Steganalysis Embedding Percentage Determination with Learning Vector Quantization

Benjamin M. Rodriguez, Gilbert L. Peterson, Kenneth W. Bauer, and Sos S. Agaian, *Member, IEEE*

***Abstract*—Steganography (stego) is used primarily when the very existence of a communication signal is to be kept covert. Detecting the presence of stego is a very difficult problem which is made even more difficult when the embedding technique is not known. This article presents an investigation of the process and necessary considerations inherent in the development of a new method applied for the detection of hidden data within digital images. We demonstrate the effectiveness of Learning Vector Quantization (LVQ) as a clustering technique which assists in discerning clean or non-stego images from anomalous or stego images. This comparison is conducted using 7 features [1] over a small set of 200 observations with varying levels of embedded information from 1% to 10% in increments of 1%. The results demonstrate that LVQ not only more accurately identify when an image contains LSB hidden information when compared to k-means or using just the raw feature sets, but also provides a simple method for determining the percentage of embedding given low information embedding percentages.**

I. Introduction

STEGANOGRAPHY is the communication of a message which is embedded within an image and shared between two people, the sender (transmitter) and the receiver. Principally, the strategy from the steganographic perspective consists of two tasks: limiting image distortions within an acceptable range, and preservation of the message. Both of these tasks are geared to ensuring that the correct message is delivered and that its transmission goes unnoticed. From the steganalysis perspective, the objective is the detection of the hidden message embedded within a transmitted signal that otherwise passes unnoticed.

As technology dictates, there must always be progression. To have a successful classification based steganography detection algorithm one must calculate features which are sensitive to the embedding process and combine the features with a strong classification algorithm capable of classifying the images using the calculated features. The difficulty for the steganalysis is that as the steganalysis tools improve, so do the steganography tools, and with so many of the steganography tools being open source it does not take a great deal for a tool to be modified and the statistical image irregularities that it generates altered. It is because of this alteration that one must detect not just the known steganography techniques but those that have never been seen before. This problem is known as the blind detection problem, and requires the classification algorithm classify normal and abnormal images having only seen normal images.

In the quest for the development of a tool to classify steganography in a blind manner, we present a new method for steganalysis that combines pixel comparisons features with several new features, and a newly modeled self-organizing map (SOM) which uses (LVQ) as a second stage for improved blind classification. The modified SOM algorithm separates the non linear data into a new feature space where it is simpler to determine if the image is clean/normal or dirty/abnormal. The key component in the new method is the clustering of three sets of features for the purpose of successfully identifying received stego containing images. The new method uses the evaluation of texture features to facilitate recognition and classification for anomalous or stego images. Experimental results show that the performance of the new proposed method is improved when compared with traditional clustering methods and self organizing maps.

This paper discusses related worked in detection and classification in Section II. In Section III the feature extraction is described. K-means clustering methods used for classification purposes are described in Section IV. Section V, describes the proposed modified LVQ which is used for classification. In Section VI, the experimental results are shown and in Section VII the conclusion is given.

II. Related Work

This section discusses work related to the blind steganalysis detection problem. Specific topics include image feature extraction methods, other blind classification approaches, and other work using hierarchical clustering.

A. Feature Extraction Methods

Several methods based on some sort of statistical measure are:

a) Methods based on pixel comparison statistic. See for example Agaian [2], Dumitrescu [3], Fridrich [4] and Johnson [5]. In spite of such, these approaches have provided significant detection accuracy of LSB embedding; but there are some limitations with these pixel comparison detection

B. M. Rodriguez is with the Air Force Institute of Technology, Wright-Patterson AFB, OH 45433 USA (e-mail: benjamin.rodriguez@afit.edu).

G. L. Peterson is with the Air Force Institute of Technology, Wright-Patterson AFB, OH 45433 USA (e-mail: gilbert.peterson@afit.edu).

K. W. Bauer is with the Air Force Institute of Technology, Wright-Patterson AFB, OH 45433 USA (e-mail: kenneth.bauer@afit.edu).

S. S. Agaian is with the University of Texas at San Antonio, San Antonio, TX 78249 USA, (e-mail: sos.agaian@utsa.edu).

methods, i.e. the inability to detect small amounts of hidden information and difficulties or inability when detecting compressed images.

a) Feature based and Histogram methods such as: Chi squared by Westfeld [6], Cosine transform coefficient histogram [7], individual histogram [7], dual coefficient histogram [7], inter-block dependencies[7], and blockiness [8], have all been applied for lossy compression detection. These methods can detect messages hidden in JPEG images using steganographic algorithms such as F5. The key elements of the histogram methods are that they compare the estimated histograms of the selected coefficients with those of the stego image. Some of these techniques were designed for only specific steganography algorithm.

b) Higher order statistic and wavelet based stego detection methods such as; Co-occurrences [7], [9] and wavelet based [10] which was presented by Lyu and Farid and Feature based methods such as: multilevel based features by Agaian et. al. The wavelet based method uses a large data base of over 40,000 images to develop feature vectors for classification. The basic motivation of this method is: Stego-images are perceptually identical to cover images, but they exhibit statistical irregularities. Detect statistical irregularities observing selected image features. Farid argues that most steganalysis attacks look at only first order statistics. But new techniques try to keep the first order statistics intact. Optimal linear predictor for wavelet coefficients and calculates the first four moments of the distribution of the prediction error. These include mean, variance, kurtosis, skewness. Various classifier types are then used to separate stego-message from cover-images.

The basic limitations of these methods are: 1) the uncertainties of how these methods works; 2) is classification accuracy compromised for a very limited image database; 3) for specific embedding methods can the hidden data be detected and 4) how computationally expensive are these methods. Embedding methods like Model Based [11] which fit the coefficient histogram to the model maximum likelihood and modify the coefficients to maintain the model are difficult to detect based on the properties of current detection methods. Unlike F5 [12] which has been proven to be "steganalysis broken [13]."

A. Blind Classification

Several articles throughout the past few years have been proposed which address various methods for detecting steganographic information within digital images. McBride and Peterson proposed a blind detection method for anomaly detection using a hyper-convex polytope to create a self class model, and a modified k-means using spherical and elliptical representations [14]. In [15], Lyn and Farid exploited color statistics to show how a one-class SVM greatly simplifies the training stage of the classifier by eliminating the need for training from stego images which makes for a blind classifier of common and future developed stego programs.

B. Concepts of Clustering Methods

Recently, several clustering methods such as hierarchical clustering and self-organizing clustering have been successfully used for data analysis, digital audio signals classification, identifying numerous voices within an audio signal and several other applications. Texture feature characterization in digital image processing applications is a well established technique. These texture features are extracted using a wide range of available methods for classification. Another clustering method used by Flietstra et. al incorporated the Radial Basis neural network as new clustering method [16].

As technology dictates, there must always be progression. To have a successful classification based stego detection algorithm one needs to calculate features which are sensitive to the embedding process and to find strong classification algorithm which is able to classify the images using the calculated features.

II. Feature Extraction

In this section we describe the difficulties encountered when generating features for detecting least significant bit modifications. These features emphasize minor changes in the LSB of a digital image with the use of weights when measuring pixel variations among adjacent pixels.

(a) (b) (c)

Fig. 1. This figure shows three images that contain various structures and patterns, (a) shows a predominantly smooth image in which steganography can be easily detected, (b) shows a landscape image that contains a large variation in pixel making it difficult to detect stego and (c) show an image that contains sharp edges from the colors black and white.

When constructing features from the spatial domain of an image, the type of image must be considered. Consider three basic image region types as smooth, transitional and large variation (sharp edges). Images with large areas of smooth pixel characteristics can be viewed as pixels with similarities in amplitude as seen in Figure 1 (b). Transitional regions within images contain prominent shifts in pixel differences which are shown in the landscaping of Figure 1 (b) are natural within an image. These regions are analyzed by each of the transition contours ensuring the difference between the pixel shifts are not mistaken for image manipulations, e.g. changes between grass and a red flower. Regions containing large

variation between adjacent pixels cause unnatural characteristics, such as the chessboard in Figure 1 (c). These blocks are analyzed by viewing the lower bit-planes which appear to be natural contour between pixels. Separating the input images into various categories allows for the feature generation to be correctly modeled without causing a reduction in classification accuracy.

The set of n ($n=7$) features are extracted by (see [1]):

$$\hat{f}_{i,j}^{n} = \frac{1}{K}\sum_{l}\sum_{k}\left|w_l x_l - w_k x_{k+l}\right|$$

$$F_n = \frac{1}{NM}\sum_{i,j}\hat{f}_{i,j}^{n}$$

$$F_n = \{f_1, f_2, ..., f_n\}$$

where, w is a weight vector, x is a set of pixels being analyzed, K is the number of adjacent pixels, k are the pixel locations of adjacent pixels, l is the focal pixel, M are the number of rows in the image, N represents the number of columns in the image and i, j subscripts represent the pixel location throughout the image.

III. K-Means Clustering Method

Using the features calculated in the previous section, it is possible to use k-means clustering on the input data. After performing k-means, a simple modification of the results provides a means by which k-means can be used for the blind classification problem and detect steganographic content within digital images.

Cluster analysis is an exploratory data analysis tool for solving classification problems that contain a large data set. The object is to sort data into groups, clusters or classes so that the degree of association is strong between the data sets of the same cluster and weak between members of different clusters. The challenge occurs when attempting to cluster data sets the required number of clusters necessary known as K (k-means) or when the data sets are not linearly separable. Clustering methods do a great job determining if the number of classes have divided the space properly. This is determined if the features for each class are separated from each other. K-means will divide the input data into k clusters if data can be divided into the number of desired clusters. With modifications the features of the received image can be used to determine steganographic content or the class in which the degree of association is strongest with the data set containing no steganographic content.

In [17], McBride and Peterson discuss the successful use of hyper-convex polytope to create a self class model, and with a modified k-means using spherical and elliptical representations to determine if an observation is clean or dirty. The blind classifier used geometric concepts which the boundaries of classes created by k-means were built. The geometric principles applied in the creation of the classifier were convex polytope, hyper-sphere, and hyper-ellipsoid classifiers. McBride and Peterson achieved the highest accuracy with the hyper-ellipsoid. It was concluded in [17] that the reason for improved detection accuracy was the tight space provided by the ellipsoid for the data class prevents an overabundance of false negatives while still retaining some generality and a minimum of false positives.

Agaian and Rodriguez use a modified k-means clustering in [18] to determine upper and lower boundaries of steganographic content to determine if the input image is within steganographic ranges. A steganographic capacity was used to separate the image into embeddable and non-embeddable areas to separate the image to determine the upper and lower boundaries of the embeddable information for both areas. The received image features were then compared to the upper and lower boundary classes to determine if the received image was within the range of the other member classes.

IV. Modified LVQ

In the previous section we introduced the use of k-means clustering used in previous detection algorithms to determine if steganographic content exists within an input image. In this section we introduce the use of linear vector quantization, which was adapted by Kohonen [19] for pattern recognition, in an attempt to find a linear separation within clean and stego data. When stego data and clean data cannot be separated a modified LVQ method is introduced in this section. The use of a modified linear vector quantization method is used to separate the data space with a non linear separation.

The presented method resolves the problem when the data space is not linearly separable. The linear vector quantization method is used to represent not individual values but arrays of the exemplars. Since the exemplars used determine bit changes when an image has been manipulated with non visual changes in the least significant bit plane. The problem with this is that changes preformed in the spatial domain cannot be viewed with statistical features that measure the modification in this domain. This results in a feature space that is not linearly separated. The features can be separated by transforming the results into another domain. Vector quantization is limited with the use of a linear transform when the presented features are used. To correct this problem the output features are separated by using a non linear transformation. The new transformed features are then separated linearly to determine if the observation contains stego or not.

A traditional learning vector quantization network consists of a competitive layer with supervision, and a second layer (linear layer) of the network separates the subclasses from the

first layer with a linear transformation. The various layers of LVQ network are shown in Figure 2.

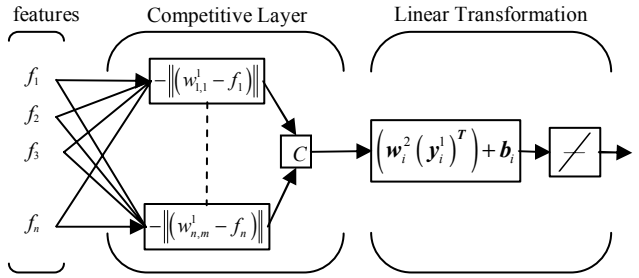


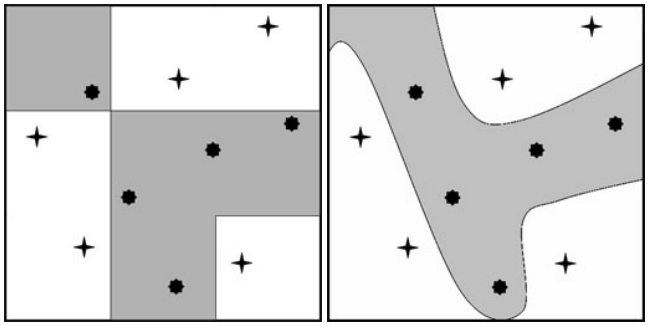
Fig. 2. This figure show the Learned Vector Quantization classifier.

This results in a classification space where new observation can be classified by the division of linear boundaries. This however does not result in a space that can guarantee accuracy of features that do not contain linear separation.

$$y_{i,j}^1 = -\left\|\left(w_{i,j}^1 - f_j\right)\right\|$$

$$\boldsymbol{y}_i^2 = \left(\boldsymbol{w}_i^2\left(\boldsymbol{y}_i^1\right)^T\right)+\boldsymbol{b}_i$$

where, the superscript 1 indicates the trained weights with the competitive layer of the learned vector quantization and the superscript 2 indicates the trained weights of the linear layer. If $\boldsymbol{y}_i^2$ is large for the corresponding target component it denotes the class to which the observation f_i belongs to.



(a) (b)

Fig. 3. (a) This figure shows two classes that have been separated with the Learning Vector Quantization Algorithm. (b)This figure shows two classes that have been nonlinearly separated with the modified Learning Vector Quantization Algorithm.

Figure 3 shows the linear separation of the classified features compared with non linear feature separation.

Since all data is not separable in a linear fashion other methods must be used for classification. We have used a modified version of Learning Vector Quantization, which achieves nonlinear separation of the observation. This is accomplished by using a nonlinear transformation instead of using a linear transformation.

$$\hat{y}_{i,j}^1 = -\left\|\left(w_{i,j}^1 - f_j\right)\right\| + b_{i,j}$$

where, $\hat{y}_{i,j}^1$ is returned features without the competitive transform. These features are trained and the new weights are used to generate $\hat{y}_{i,j}^2$. The new features $\hat{y}_{i,j}^2$ are then separated with a nonlinear transformation.

$$\hat{\boldsymbol{y}}_i^2 = \left(\boldsymbol{w}_i^2\left(\hat{\boldsymbol{y}}_i^1\right)^T\right)+\boldsymbol{b}_i$$

Figure 3 (b) shows the input features x_j separated with a nonlinear boundary. If the images in Figure 3 were representative of the differences between stego and normal images, one could imagine that the closer fit of the nonlinear boundary is more expressive and representative as the differences between the normal and abnormal classes. Figure 4 shows the proposed modified LVQ.

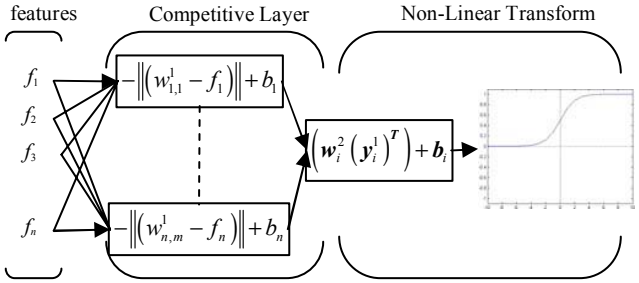


Fig. 4. This figure show the modified Learned Vector Quantization classifier.

V. RESULTS

In this section we show the results of the proposed modified Learning Vector Quantization method compared with *k*-means at handling blind classification for the steganalysis problem. The analysis was conducted with an arbitrary image data set of 200 color TIFF and RAW images formats taken with Nikon D100 and Canon EOS Digital Rebel digital cameras.

The class for the stego image data is represented as the image containing the embedded information. The embedded information was of sizes 1% to 10% in increments of 1%. This generated 20 classes of the various data which included the clean image data set. When the input image feature set resides outside of the limits for one of the stego classes the image is considered an anomaly in the detection procedure. In Table 1 we show the improvements of properly classifying a stego image with the various amounts of steganographic content when comparing LVQ and the modified classification methods. These results were calculated with a 5-fold cross validation. On average the modified LVQ shows 12% classification accuracy over LVQ.

TABLE I
IMPROVED CLASSIFICATION ACCURACY

Embedded Percentage	Classification Accuracy	
	LVQ	Modified LVQ
Clean	70%	84%
1%	71%	84%
2%	71%	84%
3%	74%	85%
4%	74%	86%
5%	75%	86%
6%	75%	87%
7%	75%	87%
8%	77%	89%
9%	78%	89%
10%	78%	89%

Overall, the modified learned vector quantization detection method shows a 10% increase when properly identifying a stego image over the k-means when used as a classifier. LVQ method was improved with the use of a nonlinear transformation of the weighted feature space which allowed for a 22% increase in classification accuracy over the k-means detection method.

VI. CONCLUSION

In this paper, we presented a new method for steganalysis that combines spatial feature extraction and newly modeled modified Learning Vector Quantization. The key component in the new method is the nonlinear separation of feature vectors for the purpose of successfully identifying received stego-containing images. Experimental results show that the performance of the new proposed method is better than other existing classification methods by an increase of 12% classification accuracy.

The modification in separating nonlinearly separable clusters allowed for classification of features which are used to determine if steganographic content exists within the digital image. Since the features were developed in the spatial domain the improvements of separating the nonlinear features with the use of a nonlinear transformation.

ACKNOWLEDGMENT

This research was partially funded by the US Air Force. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Air Force, Department of Defense or the U.S. Government. We would additionally like to express our appreciation to June Rodriguez for the contribution of a multitude of digital images for analytical support.

REFERENCES

[1] S. S. Agaian, G. L. Peterson and B. M. Rodriguez, "Multiple Masks-Based Pixel Comparison Steganalysis Method for Mobile Imaging," SPIE Defense & Security Symposium, "Mobile Multimedia/Image Processing for Military and Security Applications, 17-21 April 2006.

[2] S. S. Agaian, B. M. Rodriguez and G. Dietricht, "Steganalysis Using Modified Pixel Comparison and Complexity Measure", SPIE Symposium on Electronic Imaging, San Jose, CA, 2004

[3] S. Dumitrescu, X. Wu and Z. Wang, "Detection of LSB Steganography via Sample Pair Analysis", LNCS 2578, 355-372, 2003.

[4] J. Fridrich, M. Goljan and R. Du, "Detecting LSB Steganography in Color and Gray-Scale Images," Magazine of IEEE Multimedia Special Issue on Security, October-November 2001, pp. 22-28

[5] N.F. Johnson and S. Jajodia, "Steganalysis: The Investigation of Hidden Information", IEEE Information Technology Conference, 1998, URL: http://www.jjtc.com/pub/it98a.htm

[6] A. Westfeld, and A. Pfitzmann, "Attacks on Steganographic Systems", Proceedings of the 3rd Information Hiding Workshop, Dresden, Germany, September 28-October 1, 1999, pp. 61-75.

[7] J. Fridrich, "Feature-Based Steganalysis for JPEG Images and its Implications for Future Design of Steganographic Schemes," Proceedings of the 6th Information Hiding Workshop, Toronto, Canada, May 23-25, 2004.

[8] J. Fridrich, M. Goljan, and D. Hogea, "Attacking the OutGuess", Proceedings of the ACM Workshop on Multimedia and Security 2002, Juan-les-Pins, France, December 6, 2002.

[9] B. M. Rodriguez, S. S. Agaian and J. Rodriguez, "Co-Occurence Matrix Feature Vectors and Cluster Classification Based steganalysis", INFORM Denver, CO Oct 2004, page 221.

[10] H. Farid, "Detecting Hidden Messages Using Higher-Order Statistical Models", International Conference on Image Processing (ICIP), Rochester, NY, 2002.

[11] P. Sallee, "Model-based steganography," International Workshop on Digital Watermarking, Seoul, Korea, 2003.

[12] A. Westfeld, "F5a steganographic algorithm: High capacity despite better steganalysis," 4th International Workshop on Information Hiding, 2001.

[13] J. Fridrich, M. Goljan, and D. Hogea, "Steganalysis of JPEG Images: Breaking the F5 Algorithm", 5th Information Hiding Workshop, Noordwijkerhout, The Netherlands, 7–9 October 2002, pp. 310-323.

[14] B. McBride and G. L. Peterson, "Blind Data Classification using Hyper-Dimensional Convex Polytopes," Proceedings of the 17th International FLAIRS Conference, pp 520-525, 2004.

[15] S. Lyu and H. Farid, "Steganalysis using color wavelet statistics and one-class support vector machines," SPIE Symposium on Electronic Imaging, San Jose, CA, 2004.

[16] T. D. Flietstra, K. W. Bauer and J. P. Kharoufeh, "Integrated Feature and Architecture Selection for Radial Basis Neural Networks", International Journal of Smart Engineering System Design, Vol 5:507-516, 2003.

[17] B. McBride and G. L. Peterson, "A new blind method for detecting novel steganography", Digital Investigation, Vol 2, 50-70 ,2005.

[18] S. S. Agaian and B. M. Rodriguez, "Steganographic Capacity used for Steganalysis Cluster Classification", GSteg Pacific Rim Workshop on Digital Steganography, ACROS Fukuoka 1-1 Tenjin 1-chome, Chuo-ku, Fukuoka, 810-0001 Japan, November 17-18, 2004.

[19] T. Kohonen, *Self-Organizing and Associative Memory*, 3rd ed. Springer-Verlag, 1988.