Impact of features selected by Principal Component Analysis in featured based steganalysis in calibrated and non-calibrated images

[1]Deepa D.Shankar

ABSTRACT--*Steganalysis is helpful in finding the hidden information/data/message in cover images. In simple form, the confidential and concealed message has to be extracted efficiently in steganalysis. This paper performs universal steganalysis based on the features using F5 and Pixel Value Differencing (PVD) algorithms. The feature extraction is carried out through Discrete Cosine Transform (DCT) techniques. The dimensions or size of the feature vector/ feature set is reasonably diminished by Principal Component Analysis (PCA). The extracted features are the combined DCT and Markovian features that have 274 features. These features are inputted to the Linear Support Vector Machine (SVM) for classifying the stego and cover image. Prior to analysis, the images are calibrated so as to improve the efficiency of classifier. The analysis is done with different embedding percentages and the training and testing images are split in the ratio of 80 and 20 for SVM classifier.*

Keywords--*Steganalysis, DCT, feature set, PCA, SVM classifier*.

I. INTRODUCTION

The aim of steganography is to offer furtive data transmission. Steganalysis is a method of identifying the existence of hidden information. Most of the times, the cryptography and steganography are understood synonymously (11). Cryptography modifies the secret information from one form to the other, i.e. the message is jumbled, indecipherable and most importantly, the presence of the information is not known (35). The messages encrypted could be traced and captured but decoding them is very difficult. By this way, the message could be secured but capturing the message would probably provide a hint to the opponents and they know that some communication is going on (2). Steganography is exactly contrasting to this method and it takes all efforts to hide all the proof while communicating. The goal line of steganography is in fact, to attach a message inside a carrier signal in a way that it has not been identified by unwanted receivers (6). Steganalysis finds out the concealed signals in alleged carriers or it identifies the media that possess the hidden signals/information. The paramount concern of steganography is to identify and apply a stronger methodology for identification (12). The process of steganography and the steganalysis (7) is better understood through the image depicted in Figure 1.

[1] *Research Scholar, Banasthali Vidyapith, Rajasthan, India,sudee99@gmail.com*.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

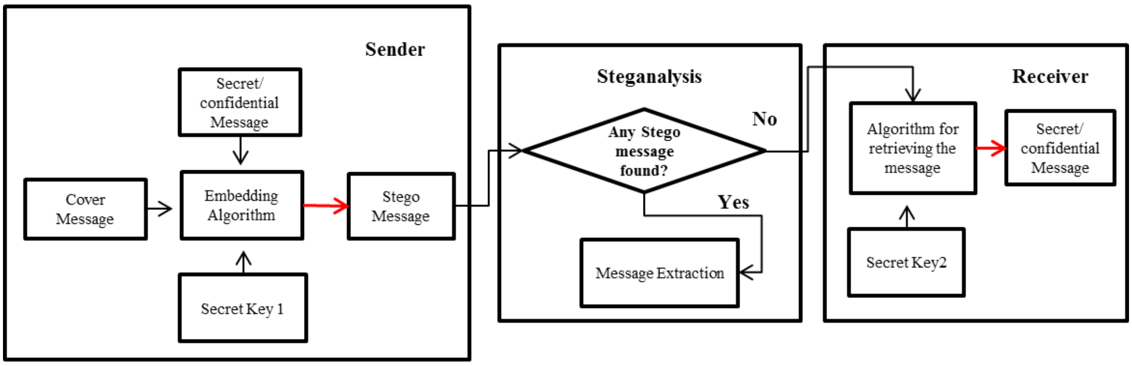

Figure 1*:* Process of Steganography and Steganalysis

II. IMAGE STEGANOGRAPHY

The Image steganography is stated as “the secret implanting of data into digital photographs”. Although steganography pelts data in any one of the digital Medias, digital photographs/images are the most prevalent as “carrier” because of their recurring usage on the web (4). As it has large sized image file, it can cover huge quantity of data. The Human Visual System cannot discriminate the usual image and the image with secreted data.

Further, as there are big numbers of redundant bits in the images of digital format, they are mostly preferred as cover objects (33). This work hence makes use of images as cover file. Various formats of digital format images include PNG, GIF, JPEG, BMP and TIFF files and any of these formats could be utilized for cover objects. The BMP or bitmap is the simplest of all. It is easy to handle data with this format because it is not compressed but it is quite difficult to manage the size with uncompressed images (13). The Graphics Interchange Format (GIF) and the Tagged Image File Format (TIFF) are rather of compressed format. They both use lossless compression (45). The GIF images have color palette to have images with color indexing but it could stock just 256 different colors and is not suited to describe composite photography with incessant tone. Portable Network Graphics (PNG) offers superior support for color with good compression ratio. Anyhow the standard format for image steganography is the Joint Photographic Expert Group (JPEG), which uses lossy compression but still maintains the image quality (26). The types of images usually dealt with are the binary or bi-level images, gray scale images and RGB or true color images. The bi-level images are allocated with bit value either one or zero (47). Each pixel of the image is assigned with a bit. This is represented either with black or white color. The colors are characterized as gray shades and every pixel is described with 8 bits (19). Thus, 256 various shades of gray is possible with these kinds of images. The pixel color is defined through intensities of a group of red, green and blue components in case of true color imageries (22). Each color is with 8 bits and the group hence has 24 bits. This study considers JPEG image for analysis.

The image steganography could be broadly sectioned into spatial domain and transform domain. The taxonomy is depicted in Figure 2 (20). The two principal types of Steganalysis are the Target and Blind analyses. Targeted Steganalysis is intended for a specific algorithm. This type is very strong as it offers better accuracy of detection while Blind steganalysis is not subjected to any individual algorithm. Therefore, it eliminates the drawback of dependency in case of targeted steganalysis that is appended to an explicit algorithm. Moreover, Blind/universal steganalysis could work with the statistical data of a particular image. Hence, it is even referred to as statistical steganalysis (38). The statistics of images are known as features of the images. Some features, which vary through the process of embedding, but do not influence the image are only chosen and extracted then.

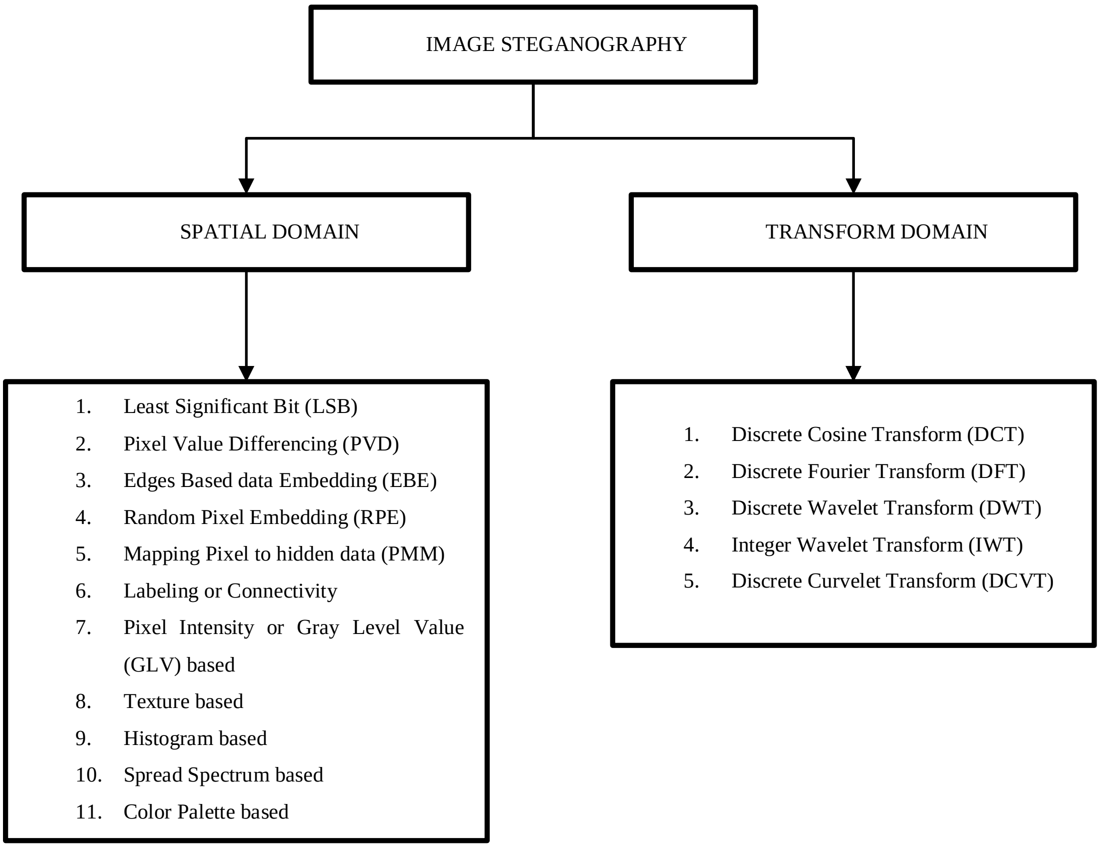
IMAGE STEGANOGRAPHY
SPATIAL DOMAIN
TRANSFORM DOMAIN
1. Least Significant Bit (LSB)
2. Pixel Value Differencing (PVD)
3. Edges Based data Embedding (EBE)
4. Random Pixel Embedding (RPE)
5. Mapping Pixel to hidden data (PMM)
6. Labeling or Connectivity
7. Pixel Intensity or Gray Level Value (GLV) based
8. Texture based
9. Histogram based
10. Spread Spectrum based
11. Color Palette based
1. Discrete Cosine Transform (DCT)
2. Discrete Fourier Transform (DFT)
3. Discrete Wavelet Transform (DWT)
4. Integer Wavelet Transform (IWT)
5. Discrete Curvelet Transform (DCVT)

Figure 2: Taxonomy of Image Steganography

This is accomplished in Transform domain, specifically; the Discrete Cosine Transform (DCT) has been applied in this work for carrying out feature centered blind steganalysis on JPEG images. The images are primarily transformed in case of transform domain, and only after that, the message/information is implanted to it. It performs data concealing by the way of manipulating mathematical functions and transformations of image. Cover image Transformation is achieved by means of coefficients tuning and transformation inverting.

III. PROCESS INVOLVED IN STEGANALYSIS

The basic steps of steganalysis involve the feature selection, feature extraction and feature classification. The features are not mined from all the images from the sender but they are extracted only for suspected images. The mined features are served to the classifier for classifying the stego and cover images. Again features are very much pivotal for correct classification, in the sense, they decide the classification accuracy. During feature extraction stage, there are some irrelevant features, which could adversely affect the efficacy of the analysis. Hence, feature reduction (dimensionality reduction by removing unwanted features) is also added in the process of steganalysis (18). In this work, the feature reduction using principal component analysis is considered. In blind steganalysis, the calibrated images can be used. Calibration could augment the sensitivity of the features towards embedding and shrinks image to image variations of features. It has been familiarized in the year 2002 as a novel idea to attack F5-algorithm. It turns out to be an indispensable fragment of many feature-dependent universal and targeted steganalysis with JPEG images as well as in spatial domain (21). Calibration also helps in handling signal to noise ratio effectively. This work will compare the analysis for both calibrated and non-calibrated images.

IV. RELATED WORK

Principal Component Analysis

One of the prevailing problems in JPEG steganalysis is high level of feature redundancy. In addition, the Complementary features are not properly utilized. This could be reduced by a new approach that applies Principal Component Analysis (PCA) and investigates about the feature complementary (16). This is accomplished by integrating the complementary features and isolated components that are redundant by means of PCA. The classification of this study has made use of RBaggSVM classifier. The numerical results of the study proved that the dimensions of characters, training and testing time have been much reduced in comparison to the existing methods.

Fazli and Zolfaghari-Nejad, (14) have taken gray-scale images for steganalysis. They had analyzed the images based on statistical characteristics. The features extracted for this purpose were the images that are highly capable of differentiation among stego images. The coefficients of Discrete Wavelet Transform with higher order statistics had been used, where the pre-processing was carried out with PCA on the features that are extracted. The classifier used for classification was the SVM. The authors could able to identify the existence of messages that hidden by the suggested approach with an accuracy of 90% with various rate of embedding.

Xuan et al. (46) suggested a class wise non principal component analysis for classifying the steganographical images in the vector space containing features of high dimension, i.e. 360 dimensional sensitive feature vectors, which were sensitive to the process of data embedding and were the derivatives of Markov model of multi direction. Three level sub-bands of wavelet along with first scale slanting sub-band coefficients through decomposition were used to make a Characteristic Function (CF). The moments of CF were utilized for the universal image steganalysis. The authors selected 102 dimensional features using the initial 3 statistical moments from the test image wavelet band as well as prediction error image. Feature reduction was carried out using PCA while the classification was done through SVM (25).

Though the image classification can be done without the acquaintance of steganographic algorithms in universal blind analysis for stego images, the amount of payload to be embedded is an issue that is usual. Here, PCA is applied to enhance the rate of False Positive (FP). The universal stego features are utilized to evaluate the payload by means of support vector regression. The SVM classifier classifies the images quantitatively for the 6 different applied algorithms (5).

DCT Techniques

Blind steganalysis senses the incidence of the messages secreted by means of several sorts of algorithms for steganography and is capable of noticing fresh unfamiliar steganography algorithms. Rabee et al. (36) suggested an authentic method to find the presence of concealed messages in JPEG imageries effectively. The functions related to image processing are straightaway executed in the DCT domain for the purpose of reducing the cost related to memory and the time of computation. The relevant features are mined by analyzing the variations of the coefficients of DCT with and without cropping. The features extracted are supplied to the SVM classifier for labeling stego and clean images.

The JPEG kind of steganographic systems are attacked in an efficient manner by the presented method using Outguess, Jsteg, DWT and F5 algorithms. This further finds the linkage among the coefficients of block-DCT through the relation amongst inter and intra blocks. The features are chosen from the characteristic function's statistical moments. The array of BDCT JPEG has been utilized for mining the features. A cross-validation SV machine has been utilized for classification (39).

A fresh scheme of steganalysis for data secreted in images of JPEG format has been suggested. The features are mined using DCT and wavelet. Then, the features are polished to be capable of discriminating the stego and clean images that have been carried out through a powerful classifier. The Markov features through DCT that is extended are integrated with wavelet sub-bands using the statistical details of SVD. The highly sensitive features in terms of embedding of data are picked by a selection procedure using SVM-RFE (48).

Feature Based

A broad-spectrum steganalysis scheme, which could attack Steganography blindly, spots hidden data regardless of the embedding algorithm and is very much valuable for practical applications. A Steganalysis method has been proposed for identifying the secreted information regardless of the databases of the image and embedding algorithm via Markov features that are modified (34). A blind classifier has been built with the help of Radial Basis Function Neural Network (RBFNN).

A novel method of steganalysis that is blind, has been recommended to find the existence of the messages/information hidden/embedded in the black and white (bi-variable) images using the steganographic tactics. The various matrix sets extracted for this work are the matrix of run length, matrix of gap length and the matrix for the difference in pixels. The distinguishing abilities are boosted through a characteristic function of the matrices extracted. The various statistics (features) such as kurtosis, mean, skewness and variance have been computed. It has been observed through the results that the suggested technique could able to detect even a small percentage of messages implanted (9).

A blind steganalysis is suggested for the JPEG type of images by a dilation procedure. This method deals with color image and therefore the RGB elements of the assumed image has been split and then they are converted into wavelet, frequency and spatial domains. The clean as well as stego images are classified through an efficient classification scheme that uses SVM. The metric for evaluating the performance is the overall success rate. Through this, it has been seen that the recommended procedure has higher rate of detection (30).

Iterative Algorithms

Feature selection is one among the key phase of pre-processing that could impact the output of steganalysis. A new-fangled feature-dependent universal steganalysis method has been projected for noticing stego images from the cover images in JPEG descriptions by means of a feature choosing practice using artificial bee colony (ABC) method. The performance of classifier and the size of the nominated feature vector are reliant on investigative info for ABC. The subsequent subset containing adaptable feature with regard to the minimum dimension feature could be chosen and the classifier's working could be enhanced (28).

Other related Studies

A unique algorithm for image steganalysis that is universal is planned with Region based Image Steganalysis (RISAB) by means of Artificial Bee Colony (ABC) algorithm. The objective of the suggested scheme is to comprehend a sub-image from cover as well as stego images via ABC in terms of density as per the cover, stego and variance images. The superior sub-image encompassing the uppermost density, regarding the altered embedding pixels is found. In addition, to opt for the finest sub-image, the features are mined, which are decided through IFAB. The features selected via ABC and that extracted via RISAB would be integrated to generate a new feature vector with improvised accurateness for steganalysis (29).

A quantifiable universal steganalysis has a drawback of dimensionality and needs many instances (samples of training) to yield outcomes with quality. A universal measureable steganalyser is proposed using spatial transform with LSB method, which involves condensed training samples and features. Designs with opulent mixture of features that are global and local are exploited as essential features. The regression trees from AdaBoost ensemble method is utilized as base learners and are helpful to assess the alteration rate instigated in stego images. A three stage approach of optimization is presented for getting minimum training instances and features with least magnitude (41).

Correlogram characteristics as features of texture have many uses in the area of steganalysis. The most widely used properties of correlogram cover energy, contrast, entropy, correlation and Homogeneity. The influence of these characteristics as metaphors of the contents of image on the universal steganalysis has been examined (3).

A medical application made on iOS platform has been presented for steganography and steganalysis jobs in images of digital formats. This mobile application is well used to perform an analysis for information of provided imageries and finds whether there is any confidential info secreted in the offered images. Also, it could accomplish the task of concealing the data imperceptibly inside the digital metaphors (24).

Problem Statement

This work is intended to perform a feature based blind steganalysis for calibrated and non-calibrated JPEG format images in Transform domain using Discrete Cosine Transform. Further, the dimensionality reduction is carried out with Principal Component Analysis method. The algorithms used for this work are F5-Algorithm and PVD. SVM classifier has been used for classification purpose. The outline of implementation of this study is displayed in Figure 3.

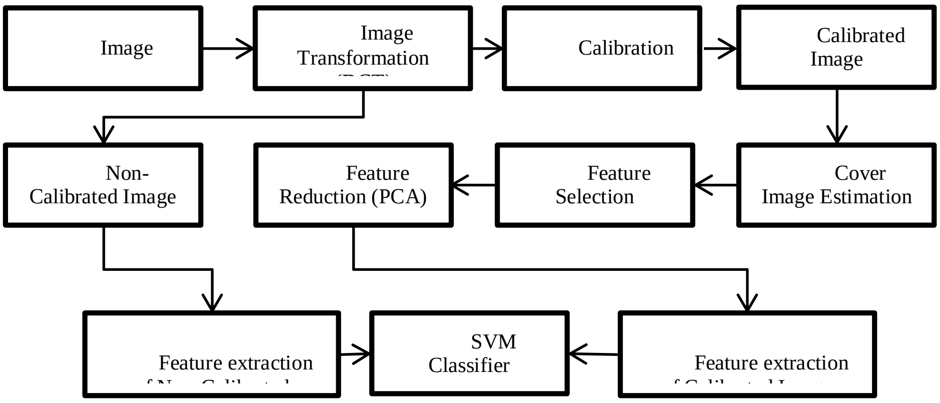

Figure 3: Implementation Flow Diagram of Feature Based Steganalysis

V. METHODOLOGY

Principal Component Analysis

Principal component analysis (PCA) is a calculation procedure that uses mathematical transformations to convert a set of probably interrelated variables into a lesser quantity of non-correlated variables and the resultant reduced variables are termed as principal components(31). It is a tool applied to diminish the dimensions wherein which, huge size variables are changed to smaller sized one yet holds most of the valid information. For better interpretation of PCA, the following two factors may be defined (10).

Variance: “It is simply a measure of variability or it measures how spread the data is”. In mathematical format, it represents the “average squared deviation from the mean score”.

$$\mathrm{var(x)} = \frac{\sum(\mathrm{x_i} - \bar{\mathrm{x}})^2}{\mathrm{N}}$$

Covariance: It is a “measure of the extent to which corresponding elements from two sets of ordered data move in the same direction”. Mathematically,

$$\mathrm{cov(x,y)} = \frac{\sum(\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})}{\mathrm{N}}$$

Here(x, y) are the two variables $x_i$and y_i are the values represented in ith dimension, while $\bar{x}, \bar{y}$ signify their mean values. The covariance may be either positive, negative and zero (43). In positive, the variables are directly related, in the sense, as x increases, y also increases. In contrast, in negative covariance, the variables are related in inverse manner. Increase in one variable decreases the other. Meanwhile, when the variables are not related at all, then it is called as zero covariance.

PCA determines a fresh dimension set in such a way that the entire dimensions are linearly independent (orthogonal) and graded as per the variance of data, i.e. more significant principle axis happens first (27). More significant here refers to more variance. The PCA initially computes the matrix of covariance, X of “n” dimensional data and computes the Eigen vectors based on the Eigen values in a descending manner. Then, it decides on “k” Eigen vectors to have new “k” dimensions. Thus the original “n” dimensional data of larger dimension has been converted to “k” dimensional data of smaller size (23). The covariance matrix can be found using the formula (8),

$C_x = \frac{1}{n-1}(X - \bar{X})(X - \bar{X})^T$; T represents Transpose.

Support Vector Machine

The Support Vector Machine (SVM) is a typical classifier mainly used for supervised learning of machine algorithms. Support vector machines tries to permit a detachable hyper-plane (surface of decision) in a linear manner over a dataset for classifying the data into two different sets (40). This hyper-plane, which is a linear/undeviating separator, could be applied for multiple dimensions. That is, it might be a line (2 Dimensional), plane (3Dimensional), and hyper-plane (multi-dimensional and greater than 3D). The optimal surface of decision that is linear is determined by decreasing the feature vectors of training. These vectors are prearranged in the feature space of larger dimensions (44). The SVM decided the highest margin hyper-plane, which splits feature vectors of the dissimilar classes with margin of high values. The performance of classification is said to be superior when the margin amongst the vectors is high. The data points very close to the hyper-plane are referred to as "support vectors". If these points are detached, it would affect/change the location of the in-between hyper-plane (1). Hence, the support vectors are the vital components of the data set. The image of linear classification by SVM is depicted in Figure 4.

Implementation

It is intended to fuse the new feature set for calibration to obtain an improved rate of detection than the other prevailing methods and to have more accurate classification with calibration. The predominant features are extracted forms the feature vector/ feature set. In fact, this set is helpful in building the linear classifier. The basic step of steganalysis is performed as an initial step by the extraction of features from the calibrated as well as non-calibrated images of JPEG format. The number of features extracted is 297 and they have been regularized for enhancing the efficacy of the algorithm. Also, the dimensionality has been reduced using PCA. The classification is to carried out through liner SVM classifier, for which, it designed to have better outcome in terms of economy, consistency and accurateness. The attained features have been used to train the SVM. The next step is to test the images. These images may be, separate images that are not used for training. Few of the testing images are taken from the training set too.

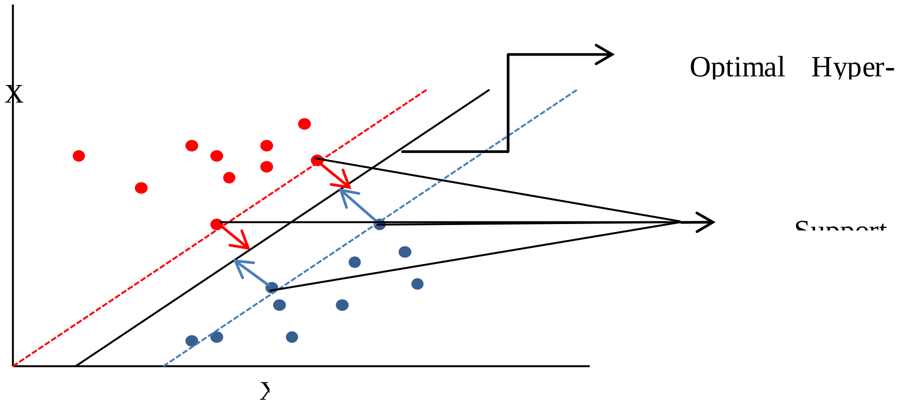

Figure 4: SVM Classifier

Feature Extraction

The extraction type utilized here is the DCT and four kinds of features have been extracted, namely, the features of first order, Markov features, DCT and Extended DCT features. The various features are given in Table 1.

Table 1: Features used for steganalysis

Type of Feature	Method
First order (Statistical Features)	Mean
	Standard Deviation
	Skewness
	Kurtosis
DCT	Variance
	Blockiness
	Co-occurrence
	Individual Histogram
	Dual Histogram
	Global Histogram
Extended DCT	Derived From DCT features
Markovian	Features through inter-block and intra-block Markov matrix

The regular features of DCT (15) will contain 23 functions, which can be made comprehensive to get extended features of DCT. 193 such functions can be extended (32). Another feature set used is the Markovian features. The dimensionality is high for this and hence the features are condensed to get only 81 vital features using PCA.

VI. LIMITATIONS OF FEATURES

Using the DCT and Markovian features as such would pose some issues, which are provided in Table 2.

Table 2: Limitations of DCT and Markovian Features.

Type of features	Limitations
DCT	Inter-block dependencies among the coefficients of DCT
Markovian	Intra-block dependencies between the constants of DCT ; fails to identify messages of short length

The limitations are addressed by integrating both of them and also to have more classification accuracy in terms of high value for rate of detection and low value for rate of false positive (FAR). Not all the features are calibrated initially.

Calibration

The macroscopic qualities of cover image could be evaluated from the stego image by the calibration technique (17). In this technique, the original JPEG image (J1) is transformed into the spatial domain following the decompression. A vertical and horizontal cut is made then for every pixel. Finally, the image has been transformed back (recompressed) to DCT by means of identical quantization matrix. These steps are taken to hold the DCT coefficients. Still, the data embedded would be deleted, changing the image similar to that of cover image. The image after the process is J2 and it would possess the features (macroscopic) just as the cover imagery as the image cropped is equivalent to the real image. The process of calibration is displayed in Figure 5.

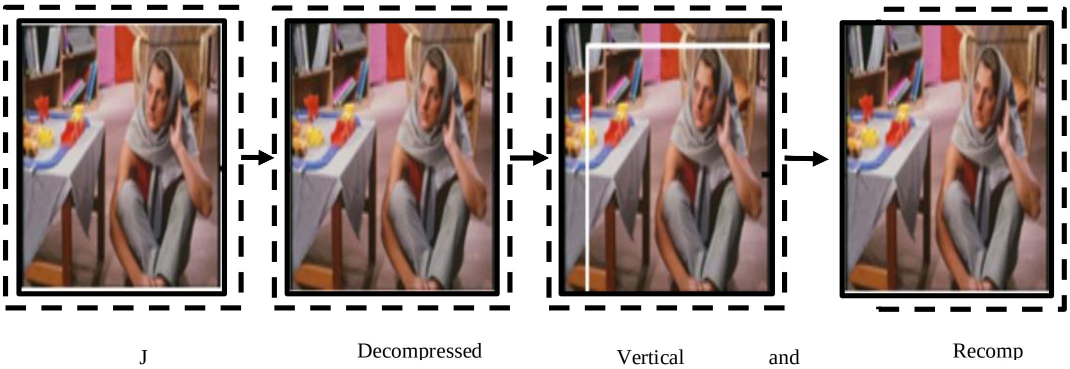

Figure 5: Calibration of JPEG Image

The DCT features have been mined and it is calculated as per the following steps:

- Calculate the difference of cover and stego images
- Consider the absolute value
- Find the L1 Norm
- The result is the DCT feature.

However, some of the pertinent features that are required for the investigation would be missed during the process of DCT extraction. Therefore, some functional with projected differences have been used in DCT, which are the features of extended DCT.

The Markovian features have been mined and it is computed as per the following steps:

- Find the absolute values of adjacent DCT constants
- Calculate the difference

The functional of Markovian itself counts to 324 features. All these features, if applied as such, would make dimensionality issues. Hence, it is converted to 4 set of dimensionality of 81. Since, the Markovian and DCT features sets are combined for the reasons stated above; the resultant combined set will carry just 274 features.

Steganographic Systems

The popular schemes like Pixel Value Differencing (PVD) and F5, which are the random algorithms for steganography, have been utilized for implanting/embedding in the JPEG images. In PVD, the cover image is merely sectioned into various blocks of two pixels that are not overlapping (37). Every single block is characterized as per the gray value difference of the two-pixels sectioned in the block. The edge region is much focused than flat/smooth region for data embedding. The input image is described as true color RGB in case of F5 (42). Different percentages of embedding are applied to the images considered, by means of PVD and F5. The complete 274 features are accounted here.

VII. RESULTS AND DISCUSSION

Dataset Details

The successful classification lies greatly on the type and size of dataset chosen. The dataset actually would possess diverse textures, images of different qualities and formats. A dataset holding a total of 420 cover and stego imageries of JPEG format having has been taken for analysis. The standardization of the dataset makes the size as 256X256. The extensive use of JPEG images on the web has made them to be preferred for analysis of this work. For performing embedding, the algorithms adapted are PVD and F5. The features are extracted using DCT and dimensionality reduction has been done with the help of PCA. The extracted features are fed to the SVM classifier that uses spate images for training and testing; obviously, the number of images taken for training would be higher than that used for testing. The number of images used for training and testing are 340 and 80 respectively (The ratio of split is 80:20 roughly).

Features Considered for Analysis

For proceeding with the experimentation, one has to consider whether all the 274 features are to be considered or just the imperative features are to be accounted. Hence, the test has been initially carried out with all features inclusive. The classification results for F5 and PVD is provided in Table 3. It might be seen from Table 3, that the classification accuracy is not appreciable. The maximum accuracy is obtained with 50 percentage of embedding. The pictorial representation of results is shown in Figure 6.

Table 3: Results of Classification with All the features inclusive

Embedding Percentage	F5	PVD
10	59.25	57
25	40.7	54.5
50	62.5	83
75	47	50
10-25-50-75	45	50

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

It might be seen from Table 3, that the classification accuracy is not appreciable. The maximum accuracy is obtained with 50 percentage of embedding. The pictorial representation of results is shown in Figure 6. The Principal Component Analysis provides better feature reduction. To check the same, an arbitrary set containing just "50" features is applied for checking the efficiency of classification without involving PCA and it just produced 50%. An identical feature set is executed with PCA and the efficiency has been improved by 40% by diminishing the feature dimensions to 20. The results are shown in Figure 7.

Hence, it is decided to use separate features (Reduction of features) by means of PCA. "Separate features" here refer to certain combination of features and not all the features. The classification has been done with various combinational features using SVM for calculating the error. The combination that provides minimum error is considered for further analysis. The results of error calculated are listed in Table 4.

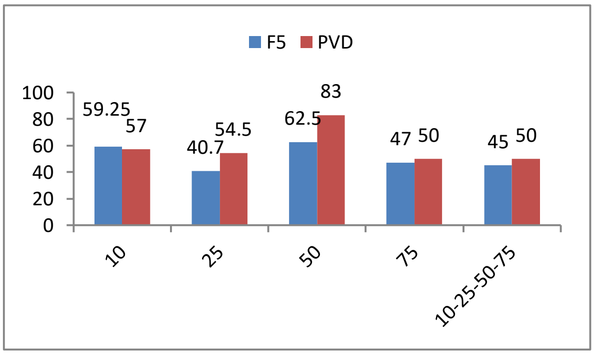

Figure 6: Classification Results with "All Features Inclusive"

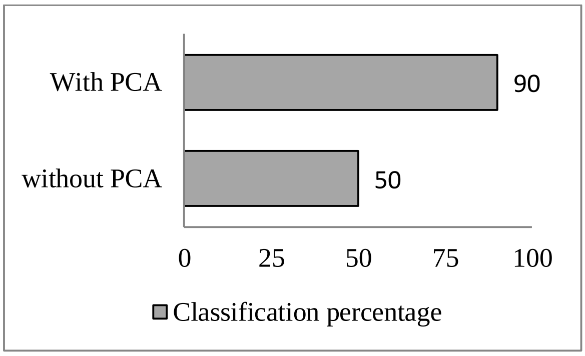

Figure 7: Efficiency Enhancement with PCA

Table 4: Results of rate of performance combination of feature sets of images using SVM

Combination of features	**Error**	**Combination of features**	**Error**
Markovian , Dual Histogram	0.1	Markovian , Variance, Moment , Dual Histogram	0.04
Blockiness, Co-occurrence, Global Histogram	0.095	Blockiness, Variance, Moment, Co-occurrence	0.0522
Variance, Moment, Dual Histogram	0.13	Markovian, Blockiness, Variance, Co-occurrence	0.12

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

Blockiness, Moment, Dual Histogram	0.123	Blockiness, Variance, Moment, Dual Histogram, Global Histogram	0.058
Variance, Moment ,Co-occurrence	0.098	Blockiness, Variance, Moment, Dual Histogram, Global Histogram	0.098
Variance, Moment, Dual Histogram, Global Histogram	0.09	Blockiness, Variance, Moment, Co-occurrence, Global Histogram	0.178
Blockiness, Moment, Co-occurrence, Global Histogram	0.112	Blockiness, Variance, Moment, Co-occurrence, Global Histogram, Markovian	0.038
Variance, Moment ,Co-occurrence ,Dual Histogram	0.13	Markovian, Blockiness, Variance, Moment , Dual Histogram	0.189
Markovian, Moment, Co-occurrence ,Dual Histogram	0.099	Blockiness, Variance, Moment, Co-occurrence, Dual Histogram, Global Histogram	0.154
Blockiness ,Variance, Moment , Dual Histogram	0.043		

Table 5: Classification Results with “Selected Features”

Percentage Embedding	First order features		Extended DCT Features								Markovian Features	
	Moment		Global Histogram		Variation		Blockiness		Co occurrence			
	PVD	F5	PVD	F5	PVD	F5	PVD	F5	PVD	F5	PVD	F5
10	50	66	62.5	60	80	80	60	100	67	83.4	80	80
25	50	75	57.1	40	60	75	60	83.3	60	66	67	71.4
50	62.5	50	66.6	57.1	75	80	60	60	71.4	69.2	75	100
75	57.1	57	80	57	60	75	50	100	66.6	83	83.3	99.8
10-25-50-75	40	57.1	50	50	75	80	60	69.2	40	71.4	80	89.4

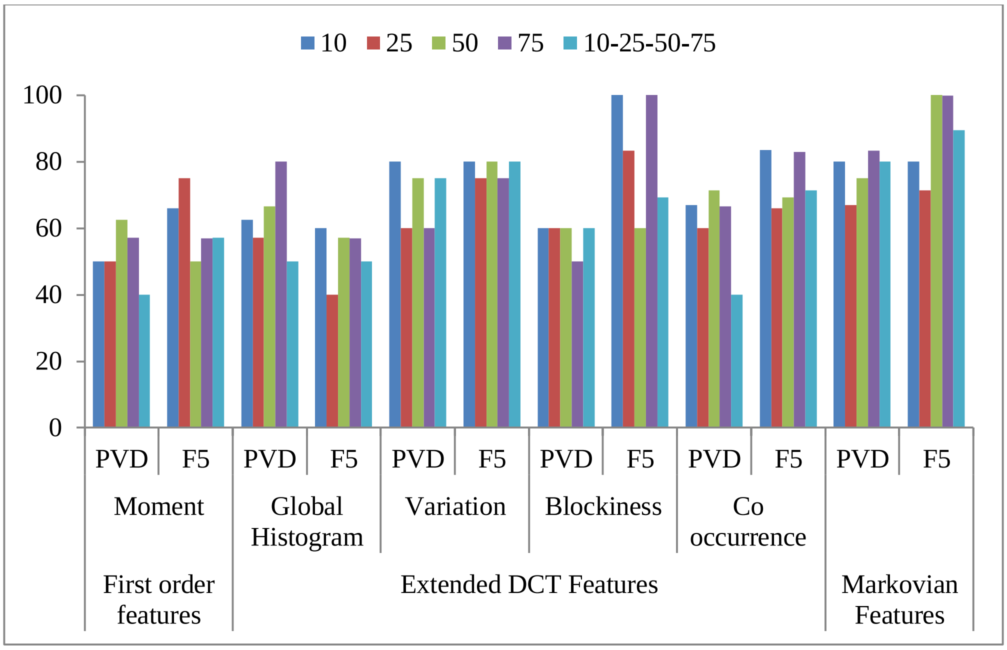

Figure 8: Classification Results with "Selected Features"

It is evident from Table 5 that the minimum error is 0.038 and hence the corresponding features have been considered for analysis. This feature reduction has been performed through PCA. Now the classification is performed with the separate features decided by PCA. The results are tabulated and given in Table 6. The results reveal that the classification performance has been improved by eliminating the redundant features and including only the vital features. The visual representation of classification using PVD and F5 is displayed in Figure 8. Now, the feature based steganalysis is checked for the importance of calibration of JPEG images of given dataset. The study has been carried out in a similar manner as that done with non-calibrated images. The final classification results with regard to F5 and PVD for both calibrated and non-calibrated images are given in Tables 6 and 7. The embedding percentages considered are 10, 25, 50 and 75 percentages. The respective graphical analysis is given in Figures 9 and 10.

Table 6: Classification Results of non-calibrated JPEG Images

Without Calibration				
Percentage of embedding	10	25	50	10-25-50-75
F5	59.5	46.6	60	50
PVD	57.2	54.5	83	50

Table 7: Classification Results of Calibrated JPEG Images

With Calibration				
Percentage of embedding	10	25	50	10-25-50-75

F5	83.3	89.3	95	75
PVD	83	91	98	75

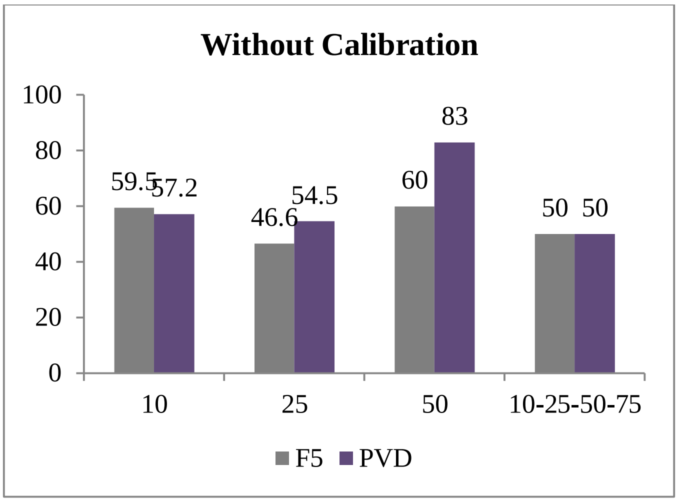

Figure 9: Classification Results of non-calibrated JPEG Images

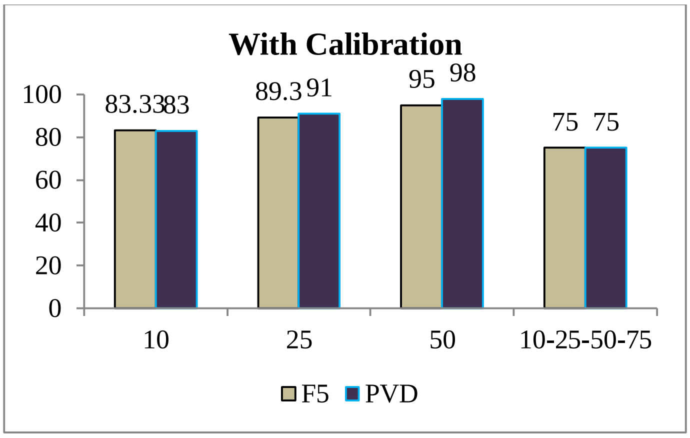

Figure 9: Classification Results of Calibrated JPEG Images

From the Tables 6 and 7, it is revealed that the classification performance is amplified by the use of calibration.

VIII. CONCLUSION

A feature based steganalysis has been performed using DCT, Extended DCT and Markovian features of JPEG imageries. The impact of features has been studied broadly and the unwanted features are eliminated using PCA. The JPEG images are used for the study because of its availability and fame. The analysis has been done without doing any calibration. Then, the calibration has been performed and the same analysis is performed. The extracted features that influence the study alone are fed to the SVM classifier, wherein which, the linear hyper-plane separated the cover and stego images in an efficient manner. The algorithms used entail the F5 and PVD. The comparative analysis has been conducted in two different views; one “with all features” and “with selected features” and the other “with calibration” and “without calibration”. It has been also analysed the effect of PCA in

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

lessening of dimensions. The 50% embedding of have given highest classification accuracy/efficiency in both calibrated and non-calibrated JPEG imageries. The results of calibrated case are superior to that of non-calibrated case.

From the entire analysis the following conclusions have been arrived:

- Features of the image have high impact on deciding the classification accuracy.
- Not all the extracted features are needed for classification, as it would increase the size/dimensions of the feature vector.
- Feature reduction is hence a crucial part in steganalysis. Anyhow, care should be taken while removing the features. The redundant and unwanted features that do not affect the classification performance alone should be removed. Principal Component Analysis is a proven method for efficient feature reduction.
- Instead of using a single feature, a combinational feature (not all the features) approach will yield better efficiency.
- The classification methodology depends on the type of problem and its application. Mostly, SVM classifier provides a competent solution.
- The calibration improves the classification efficacy.

REFERENCES

1. Al-Shaaby, A., & Al-Kharobi, T. (2017). Cryptography and Steganography: New Approach. Transactions on Networks and Communications (Vol. 5).
2. Alimoradi, D., & Hasanzadeh, M. (2014). The Effect of Correlogram Properties on Blind Steganalysis in JPEG Images. Journal of Computing and Security, 1(1), 39–46.
3. Amritha, P., & Adathil, A. (2014). Payload Estimation in Universal Steganalysis. Defence Science Journal, 60(4), 412–414.
4. Bachrach, M., & Shih, F. Y. (2011). Image steganography and steganalysis. Wiley Interdisciplinary Reviews: Computational Statistics, 3(3), 251–259.
5. Badr, S. M., Smaial, G., & H. Khalil, A. (2014). A Review on Steganalysis Techniques: From Image Format Point of View. International Journal of Computer Applications, 102(4), 11–19.
6. Cao, G., & Bouman, C. (2008). Covariance estimation for high dimensional data vectors using the sparse matrix transform. Advances in Neural Information Processing Systems (NIPS), 1–9.
7. Clark, D. (2006). Variance and covariance due to inflation. CAS Forum, 61–95.
8. Das, S., Das, S., Bandyopadhyay, B., & Sanyal, S. (2011). Steganography and Steganalysis: Different Approaches.
9. Duric, Z., Jacobs, M., & Jajodia, S. (2005). Information Hiding: Steganography and Steganalysis (pp. 171–187).
10. Eltyeb, E. (2013). Comparison of LSB Steganography in BMP and JPEG Images. International Journal of Soft Computing and Engineering (IJSCE), 3(5), 91–95.
11. Fazli, S., & Zolfaghari-Nejad, M. (2012). A New Steganalysis Method for Steganographic Images on DWT Domain. International Journal of Science and Engineering Investigations, 1(2), 1–4.

12. Fridrich, J. (2004). Feature-Based Steganalysis for JPEG Images and Its Implications for Future Design of Steganographic Schemes (pp. 67–81)

13. He, F. Y., Zhong, S. P., & Chen, K. Z. (2012). JPEG Steganalysis Based on Feature Fusion by Principal Component Analysis. Applied Mechanics and Materials, 263–266, 2933–2938.

14. Huang, F., & Huang, J. (2009). Calibration based universal JPEG steganalysis. Science in China Series F: Information Sciences (Vol. 52).

15. Jain, D., & Singh, V. (2018). An Efficient Hybrid Feature Selection model for Dimensionality Reduction. Procedia Computer Science, 132, 333–341.

16. Kanan, C., & Cottrell, G. W. (2012). Color-to-Grayscale: Does the Method Matter in Image Recognition? PLoS ONE, 7(1), e29740.

17. Kodovský, J., & Fridrich, J. (2009). Calibration revisited, 63.

18. Kumar, T., & Verma, K. (2010). A Theory Based on Conversion of RGB image to Gray image. International Journal of Computer Applications (Vol. 7).

19. Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. Nature Methods, 14(7), 641–642.

20. Li, E., & Yu, J. (2017). A Forensic Mobile Application Designed for both Steganalysis and Steganography in Digital Images. Electronic Imaging, 2017(6), 84–89.

21. Liu, Q., Sung, A. H., Qiao, M., Chen, Z., & Ribeiro, B. (2010). An improved approach to steganalysis of JPEG images. Information Sciences, 180(9), 1643–1655.

22. Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., Laishram, M. (2017). Principal Component Analysis. International Journal of Livestock Research, 1.

23. Mohammadi, F. G., & Abadeh, M. S. (2014). Image steganalysis using a bee colony based feature selection algorithm. Engineering Applications of Artificial Intelligence, 31, 35–43.

24. Mohammadi, F. G., & Sajedi, H. (2017). Region based Image Steganalysis using Artificial Bee Colony. Journal of Visual Communication and Image Representation, 44, 214–226.

25. Pathak, P., & Selvakumar, S. (2014). Blind Image Steganalysis of JPEG images using feature extraction through the process of dilation. Digital Investigation, 11(1), 67–77.

26. Paul, L. C., Suman, A. Al, & Sultan, N. (2013). Methodological Analysis of Principal Component Analysis (PCA) Method. IJCEM International Journal of Computational Engineering & Management ISSN, 16(2), 2230–7893.

27. Pevny, T., & Fridrich, J. (2007). Merging Markov and DCT features for multi-class JPEG steganalysis. Security, Steganography, and Watermarking of Multimedia Contents IX, 6505, 650503.

28. Prasad, S., & Pal, A. K. (2017). An RGB colour image steganography scheme using overlapping block-based pixel-value differencing. Royal Society Open Science, 4(4), 161066.

29. Priya, R. L., Eswaran, P., & Kamakshi, S. L. P. (2013). Blind Steganalysis with Modified Markov Features and RBFNN, 2(5), 2031–2038.

30. Pujari, M. A. A., & Shinde, M. S. S. (2016). Data Security using Cryptography and Steganography. IOSR Journal of Computer Engineering, 18(04), 130–139.

31. Rabee, A. M., Mohamed, M. H., & Mahdy, Y. B. (2018). Blind JPEG steganalysis based on DCT coefficients differences. Multimedia Tools and Applications, 77(6), 7763–7777.

32. Sabeti, V., Samavi, S., Mahdavi, M., & Shirani, S. (2007). Steganalysis of Pixel-Value Differencing Steganographic Method. In 2007 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (pp. 292–295). IEEE.

33. Sabnis, S. K., & Awale, R. N. (2016). Statistical Steganalysis of High Capacity Image Steganography with Cryptography. Procedia Computer Science, 79, 321–327.

34. Swagota Bera, M., & Sharma, M. (2015). A Blind Steganalysis on JPEG Gray Level Image Based on Statistical Features and its Performance Analysis. International Journal of Engineering Research and Development, 11(09), 2278–67.

35. Tanwar, R., & Malhotrab, S. (2017). Scope of Support Vector Machine in Steganography. IOP Conference Series: Materials Science and Engineering, 225, 012077.

36. Veena, S. T., & Arivazhagan, S. (2018). Quantitative steganalysis of spatial LSB based stego images using reduced instances and features. Pattern Recognition Letters, 105, 39–49.

37. Vigil, A., Rathor, S. S., & Singh, J. (2016). Secure binary image steganography using F5 algorithm based on data hiding and diffusion techniques (Vol. 9).

38. Voropaev, M. (2009). Variance-covariance based risk allocation in credit portfolios: analytical approximation, (May), 1–9.

39. Watanabe, S., Murakami, K., Furukawa, T., & Zhao, Q. (2016). Steganalysis of JPEG image-based steganography with support vector machine. In 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD) (pp. 631–636). IEEE.

40. Wiggins, R. H., Davidson, H. C., Harnsberger, H. R., Lauman, J. R., & Goede, P. A. (2001). Image File Formats: Past, Present, and Future. RadioGraphics, 21(3), 789–798.

41. Xuan, G., Cui, X., Shi, Y. Q., Chen, W., Tong, X., & Huang, C. (2007). JPEG Steganalysis Based on Classwise Non-Principal Components Analysis and Multi-Directional Markov Model. Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, ICME 2007.

42. Zhang, J., Lu, W., Yin, X., Liu, W., & Yeung, Y. (2019). Binary image steganography based on joint distortion measurement. Journal of Visual Communication and Image Representation, 58, 600–605.