IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Improvements of Steganography Parameter in Binary Images and JPEG Images against Steganalysis

Manu Devi[*1], **Ms. Nidhi Sharma**[2]
[*1, 2] Department of Computer Engineering, Mahrashi Dayanand University,Rohtak
The Technological Institute of Textile & Sciences, Bhiwani, Haryana, India
Manughanghas26@gmail.com

Abstract

Steganography is a science of hiding messages into multimedia documents. A message can be hidden in a document only if the content of a document has high redundancy. Although the embedded message changes the characteristics and nature of the document, it is required that these changes are difficult to be identified by an unsuspecting user. On the other hand, steganalysis develops theories, methods and techniques that can be used to detect hidden messages in multimedia documents. The documents without any hidden messages are called cover documents and the documents with hidden messages are named stego documents. The work of this research paper concentrates on image steganalysis. We present four different types of steganalysis techniques. These steganalysis techniques are developed to counteract the steganographic methods that use binary (black and white) images as the cover media. Unlike grayscale and color images, binary images have a rather modest statistical nature. This makes it difficult to apply directly the existing steganalysis on binary images.

Keywords: Gray scale, Compression, Extraction, GLRL (Gray Level Run Length) matrix.

Introduction

Since the rise of the Internet one of the most important factors of information technology and communication has been the security of information. As the number of Internet users rises, the concept of Internet security has also gain importance. The fiercely competitive nature of the computer industry forces web services to the market at a breakneck pace, leaving little or no time for audit of system security, while the tight labor market causes Internet project development to be staffed with less experienced personnel, who may have no training in security. This combination of market pressure, low unemployment, and rapid growth creates an environment rich in machines to be exploited, and malicious users to exploit those machines.

Steganography supports different types of digital formats that are used for hiding the data. These files are known as carriers. Depending upon the redundancy of the object, suitable formats are used. Redundancy is the process of providing better accuracy for the object that is used for display by the bits of object. The main file formats that are used for steganography are Text, images, audio and video. Steganography is also used for the less dramatic purpose of watermarking. The applications of watermarking mainly involve the protection of intellectual property such as ownership protection, file duplication management, document authentication (by inserting an appropriate digital signature) and file annotation. A larger part of steganalysis works published so far deals with grayscale and color images. We consider a less explored area of binary image steganography, which becomes more and more important for electronic publishers, distribution, management of printed documents and electronic libraries. Note that there are two aspects of steganalysis. The first relates to the attempt to break or attack a steganography; the second uses it as an effective way of evaluating and measuring steganography security performance. This work studies steganalysis in terms of the first aspect. In particular, we aim to carry out different levels of analysis to extract the relevant secret parameters. Steganography is an alternative method for privacy and security. Instead of encrypting, we can hide the messages in other innocuous looking medium (carrier) so that their existence is not revealed. Clearly, the goal of cryptography is to protect the content of messages, steganography is to hide the existence of messages. An advantage of steganography is that it can be employed to secretly transmit messages without the fact of the transmission being discovered. Often, cryptography and steganography are used together to achieve higher security.

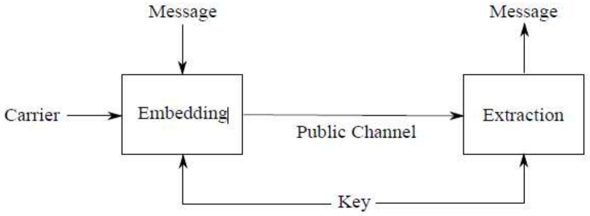


Fig 1: General model of steganography

Steganography can be mathematically defined as follows:

Emb: C ×M × K → S
Ext: S × K → M,

Digital Images

A digital image is produced through a process called digitisation. Digitising an image involves converting analogue information into digital information [1]; thus, a digital image is the representation of an original image by discrete sets of points. Each of these points is called a picture element or pixel. Pixels are normally arranged in a two-dimensional grid corresponding to the spatial coordinates in the original image. The number of distinct colours in a digital image depends on the number of bits per pixel (bpp). Hence, the types of digital image can be classified according to the number of bits per pixel. There are three common types of digital image:

i..Binary image. In a binary image, only one bpp is allocated for each pixel. Since a bit has only two possible states (on or off), each pixel in a binary image must represent one of two colours. Usually, the two colours used are black and white. A binary image is also called a bi-level image.

ii. Greyscale image. A grayscale image is a digital image in which the only colors are shades of grey. The darkest possible shade is black, whereas the lightest possible shade is white. Normally, there are eight bits per pixel assigned for a greyscale image. This creates 256 possible different shades of grey.

iii.Colour image. In general, a pixel in a colour image consists of several primary colours[3]. Red, green and blue are the most commonly used primary colours.

Proposed Steganalysis Method

The ultimate goal of steganalysis is to extract the full hidden message. This task, however, may be very difficult to achieve. Thus, we may start with more realistic and modest goals, such as identifying the type of steganographic technique used for the embedding. We want to improve our existing technique so that we can identify the embedding algorithm.Our analysis includes feature extraction and data classification. The first stage is crucial and we show how to construct the features. The second stage uses the SVM [23] as the classifier. SVM is based on the idea of hyper plane separation between two classes. It obtains an optimal hyperplane that separates the feature set of different classes into different sides of the hyper plane. Based on the separation, the class an image belongs to can be determined.

Grey Level Run Length Matrix

The feature we want to extract from images is based on the grey level run length (GLRL). The length is measured by the number of consecutive pixels for a given grey level g and direction θ. Note that 0 ≤ g ≤ G − 1, G is the total number of grey levels and θ, where 0◦ ≤ θ ≤ 180◦, indicates the direction. The sequence of pixels (at a grey level) is characterised by its length (run length) and its frequency count (run length value), which tells us how many times the run has occurred in the image. Thus, our feature is a GLRL matrix that fully characterises different grey runs in two dimensions: the grey level g and the run length ℓ. The GLRL matrix is defined as follows:

$$\begin{aligned} r(g,\ell|\theta) = \ \# \{(x,y) \mid p(x,y) = p(x+s,y+t) = g; \\ p(x+u,y+v) \neq g; \\ 0 \leq s < u \quad \& \quad 0 \leq t < v; \\ u = \ell\cos(\theta) \quad \& \quad v = \ell\sin(\theta); \\ 0 \leq g \leq G-1 \quad \& \quad 1 \leq \ell \leq L \quad \& \quad 0^\circ \leq \theta \leq 180^\circ\}, \end{aligned}$$

where # denotes the number of elements and p(x, y) is the pixel intensity (grey level) at position x, y. G is the total number of grey levels and L is the maximum run length.

Pixel Differences

The pixel difference is the difference between a pixel and its neighbouring pixels. Given pixel p(x, y) of an image, with x ∈ [1,X] and y ∈ [1, Y], where X and Y are the image width and height, respectively, the vertical difference for the pixel p(x, y) in the vertical direction is defined as follows:

$$pv\ (x', y') = p(x, y + 1) - p(x, y)$$

where x′ ∈ [1,X − 1] and y′ ∈ [1, Y − 1]. The pixel differences in the horizontal, main diagonal and minor diagonal directions are defined similarly.

Gray level co occurrence matrix

We replaced the grey level gap length (GLGL) matrix proposed in our previous work with the grey level co-occurrence matrix (GLCM)[9]. From empirical studies, we found that GLCM performs better in multi-class classifications than GLGL. GLCM can be considered an approach for capturing the inter-pixel relationships. More precisely, the elements in a GLCM matrix represent the relative frequencies of two pixels (with grey level g1 and g2, respectively) separated by a distance, d. GLCM can be defined as follows:

$$\begin{aligned} o(g_1, g_2, d|\theta) = \# \{(x,y) \mid p(x,y) = g_1; \\ p(x+u, y+v) = g_2; \\ u = d\cos(\theta) \quad \& \quad v = d\sin(\theta); \\ 0 \le g_1, g_2 \le G-1 \quad \& \quad 1 \le d \le D \quad \& \quad 0^\circ \le \theta \le 180^\circ\}, \end{aligned}$$

Cover Image Estimation

Cover image estimation is the process of eliminating embedding artefacts1 in a given image with the objective of getting close to a "clean image". Cover image estimation was first proposed by Fridrich and known as image. For brevity, consider the following proposition:

Let Ic and Is represent the cover image and stego image, respectively.

$$\sum |I_c - I_c'| < \sum |I_s - I_s'|, \ then$$

$$\phi(I_c) - \phi(I_c') < \phi(I_s) - \phi(I_s'),$$

where Ic and I′s are the estimated cover images from Ic and Is, respectively. I − I′ is the pixel-wise difference between two same resolution images. | · | represents absolute value and φ() indicates the feature extraction function.

Hidden Message Length Estimation

The field of information hiding has two facets. The first relates to the design of efficient and secure data hiding and embedding methods. The second facet, steganalysis, attempts to discover hidden data in a medium. Under ideal circumstances, an adversary who applies steganalysis wishes to extract the full hidden information. This task, however, may be very difficult or even impossible to achieve. Thus, the adversary may start steganalysis with more realistic and modest goals. These could be restricted to finding the length of hidden messages, identification of places where bits of hidden information have been embedded, estimation of the stegokey and classification of the embedding algorithms. Achieving some of these goals enables the adversary to improve the steganalysis, making it more effective and appropriate for the steganographic method used.

Boundary Pixel Steganography

The steganography developed in [69] is a variant of boundary pixel steganography. This method uses a binary image as the medium for secret message bits. A set of rules is proposed to determine the data carrying eligibility of the boundary pixels. This plays an important role in ensuring that embedding produces minimum distortion and obtaining error free message extraction. In addition, the embedding algorithm generates no isolated pixels. This method also employs a PRNG to produce a random selection path for embedding. As the embedding algorithm modifies only boundary pixels, the visual distortions are minimal and there is no pepper-and-salt like noise. However, if we take a close look at an image with an embedded message, we can observe small pixel-wide notches and protrusions near the boundary pixels[15]. We use these small distortions to launch an attack on the steganographic algorithm. In our attack, we first detect the existence of a hidden message and then estimate its length.

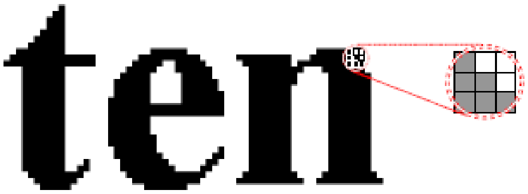


Fig 2: Illustration of a boundary pixel in the magnified view on some portion of the 'n' character

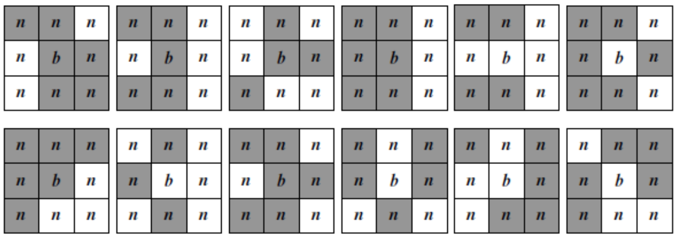


Fig 3: Examples of the patterns formed by a single boundary pixel (denoted by b) and its eight neighbouring pixels (denoted by n) from a binary image.

Improving JPEG Image Steganalysis

To do this we propose to minimize the image-to-image variations, which increases the discriminative ability of a feature set. We will illustrate the efficiency of the proposed method by incorporating it into several existing JPEG image steganalysis techniques. The experimental results presented will verify the feasibility and applicability of the proposed technique for improving existing techniques.

Steganography as Additive Noise

Let X denote an instance of a JPEG cover image and let PC(x) denote the probability mass function of a cover image. In a JPEG image, the probability mass function can be considered as the frequency count of the quantised DCT coefficients

The secret message probability mass function can be defined as the distribution of additive stego noise, which is defined as follows:

$$P_N(n) \equiv P(x' - x = n),$$

where x and x′ are quantised DCT coefficients before and after embedding, respectively.

Image-to-Image Variation Minimization

Defining a discriminative feature set in image steganalysis is a challenging task because the defined feature set should be optimally sensitive to steganographic alteration and not to image-to-image variations. Image-to-image variation is defined as the difference between the underlying statistic of one image and that of another. The underlying statistic can be the histogram distribution of the DCT coefficients or the pixel intensities. For example, the images shown in Figure 4 are obviously different and, therefore, their underlying statistics (histogram distributions shown below each image) differ. This difference is the image-to-image variation. In other words, the image-to-image variation is caused by the difference of the image content.

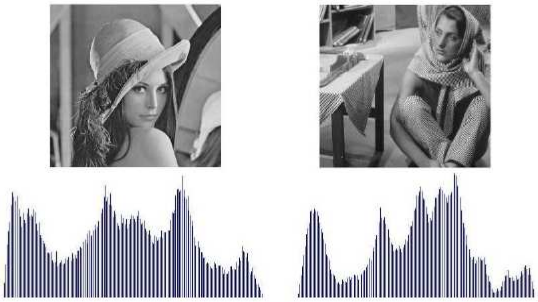


Fig 4: Two images with their respective underlying statistics

If we apply feature extraction directly to the histogram distribution, then the extracted feature will have poor discriminative capability because the image-to image variation is large.

Experimental Results

The experimental settings are described below:

- The embedding algorithm used to create the stego images is the steganography.
- The total embeddable pixel per image produced by this embedding algorithm is about 25 per cent of the   total boundary pixels.
- The maximum message length (100 per cent length) is defined as the total number of embeddable pixels per image.
- Eight sets of stego images (i.e., 10, 20, 30, 40, 50, 60, 70 and 80 per cent) are created from 659 binary cover images.
- The cover images are all textual documents with a white background and black foreground.
- The resolution of all binary images is 200 dpi and with image size of 800×800.
- The prototype is implemented in Mat lab.

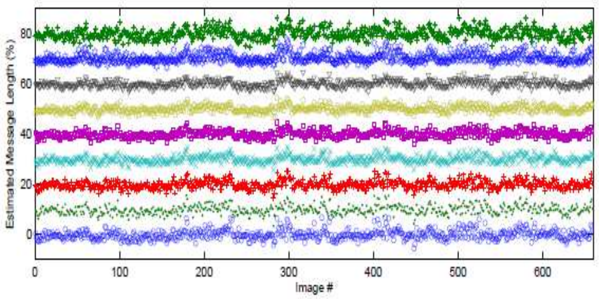
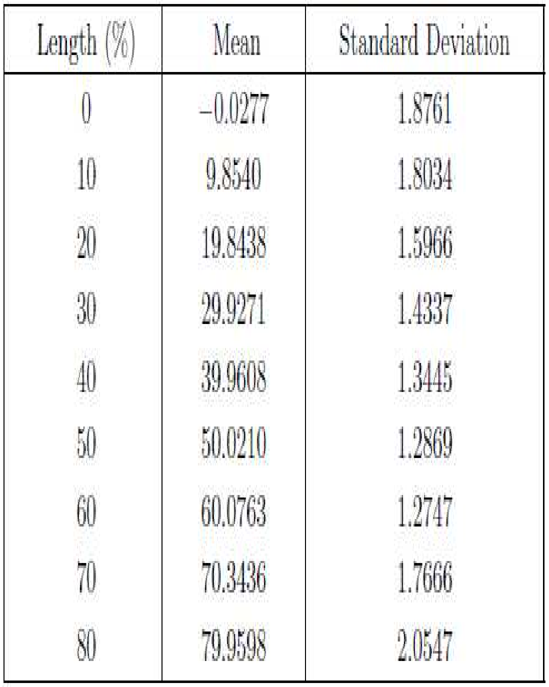


Fig 5: Estimated length of hidden messages for all binary images.

Results of the Estimation

From the 5931 mixture of cover and stego images, we estimate the length of the embedded message and compare it with the actual embedded lengths of 0, 10 , 20, 30, 40, 50, 60, 70 and 80 per cent, using our proposed method. Zero per cent represents a cover image. The estimation results are shown in Figure 6.6. The estimated lengths are very close to the actual lengths. The estimations for large embedded messages, such as 80 per cent are not as close as those of other estimations, although they retain good accuracy. At such a high percentage, some stego images are quite distorted and the pixels exhibit a high degree of randomness[18]. We believe this randomness causes slight instability of our proposed method; however, this phenomenon does not pose a serious problem because we can easily spot the embedding artefacts in such a highly distorted stego image shows a highly distorted stego imag lengths according to the actual embedded message lengths. The average value for each estimated length is very close to the actual length. The standard deviation is also very small—only about one or two per cent. This implies that the estimated lengths do not deviate much from the actual lengths.

TABLE 1:Mean And Standard Deviation Of The Estimation



Length (%)	Mean	Standard Deviation
0	−0.0277	1.8761
10	9.8540	1.8034
20	19.8438	1.5966
30	29.9271	1.4337
40	39.9608	1.3445
50	50.0210	1.2869
60	60.0763	1.2747
70	70.3436	1.7666
80	79.9598	2.0547

The estimation errors are displayed in Figure 6.The estimation error for each binary image is computed as the difference between the estimated and the actual embedded message length in percentage terms. The estimation errors are relatively low and concentrated around 0.00 per cent. The highest estimation error is occasionally found and only about 6.00 per cent, except that one outlier has an error of 7.43 per cent.

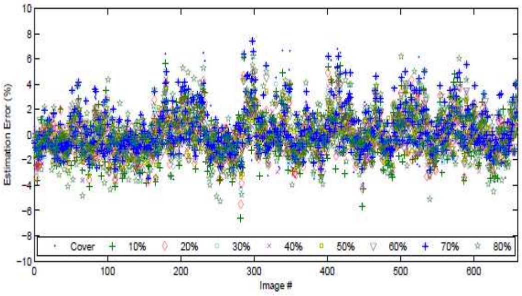


Fig 6: Estimation error of hidden message length for all binary images.

Conclusion

In this paper, we investigated steganalysis that extract information related to a secret message hidden in multimedia document. In particular, we focused our analysis on steganographic methods that use binary images as the medium for a secret message. We organised our work according to the amount of information extracted about the hidden message In conclusion, our proposed technique has improved the selected steganalysis techniques by minimising image-to-image variations. To minimise the image-to image variation, we estimate the cover image from the stego image and then compute the difference between the two. Finally, we extract the feature set from this difference. The experimental results prove the effectiveness of using the proposed technique.The method prposed in this work can detect the steganography developed in and estimate the ength of the embedded message. We observe that it is insufficient only using a set of rules to ensure suitable datacarrying pixels because the notches and protrusions produced from embedding still can be utilised to mount an attack. To alleviate this shortcoming in the steganography, we suggest incorporating an adaptive pixel selection mechanism for the identification of suitable data-carrying pixels. we revisited some of the existing blind steganalysis techniques for analysing JPEG images. We combined several types of features and applied a feature selection technique for the analysis, which not only improves the detection accuracy, but also reduces the computational resources. We showed that an enhancement can be obtained by minimising the influence of image content. In other words, we increased the feature sensitivity with respect to the differences caused by steganographic artefacts, rather than the image content.

References

[1] J. Fridrich, "Feature-Based Steganalysis for JPEG Images and its Implications for Future Design of Steganographic Schemes", *Information Hiding, 6th International Workshop*, *LNCS 3200*, PP 67-81, 2004.

[2] Tomas Pevny, Jessica Fridrich, "Multi-Class Detector of Current Steganographic Methods for JPEG Format", IEEE, 2008.

[3] Tomas Pevny, Jessica Fridrich, "Merging Markov and DCT Features for Multi-Class JPEG Steganalysis", SPIE, 2006

[4] Dongdong Fu, Yun Q. Shi, DekunZou, GuorongXuan, “JPEG Steganalysis Using Empirical Transition Matrix in Block DCT Domain”, Department of Electrical and Computer Engineering, New Jersey Institute of Technology, 2005.

[5] Xiang-Yang Luo, Dao-Shun Wang, Ping Wang, Fen-Lin Liu, “A review on blind detection for image steganography”, Signal Processing, 2008.03.016, 2008.

[6] S. Lyu and H. Farid, "Detecting Hidden Messages Using Higher-Order Statistics and

Support Vector Machines" Proc. 5th Inrernarional Workshop on Informarion Hiding, 2002.

[7] TarasHolotyak, Jessica Fridrich, SviatoslavVoloshynovskiy, "Blind Statistical Steganalysis of Additive Steganography Using Wavelet Higher Order Statistics", CMS2005.

[8] J. Fridrich, M. Goljan, D. Hogea, "Steganalysis of JPEG Images: Breaking the F5 algorithm", 1 Department of Electrical and Computer Engineering, SUNY Binghamton, Binghamton, NY 13902-6000, USA, 2003.

[9] R. J. Anderson. Stretching the Limits of Steganography. *1st International Workshop on Information Hiding*, 1174:39–48, 1996.

[10] R. J. Anderson and F. A. P. Petitcolas. On the limits of steganography. *IEEE Journal of Selected Areas in Communications*, 16(4):474–481, 1998.

[11] I. Avcibas, M. Nasir, and B. Sankur. Steganalysis based on image quality metrics. *IEEE 4th Workshop on Multimedia Signal Processing*, pages 517– 522, 2001.

[12] S. Badura and S. Rymaszewski. Transform domain steganography in DVD video and audio content. *IEEE International Workshop on Imaging Systems and Techniques*, pages 1–5, 2007.

[13] J. D. Ballard, J. G. Hornik, and D. Mckenzie. Technological Facilitation of Terrorism: Definitional, Legal, and Policy Issues. *American Behavioral Scientist*, 45(6):989–1016, 2002.

[14] R. B¨ohme and A. Westfeld. Breaking cauchy model-based JPEG steganography with first order statistics. *9th European Symposium on Research Computer Security*, 3193:125–140, 2004.

[15] C. Cachin. An Information-Theoretic Model for Stegangraphy. *2nd International*

[16] *Workshop on Information Hiding*, 1525:306–318, 1998.

[17] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[18] C.-C. Chang, C.-S. Tseng, and C.-C. Lin. Hiding data in binary images. *1st International Conference on Information Security Practice and Experience*, 3439:338–349, 2005.

[19] S. Chatterjee and A. S. Hadi. *Regression Analysis by Example*. John Wiley and Sons, 4th edition, 2006.