Performance Analysis of various feature sets in Calibrated Blind Steganalysis

Deepa D.Shankar[†] **and Prabhat Upadhyay**[††]

Birla Institute of Technology, Ras Al Khaimah, UAE

Summary
The objective of the science of steganalysis is to detect the message hidden in an image. The steganalyst needs to keep his ultimate goal to detect the image which has data hidden in it and retrieve the message. This paper makes use of statistical data analysis using different feature sets combination at random and a comparative study of each. Since blind steganalysis techniques are used, the technique of calibration to retrieve an estimate of the cover image is used. The steganalytic technique is feature based as well. This is due to the fact that the features that is sensitive to the embedding changes that are employed for steganalysis. The domain used in this paper will be Discrete Cosine Transform. The feature set will be a combination of first order, second order and Markovian set of features. The performance rate is calculated by the error detection percentage of the combination of the feature sets. The extracted features are fed into a classifier which helps to distinguish between a stego and cover image. Support Vector Machine (SVM) is used as a classifier here.Principal Component Analysis (PCA) is used for feature reduction .

Key words:
Steganalysis, feature set, calibration, Markov, DCT, SVM,PCA

1. Introduction

Steganography is a mode of hidden communication. When steganography is utilized to hide a message, anyone inspecting the image will be unable to detect the presence of a hidden message. The message can be hidden within an image, audio or video. Thus the presence of message is invisible to the outside world.

To launch an attack on steganographic schemes, it is necessary to show that the probability of hidden data is more than just random guessing. In this paper, the statistical steganalysis is applied to different steganographic schemes. The paper refers only to embedding data into images. The image format used in this paper is restricted to JPEG. This decision is taken based on the fact that JPEG is the common image format used over the internet. Steganalysis are broadly divided into two – Targeted and Blind. Targeted steganalysis is designed for a particular steganographic scheme. Hence Targeted Steganalysis is robust for that particular steganographic scheme. This is because the detection accuracy for the hidden message is better. Blind steganalysis does not depend on any particular steganographic schemes. They are independent in their behavior with different embedding techniques. To achieve this, a set of distinguishing statistics that are sensitive to embedding techniques, are determined and collected. These statistics are retrieved from stego and calibrated stego image which would act as an estimate of the cover image. The extracted statistics can be used to train a classifier, which can subsequently be used to decide whether the given image is cover or it has a message embedded in it. Since the dependency on the embedder is removed, blind steganalysis can be used for analysis of an image and the classifier can be used to clearly distinguish between a cover and a stego. The extracted statistics are combined and the output error is checked to decide on the performance of the feature set. This combination is done to decide which set of features can best be used to detect a hidden message.

Blind steganalysis, otherwise known as Universal steganalysis, is composed of three parts. They are feature selection, feature extraction and feature classification. In feature selection a set of features are selected from different statistical features. The features used in this page are first order features, second order features and Markovian features. These features are mainly constructed in DCT domain. These features are proposed by observation on general image features which exhibit potential variation when embedding is done on it. The features are extracted from both stego and cover images. In blind steganography we do not have cover image. Hence an estimate of cover image is created by means of a technique called calibration [5]. This technique is used for feature extraction.

Calibration is done as discussed below.

The stego image Si is decompressed and restored in spatial domain. Si is cropped by 4 pixels horizontally and vertically, thus removing the embedding [ ]. Si is recompressed back to DCT domain using the same quantization factor to create Ci. Any functional in the feature set is obtained from the L1 norm of the difference of the fundamentals.

$$F_{final} = |f_{(Si)} - f_{(Ci)}|$$

Manuscript received August 5, 2016
Manuscript revised August 20, 2016

Fig.1.Technique of calibration

Thus the features are extracted. The third component is feature classification. The classification is done in two phases. The classifier needs to be trained first using the image dataset. Both cover image and stego image statistics are used for this purpose. The trained classifier is then used to classify an input image as either a clean image or the one with the hidden message embedded in them.

Thus the feature based steganalysis is done by using the features that is sensitive to embedding changes and insensitive to image content. Previous literature [1, 2] had either the DCT features or Markovian features extracted for analysis. In this paper, we make use of both DCT features and Markovian features. This is to eliminate the drawbacks of both.

The features are extracted from calibrated images. The features are normalized to decrease the computational complexity. Complexity is further reduced by using Linear Support Vector Machine as a classifier. By merging the major first order, extended DCT features and Markovian features, [3] a total of 274 functional features are recovered which are distinguishable.

In section 2,the general architecture of the system is discussed. The experimental results will be discussed in section 3.Section 4 gives an overview of the classifier. Section 5 gives a short note on the future work.

2. Implementation

The system architecture of the feature based steganalytical system using DCT is as given below. This system deals with the normalized features being classified as cover or stego using a linear classifier. The features are calculated from DCT domain for a more straight forward interpretation of the changes in embedding.

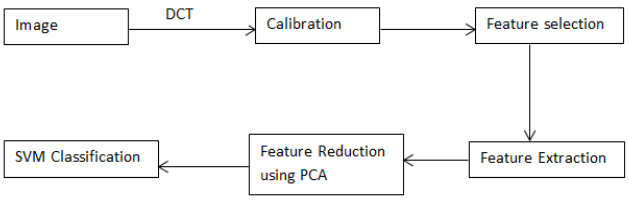

Fig.2 . System Architecture

2.1 Feature Extraction

The goal of this paper is to formulate a new feature set by merging two standard feature sets. These feature sets are obtained from calibrated image, thus giving a good detection rate and a better classification rate. The features are initially extracted in the DCT domain. These features are normalized to eliminate redundancy and hence computational complexity. The classifier used here is linear support vector machine. The choice of the classifier was based on its reliability, accuracy, cost and efficiency. The SVM is trained using the major share of the extracted features. After the training the remaining dataset is used for testing. The system architecture for the representation is as in fig.2.

The images are converted to its DCT domain, divided to 8 X 8 pixel blocks and the features are extracted. DCT is used since JPEG can undergo lossy compression and the signal information is mostly concentrated in the low frequency component of DCT. The extracted features are first order, second order, extended DCT features [1] and Markovian features [2]. The basic DCT features consist of 23 functionals. The extended features have 193 functional in them. The original Markovian features have 324 functionals which dimensionally is huge. Hence the feature dimensionality is reduced to 81. Since the pixels are divided to 8 X 8 pixel blocks, the inter block dependencies between DCT coefficients and intra block dependencies are captured which is projected as DCT features and Markovian features respectively. The merging also help to classify features which otherwise would give a complementary performance. The quantization factor used here is 50 and all images are calibrated.

Calibration is a process by which one can estimate the macroscopic properties of the cover image from the given stego image. Calibration is usually done in DCT domain to get the estimate of the cover image. The resultant calibrated image will be perceptibly similar to the original image.

The features to be extracted are dependent on blocks [3]. Hence, the transformed image is divided into 8 X 8 blocks. Thus, four types of statistical features are extracted. They are the first order DCT, extended DCT and Markovian functions. The first order DCT is global histogram and dual histogram. The second order features are variance, blockiness and co-occurrence. Assume that a stego image is represented by DCT coefficient array dk(i,j) [1],where i and j are coefficients and k is the block [1].

The global histogram of 64k blocks can be represented by Gr where r=A,…,B where A = min k,i,j (dk(i,j)),B = max k,i,j (dk(i,j))

The dual histogram gives an idea on how a coefficient is distributed among different DCT modes. It can be represented by

$$g_{ij}^{d} = \sum_{k=1}^{B} x\left(d, d_{k(i,j)}\right)$$

where B is the number of blocks

Variance that captures the inter block dependencies [1] can be represented by

$$V = \frac{\sum_{i,j=1}^{8} \sum_{k=1}^{|I_r|-1} \left|\boldsymbol{d}_{Ir(k)(ij)} - \boldsymbol{d}_{Ir(k+1)(ij)}\right| + \sum_{i,j=1}^{8} \sum_{k=1}^{|I_c|-1} \left|\boldsymbol{d}_{Ic(k)(ij)} - \boldsymbol{d}_{Ic(k+1)(ij)}\right|}{|\boldsymbol{I_r}| + |\boldsymbol{I_c}|}$$

Where Ir and Ic are vectors of block indices when scanned by rows and columns.

The blockiness is also an inter-block dependency functional which is calculated in decompressed DCT images. It is represented by

$$B_{\alpha} = \frac{\sum_{i=1}^{\left|{}^{(M-1)}/_{8}\right|} \sum_{j=1}^{N} \left|x_{8i,j} - x_{8i+1,j}\right|^{n} + \sum_{i=1}^{\left|{}^{(N-1)}/_{8}\right|} \sum_{j=1}^{M} \left|x_{8i,j} - x_{8i+1,j}\right|^{n}}{N\left|{}^{(M-1)}/_{8}\right| + M\left|{}^{(N-1)}/_{8}\right|}$$

Where M and N are dimensions of the image

The co-occurrence matrix which determines the probability distribution of pairs of neighbouring DCT coefficients is represented as below.

$$C_{st} = \frac{\sum_{k=1}^{|I_r|-1} \sum_{i,j=1}^{8} \delta(s, d_{i,(k)}(i,j))\delta(t, d_{i,(k+1)}(i,j) + \sum_{k=1}^{|I_c|-1} \delta(s, d_{i,(k)}(i,j))\delta(t, d_{i,(k+1)}(i,j)}{|I_r| + |I_c|}$$

The original DCT features have 23 functional in them. These features can be extended to form extended DCT features, which contains 193 functional to be extracted. The final set of statistical features is the Markov features.

Markov features have the dimensionality of 324. The dimensionality is reduced to 81, to remove the redundant features. The extended DCT features capture the inter-block dependencies among DCT coefficients whereas the Markov features capture the intra block dependencies (of similar spatial frequency with in the same 8 X 8 blocks) between frequencies between the consecutive blocks [2]. The merging of the two features would help to eliminate the drawback for both. Another reason for merging the set is that the classifiers that are employed for each feature set individually has complimentary performance. More over the Markov features are unable to detect embedded messages that are of short length. The features are extracted from calibrated images.

Calibration of images is done to get an estimate of the cover image. Calibration is usually done in the DCT domain. Calibration is the process or technique by which a JPEG image J1 is decompressed, converted to spatial domain. After conversion the image is cut horizontally and vertically by 4 pixels each [5]. The image is later compressed back to the DCT domain using the same quantization matrix used earlier. This technique is equivalent to shifting of an image, thus helping it to retain the DCT coefficients. But the change erases the data embedded in the image, thus creating a close estimate of the cover image. The newly created calibrated JPEG image J2 will have the macroscopic features that are similar to the original cover image since the cropped image is visually similar to the original image. The DCT features are extracted as the L1 norm of the absolute value of the difference between the cover image and the stego image. The Markov feature set [2] models the difference between the absolute values of the neighboring DCT coefficients used as a Markov process. The original Markovian functional consists of 324 features. These features will have increased dimensionality, which needs to be reduced. This is done by taking the average of the four 81 dimensionality features which are measured

Thus the 193 functional of extended DCT features along with 81 dimensionality of the reduced Markov features combined together to form the needed feature set of 274 functionals.

2.2. Feature Classification

Once the features are extracted, it needs to be classified. The extracted features determine whether the image is cover or stego. Hence, this paper uses the technique of machine learning and the classifier used here is Support Vector Machine (SVM).

2.2.1 SVM Prediction

In the field of machine learning, many techniques or algorithms are used to classify a set of data. Some techniques that are involved are Neural Network, Perception, Fisher Linear discriminant, Support Vector Machines etc. SVM is a classifier that is defined by a separating hyper plane [4]. It is a supervised learning technique. Hence, after training the SVM, a hyper plane is output which classifies the test data. The hyper plane should be so created that the distance of the plane should be maximum with the training samples. If the hyper plane is close to the sampling data, it will be sensitive to noise, and the classification will not happen correctly.

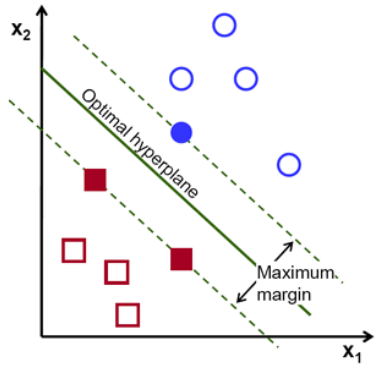

Fig 3. SVM Classification

The hyper plane can be computed using the formula

$$F(x) = w_0 + w^T x$$

Where w is the weight vector and w0 is the bias. X is the training sample. For an optimal hyper plane,

$$| w_0+w^Tx| = 1$$

Hence the distance between a point x and a hyper plane (w, w0) is

$$\frac{| w_0+w^Tx|}{||w||}$$

SVM is widely popular among the classification algorithm since it maps the features from the original space to a high dimensional feature space. If the kernel function is linear, the resulting SVM classification is a maximum margin hyper plane. Given a training sample, the hyper plane would split it in a way that the distance from the closest values to the hyper plane is maximum. The complexity of SVM depends on the features used for training it. Thus the SVM can guarantee a generalization to a great extent.

2.2.2. Principal Component Analysis

Principal Component Analysis (PCA) is a technique of feature reduction from a high dimension feature space. Thus Principal Component Analysis can help eliminate redundant features and highlight the important features. The Principal component analysis uses standard deviation, covariance, eigen values and eigen vectors to evaluate the set of relevant features. Hence PCA can be termed as a simple and robust was of feature reduction without much loss in overall data. This is done by transforming a set of principal components which has most of characteristics of all the original features.

PCA helps to pre-process data so that computational cost in a linear classifier will be less since the dimensionality has been reduced [10].

The original data will have its mean calculated across both dimensions. The covariance matrix is then calculated. The covariance matrix for a data with n dimensions is

$$C^{nXn} = (c_{i,j}, c_{i,j} = cov(Dim_i, Dim_j))$$

CnXn is a matrix with n rows and n columns,Dimx is the xth dimension [10].

After the calculation of covariance matrix, eigen values and eigen vectors of covariance are calculated. By calculating the eigen vectors of the covariance, the plotting lines can be drawn which represent the data. The eigen vector of the highest eigen value is the Principal Component. Once the eigen value is retrieved, it can be arranged in the order of highest to the lowest. The lowest significant values can be ignored. This is the final dataset , which can be called as feature vector. The final data set will be the transpose of the feature vector , which is multiplied with the transpose of the mean adjusted original data.

3. Experimentation results

3.1 Database of images

An important aspect of any performance evaluation research is the quality of the database that is employed for the experiments. The dataset contain different image formats, quality and texture. A set of 420 JPEG images are compressed to the size of 256 X 256 JPEG format retrieves the RGB representation of an image [4]. It would then calculate the quantization table corresponding to the quality factor Q and thus compresses the image while storing the quantized DCT coefficients [4]. The estimated capacity is then computed. A password which is user specified is used to generate a seed for PRNG that creates randomness for the embedding message. The message is divided into K bits and embedded in 2 K-1 coefficients along the random walk. If the message fits onto the estimated capacity, the embedding would proceed, else would show an error message.

The results derived are based on sample from the testing dataset. These datasets are not used in any form during the training of the classifier.F5 and pixel value differencing (PVD) are used as steganographic schemes[7][8]. A dataset of 420 images of cover and stego is used for analysis. These images are used to extract the different features. Many features of the datasets are irrelevant. Hence these features maybe removed for better performance. The extracted features after removal of irrelevant ones are then fed into a linear classifier. Linear SVM is used for this purpose. From the 420 images selected, about 340 images are used to train the SVM. The remaining 80 datasets are used for testing. The main features considered for the analysis are first order features, extended DCT features and Markovian features. Combination of each feature is taken, compared and tabulated below.

Table 1:Table of comparison of the performance rate of combination of feature sets of calibrated JPEG images using SVM

Sl No	Combination of features	Error Detection
1	Markovian , Dual Histogram	0.1
2	Blockiness, Cooccurance,Global Histogram	0.095
3	Variance,Moment,Dual Histogram	0.13
4	Blockiness,Moment,Dual Histogram	0.123
5	Variance, Moment ,Co-occurrence	0.098
6	Variance,Moment,Dual Histogram ,Global Histogram	0.09
7	Blockiness, Moment, Cooccurance,Global H histogram	0.112

8	Variance, Moment ,Co-occurrence ,Dual Histogram	0.13
9	Markovian, Moment ,Co-occurrence ,Dual Histogram	0.099
10	Blockiness , Variance, Moment , Dual Histogram	0.043
11	Markovian , Variance, Moment , Dual Histogram	0.04
12	Blockiness, Variance, Moment, Co-occurrence	0.0522
13	Markovian, Blockiness,Variance,Cooccurance	0.12
14	Blockiness, Variance,Moment,Dual Histogram, Global Histogram	0.058
15	Blockiness, Variance,Moment,Dual Histogram, Global Histogram	0.098
16	Blockiness, Variance, Moment, Cooccurance,Global Histogram	0.178
17	Blockiness, Variance, Moment, Cooccurance,Dual Histogram	0.048
18	Markovian, Blockiness, Variance, Moment , Dual Histogram	0.189
19	Blockiness, Variance, Moment ,Co-occurrence, Dual Histogram, Global Histogram	0.154

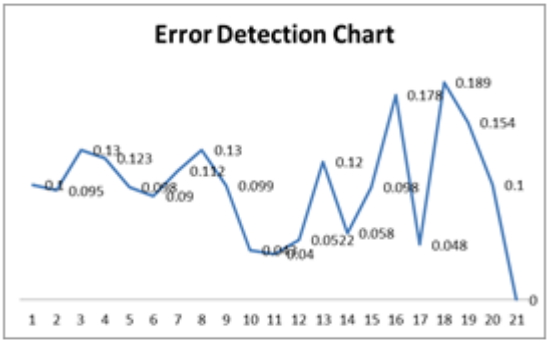

Fig 4:Graphical representation of error detection

From the tables it can be concluded that combination of certain features would give a lesser error detection than the other. The global histogram used is normalized to 11 features, from - 5 to 5. For dual histogram of 99 functionalities the upper 9 triangular values are taken. For the co-occurrence matrix of 121 features, the co-index of s and t values are normalized to values from -2 to 2.

Sample outputs of global histogram, dual histogram and co-occurrence are given below. The plot is based on the range of specified normalized values against the occurrence of pixels.

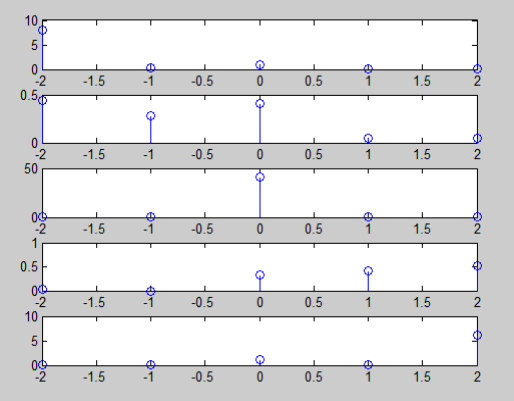

Fig 5 :sample plot on cooccurance

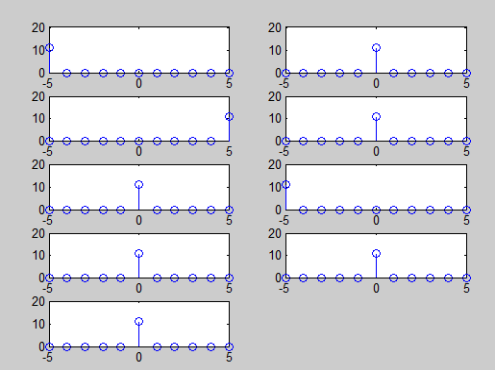

Fig 6: sample plot on dual histogram

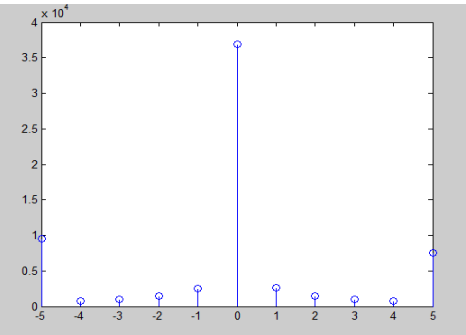

Fig 7: sample plot on global histogram

4. Linear Support Vector Machine

Even though support vector machines are effective in high dimension spaces, the dimensionality of the feature set needs to be reduced since the classifier is likely to give poor performance if the number of features is much greater than the number of samples. Moreover the feature reduction could reduce the cost and computational complexity. Since the system of steganalysis is Blind, and the learning system used is supervisory, SVM needs to be

trained before any test data is input. Out of the 420 JPEG images used, 340 images of different textures are used to train the SVM [3]. The other 80 is used to test the data. The stego data used to train and test the SVM is created by embedding data using different steganographic schemes and the percentage of embedding ranges from 10%,25%,50%,and75%.The quality factor used to transform the image also varies from 50 to 70.

5. Conclusion

A set of images were converted from spatial to a transform using different quality factors and is used to extract a set of relevant features, is developed. The relevancy of the features was based on how they change the DCT coefficient when an embedding is done on the image. The feature set is obtained by merging two separate features which had been proposed individually, and that are complementary to each other. The DCT features are obtained by replacing the L1 norm of their differences. The Markov features were extracted and their dimensionality was reduced for effectiveness. The features are randomly combined, and the results are obtained. The combination of the merged feature set provides a better result than the combination of individual features.

The feature set extracted, is being input into the Support Vector Machine, to classify whether a new input image is a cover or stego. An enhancement of the research can be done using the feature set extracted from uncalibrated images. Feature selection can be done by means of Principal Component Analysis [9]. A comparative study of calibrated images and uncalibrated images can be done in this case.

Another enhancement can be the determination of the message length. This can be done in the spatial domain. In the digital processing perspective, the embedding can be considered as an addition of noise. The data to be embedded is kept low in order to fulfill the detectability factor. This method work in three steps. First, the estimate of cover image is recovered. The stego image is then recovered from the image, and its length is estimated. Since prior knowledge of the length of the message is not known, the Maximum Likelihood algorithm can be used for the estimation of message length.

References

[1] J.Fridrich “Feature Based Steganalysis for JPEG Images and its Implications for Future Design of Steganographic Schemes”.6th International Workshop, volume 3200 of lecture notes in Computer Science, Pages 67-81, 2005

[2] T.Pevn’y and J.Fridrich. “Merging Markov and DCT Features for Multiclass JPEG Steganalysis” Proceedings SPIE, Electronic Imaging, Security, Steganography and Watermarking of Multimedia Contents IX,Volume 6505, Pages 301-314, San Jose,CA,January 29-February 1, 2007

[3] Deepa D.Shankar,T.Gireeshkumar,Hiran V. Nath, “Steganalysis for calibrated and Lower Embedded Uncalibrated Images”, Lecture Notes on Computer Science,Springer,Volume 7077,2011,pp-294-301

[4] Mehdi Kharrazi, Husrev T.Sencar, Nasir Memon “Performance study of common image steganography and steganalysis techniques”Journal of Electronic Imaging15 (4), (2006)

[5] Jan Kodovský, Jessica Fridrich,"Calibration Revisited”, Proceedings of the 11th ACM workshop on Multimedia and security pp.63-73(2009)

[6] Tomas Pevny Jessica Fridrich, Andrew D.Ker “From Blind to Quantitative Steganalysis”, SPIE, Electronic Imaging, Media Forensics and Security XI, San Jose, CA, January 18-22, 2009, pp. 0C 1-0C 14.

[7] Da-Chun Wu and Wen-Hsiang Tsai “A steganographic method for images by pixel-value differencing” Pattern Recognition Letters, Vol.24, pp.1613-1624, (2003).

[8]] J.Fridrich, M.Goljan and D.Hogea, “Steganalysis of JPEG images: breaking the F5 algorithm” in Proc. 5th International Workshop on Information Hiding (IH’02), pp 310-323, Noordwijkerhout, Netherlands, October (2002)

[9] Miranda, A.A., Le Borgne, Y.A., Bontempi:”New Routes from Minimal Approximation Error to Principal Components.” Neural Processing Letters 27(3), 2008

[10] Lindsay Smith,” A tutorial on Principal Component Analysis”, 2002.

Deepa D.Shankar has completed her Masters in Cyber Security from Amrita Viswa Vidyapeetham. Currently she is the Assistant Professor of Computer Science within the Department of Computer Science at Birla Institute of Technology offshore Campus, Ras al Khaimah. She has authored 5 papers in the field of statistical steganalysis. Her work focuses on feature based steganalytic scheme.

Prabhat Upadhyay Received his U.G and P.G degrees from Ranchi University, India. He obtained his Ph. D. (Technology) degree from Birla Institute of Technology, Ranchi; India.He has been working as a faculty in EEE Department of BIT Mesra since last 15 years. He has published more than twenty research papers in various Journals and conferences. His current research interests include Brain signals, Signal processing and soft computing applications