A}^{-1}$ can be identified then we can obtain an estimate of $\mathbf{r}(k)$, say, $\mathbf{r}_1(k)$, *i.e.*, the steganalysis

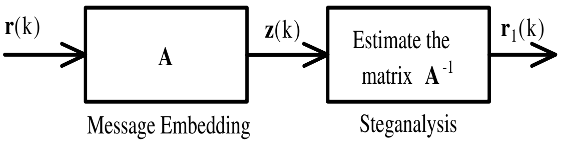

Fig. 1. Steganalysis as a blind system identification problem.

problem is to find a linear transform such that the components of $\mathbf{r}(k)$ can be retrieved. We also notice the similarity between this version of steganalysis and a blind source separation (BSS) problem [9].

Some of the advantages of using a blind system identification approach to steganalysis are the following.

- In this formulation of steganalysis we note that there is no training data. Each image is analyzed individually based on the computed statistics. This is good in the sense that the estimated true statistics of the image are available to the steganalysis detector rather than an average as in learning based steganalysis. Therefore the computed statistics reflect the characteristics of the image more accurately.
- It is possible to extract the hidden message [4] rather than a simple detection of its presence or absence.
- Since the blind system identification framework is quite general several stego embedding algorithms can be detected by modelling them within this framework.
- It is possible to derive analytical results that suggest the feasibility of successful steganalysis for certain types of statistical models for the original image and the secret message. For instance, it is shown in [4] that for the linear spread spectrum message embedding in the discrete cosine transform domain the following identifiability conditions must hold:
  - At least the discrete cosine transform coefficients of the host image or the message carrier must be non-Gaussian.
  - The matrix $\mathbf{A}$ must be of full column-rank.

While the advantages are several as described above there are also some problems with this type of steganalysis as discussed below.

- Digital images are known to be statistically non-stationary. This causes practical issues in implementing algorithms based on the blind identification model since blind identification inherently assumes stationarity of data.
- When the stationarity condition is violated additional effort is needed to make steganalysis work. This may need some heuristic approaches such as moving window based statistics computation, piece-wise stationarity assumption etc.
- If the message embedding algorithm is nonlinear then the blind identification problem becomes more difficult. Additionally, computation of several higher order statistics may be necessary for successful inverse computation.
- If the assumptions on prior statistical models for the host image, stego image and the secret message are not accurate, then there could be a severe performance loss.

4. PARAMETRIC STATISTICAL DETECTION BASED STEGANALYSIS

Using parametric statistical detection techniques several cases of steganalysis can be studied. Specifically the following cases can be investigated:

- **Completely known statistics:** This case arises as a result of Kerchoff's principle where the assumption is that, the stego embedding algorithm is made public and only the secret key is not. Therefore, the image statistics are completely available to the steganalysis detector.
- **Partially known statistics:** A (noisy) estimate of statistics may be obtained using a large training set obtained before and after embedding when the stego embedding algorithm itself may only be known as a black box (e.g., only the executable code of a steganography software may be available.).
- **Completely unknown statistics:** This is true for applications such as steganographic covert communications where only the stego image may be available to the steganalysis detector with no further knowledge.

For the completely known statistics case the parametric models for the stego image, host image and the secret message are accurately known. For the partially known case, the parametric probability models are available but not the parameters themselves. These parameters can be estimated. Finally, for the completely unknown statistics case it is possible to assume Bayesian prior models and then develop detectors.

Assuming that a parametric probability distribution model is available to the steganalysis detector we note the following advantages in this class of steganalysis techniques:

- Parametric statistical detection theory is a well developed subject area. Therefore many of the known results in this area can be applied in a straightforward manner to investigate steganalysis detection rules.

- Receiver operating characteristic completely specifies the performance of the steganalysis detector. This is a curve with false alarm probability on the X-axis and detection probability on the Y-axis. Therefore the achievable error rates can be easily deducted. Depending on the user preference the steganalysis detector can be made to operate on point on the receiver operating characteristic curve.
- A steganalyst has control over the desired detection error probability. The detection thresholds can be computed in closed-form for a given error probability constraint. Sometimes it may be even possible to specify constraints on both false alarm and detection probability [5].
- Estimating secret key, message locations, message length etc. is also possible [5].

Some the of drawbacks of parametric statistics based steganalysis detection are the following.

- By nature, parametric steganalyzers are sensitive to inaccuracies in statistical estimates of certain parameters. That is the steganalyzers performance could suffer if the estimated statistics do not truely reflect the image statistics.
- Assuming probabilistic priors is a contentious issue. Priors are typically subjectively chosen. Therefore, this involves a higher degree of user involvement in the steganalysis process.
- Statistical non-stationarity of digital images pose a serious practical problem.

5. CONCLUSION

There are two extremes in current steganalysis detection algorithms: (a) techniques that assume no statistical information about the stego image, host image and the secret message and (b) techniques that make significant assumptions about the statistics. Machine learning theory based steganalysis is a popular choice for the first class of detection algorithms and parametric statistical detection for the second class. Each of these methodologies have pros and cons. Therefore, it is up to the user (steganalyst) to choose an appropriate methodology based on the amount of side information that is available a priori.

6. REFERENCES

[1] I. Avcibas, N. Memon, and B. Sankur, “Steganalysis using image quality metrics,” *IEEE Trans. on Image Processing*, vol. 12, no. 2, pp. 221–229, Feb. 2003.

[2] S. Lyu and H. Farid, “Steganalysis using color wavelet statistics and one-class support vector machines,” *SPIE Symposium on Electronic Imaging*, 2004.

[3] J. Fridrich, R. Du, and M. Long, “Steganalysis of lsb encoding in color images,” *IEEE ICME*, 2000.

[4] R. Chandramouli, “A mathematical framework for active steganalysis,” *ACM Multimedia Systems*, vol. 9, no. 3, pp. 303–311, September 2003.

[5] S. Trivedi and R. Chandramouli, “Secret key estimation in sequential steganography,” *Supplement on Secure Media, IEEE Trans. on Signal Processing*, Feb. 2005.

[6] J. Harmsen and W. Pearlman, “Steganalysis of additive noise modelable information hiding,” *Proc. SPIE Electronic Imaging*, 2003.

[7] N. Provos and P. Honeyman, “Detecting steganographic content on the internet,” .

[8] A. Westfeld and A. Pfitzmann., “Attacks on steganographic systems,” *Third Information Hiding Workshop*, Sept. 1999.

[9] J.-F Cardoso, “Blind signal separation: statistical principles,” *Proceedings of the IEEE*, vol. 90, no. 8, pp. 2009–2026, Oct. 1998.