JPEG Steganalysis Using Empirical Transition Matrix in Block DCT Domain

Dongdong Fu, Yun Q. Shi, Dekun Zou
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07102, U. S. A.
{df7,shi}@njit.edu

Guorong Xuan
Department of Computer Science
Tongji University
Shanghai, China

Abstract— **This paper presents a novel steganalysis scheme to effectively attack the JPEG steganographic schemes. The proposed method exploits the correlations between block-DCT coefficients in both intra-block and inter-block sense. We use Markov empirical transition matrices to capture these dependencies. The experimental results demonstrate that the proposed scheme is superior to the existing steganalyzers in attacking OutGuess, F5, and MB1.**

Keywords— steganalysis; steganography; conditional probability matrix; support vector machine (SVM)

Topic area— Multimedia security

I. INTRODUCTION

The past few years have witnessed the advance in both steganography and steganalysis techniques. Steganography is the art and science of covert communication. It attempts to hide information in such a way that no one else would be aware of the existence of the hidden information except the intended recipient. The aim of steganalysis, on the other hand, is to detect the very existence of information hidden by steganography. As steganographic techniques become more and more sophisticated, more advanced steganalysis schemes are needed in practice. Nowadays, digital image is very popular and can be obtained almost everywhere. It can be easily modified and transferred through Internet. It also can tolerate large amount of perceptually redundant information without any visible clue. Therefore, digital image is an ideal cover media for steganography yet a big challenge for steganalysis. Since the JPEG (Joint Photographic Experts Group) format is the most dominant image format for image storage and exchange at this time, the JPEG steganographic techniques have attracted more and more attentions. In this paper, we focus on attacking JPEG related steganographic schemes.

Several steganographic techniques working on JPEG images have been proposed in recent years. In this paper we focus on attacking three recently published and most advanced steganographic methods, i.e., Outguess [1], F5 [2], and the model-based steganography (MB) [3]. To attack them, we will discuss these methods briefly at first.

OutGuess [1] is a universal steganographic scheme that embeds hidden information into the redundant bits of data sources. For JPEG images, it embeds messages in DCT domain by first identifying the redundant block-DCT coefficients to be modified. To resist the steganalysis attacks based on detecting the change of block-DCT coefficients histogram, OutGuess then adjusts the reserved coefficients during the embedding procedure to preserve the original global histogram of the DCT coefficients after embedding.

F5 [2] works on JPEG by modifying the block-DCT coefficients to embed messages. Instead of flipping the LSBs of the DCT coefficients, it always reduces the absolute value of a non-zero DCT AC coefficient by one. F5 improves the embedding efficiency by utilizing the matrix coding. However, F5 cannot maintain the original histogram of the block-DCT coefficients.

MB [3] preserves the marginal statistics of the cover image more successfully. In MB1 [3], which implements MB principle on JPEG images, a generalized Cauchy distribution is used to model the block-DCT mode histogram. By using the non-adaptive arithmetic decoder, the embedding procedure keeps the lower precision version of each block-DCT mode histogram unchanged. It is claimed that MB can resist first order statistical attacks.

As the battle between steganography and steganalysis always continues, several steganalysis techniques have been proposed in literature.

In [4], Farid presented a general steganalysis scheme based on image's high order statistics in wavelet domain. These statistics are based on decomposition of images with separable quadrature mirror filters. The subbands' statistical moments are obtained as features for steganalysis. This scheme (denoted by MW in this paper), though successful to some extent, does not perform adequately especially for attacking advanced JPEG steganographic techniques such as we discussed above in this paper.

Shi *et al.* proposed a universal steganalysis system in [5]. The statistical moments of characteristic functions of the image, its prediction-error image, and their discrete wavelet transform (DWT) subbands are selected as features. All of the low-low wavelet subbands are also used in their system. This steganalyzer (denoted by MC in this paper) can provide a better performance than [4] in general, but, still cannot attack JPEG steganography adequately.

In [6], Fridrich developed a steganalysis scheme specifically designed for JPEG steganography. The main part

of this scheme is to estimate the statistics of the original image. A set of well-selected features for steganalysis are generated from the statistics of the JPEG image and its estimated version. This scheme (denoted by JF in this paper) performs better than [4] and [5] in attacking JPEG steganography mentioned above [1-3].

Sullivan *et al.* [7] presented a steganalysis technique based on Markov chain model which captures the inter-pixel dependencies in the image. Because the size of the calculated empirical transition matrix is very large, e.g., the 65536 elements for a gray-level image for a bit depth of 8, it cannot be used as features directly. The authors select several largest probabilities along the main diagonal together with their neighbors, and randomly select some other probabilities along the main diagonal as features, resulting in a 129-dimensional (129-D) feature vector. This technique, though was designated to attacking spread spectrum (SS) data hiding, provides some insights that have motivated our steganalysis method in attacking JPEG steganography. Specifically, the Markov empirical transition matrix, a second order statistics, inspires us to investigate higher order statistics in DCT domain to attack modern sophisticated JPEG steganographic techniques which generally maintain the first order statistics of the cover data.

In this paper, we propose a novel steganalysis scheme to effectively attack the advanced JPEG steganographic methods. In our scheme, Markov empirical transition matrices are proposed to capture both intra-block and inter-block dependencies between block-DCT coefficients in JPEG image. Since the hidden messages are sometimes independent to the cover data, the embedding process often decreases the dependencies exiting in original cover data to some extent. Therefore, the proposed second order statistics can capture such kind of changes. To reduce high dimensionality of the proposed empirical transition matrices, a threshold technique is applied to generate efficient features. Finally we evaluate the proposed features with support vector machines (SVM) as classifier by conducting experiments over a diverse data set of 7560 JPEG images. The superior results have demonstrated the effectiveness of our proposed scheme.

The rest of this paper is organized as follows. Section II discusses the proposed scheme for feature generation. Classification experiments are presented in Section III and conclusions are drawn in Section IV.

II. Proposed Scheme for Feature Generation

The steganalysis is tackled under the two-class pattern recognition framework in our work. In other words, a given image needs to be classified as either a stego image (with hidden data) or a non-stego image (without hidden data). Therefore, the generation of appropriate features is the most important part in our scheme.

We observed that the above mentioned JPEG steganographic methods such as OutGuess and MB have made great efforts to maintain the marginal histogram of the block-DCT coefficients (first order statistics). Although F5 does not preserve the block-DCT coefficients histogram explicitly, it still tries to keep the histogram appear unchanged by decreasing or increasing the DCT coefficient values by only one. This fact suggests that the steganalysis schemes based only on first order statistics is not sufficient. In this paper, we propose to employ higher order statistics for steganalyzing the JPEG steganography. To avoid increasing the computational complexity dramatically, only the second order statistics are used in this work.

First order statistics refer to the statistical measures which capture only the properties of individual sample, ignoring the inter-dependencies between data in the dataset. On the other hand, second or higher order statistics consider not only the values of two or more observations but also their position relative to one another in the dataset [8]. In this sense, mean, variance and histogram are typical first order statistics, whereas covariance, co-occurrence matrix and empirical transition matrix are typical second order statistics.

The local dependency between image pixels is a well-known property in image processing applications. This property has been also used by [7] for steganalysis to detect spread spectrum data hiding as we have mentioned in Section I. The dependency among quantized block-DCT coefficients, though may not be widely used, compared with its counterpart in spatial domain, for steganalysis, have been actually exploited by researchers for a long period of time for image compression.

There are three kinds of dependencies between quantized block-DCT coefficients which have been summarized in [9]: intra-block correlation, inter-block correlation and sign correlation.

1. The intra-block correlation states that although the block-DCT coefficients are decorrelated by the transform to some extent, two block-DCT coefficients within one block are not independent. A coefficient still has correlation with its neighbors in same block [9]. It is also well-known that the general trend in magnitude of the block-DCT coefficients in each block is non-increasing along the zig-zag scan order of all of the DCT coefficients in the block if we ignore some up-and-down of small magnitudes. This fact inspires us to arrange 8x8 block-DCT coefficients in one block into a 1-D vector in zig-zag order such that the intra-block dependency can be exploited more efficiently.

2. Inter-block correlation: a block-DCT coefficient in one block has somewhat strong correlation with the coefficient of the same position in one of the neighbor blocks [9]. Fridrich [6] has taken into account this kind of inter-block correlation. Several features derived from second order statistics, co-occurrence matrix, are used to capture the dependency between the DCT coefficient pairs from neighboring blocks. In her observation, the co-occurrence matrix of DCT coefficients is “one of the most influential features”. This fact has also justified that second order statistics based features outperform the features based on first order statistics in attacking marginal statistics (first order) preserving JPEG steganography. Although Fridrich’s approach is successful to some extent, the intra-block correlation is not considered in [6].

3. Sign correlation is also possible to be exploited [9]. However, today's steganographic schemes generally do not touch the signs of the DCT coefficients. Hence, we do not exploit the sign correlation in the proposed steganalysis scheme. Therefore, only the absolute value (i.e., magnitude) of block-DCT AC coefficients are consider in the proposed scheme if not denote explicitly otherwise. The DC components are not considered because the DC components are generally not changed in JPEG steganograph.

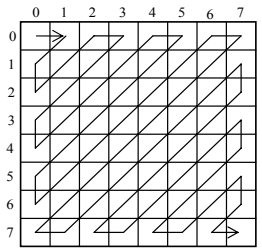

Figure 1 Block-DCT coefficients zig-zag scanning pattern

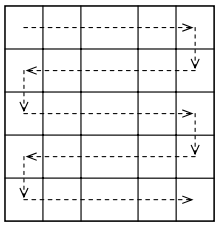

Figure 2 Block scanning pattern

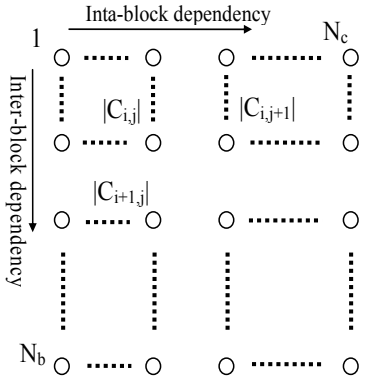

Figure 3 Reordered block-DCT AC coefficients 2-D array

Based on the discussions above, we propose a feature extraction algorithm as follows.

a) According to the zig-zag scanning pattern shown in Figure 1, expand each set of 8×8 block-DCT coefficients into a 1-D row vector. ***b)*** Arrange the blocks by following the scanning pattern shown in Figure 2. ***c)*** A new DCT AC coefficients 2-D array is formed as shown in Figure 3. In this 2-D array, we have coefficients of same block in the same row and coefficients of same frequency (or mode) in the same column. In this way, the dependencies among neighboring coefficients in both horizontal and vertical directions can be conveniently investigated. Note that we discard the DC components and sign as discussed above. Thus, the elements of the 2-D array in Figure 3 are actually the absolute value (magnitude) of the block-DCT AC coefficients. We further get rid of some high frequency coefficients in each row and only keep the L lowest AC frequency coefficients in zig-zag order because most of the high frequency coefficients are quantized to zero, thus not so much information can be exploited. The L can be adjusted according to the Q-factors of the testing images. ***d)*** We propose to model the special 2-D array shown in Figure 3 by using Markov random process. To capture the dependencies between intra-block coefficients in each row, we calculate the horizontal Markov empirical transition matrix of the 2-D array in Figure 3. Its element is given by:

$$p_h(|C_{i,j+1}| = n \mid |C_{i,j}| = m) = \frac{\sum_{i=1}^{N_b-1}\sum_{j=1}^{N_c-1}\delta(|C_{i,j}| = m, |C_{i,j+1}| = n)}{\sum_{i=1}^{N_b-1}\sum_{j=1}^{N_c-1}\delta(|C_{i,j}| = m)}$$

where $|C_{i,j}|$ and $|C_{i,j+1}|$ are absolute neighboring DCT AC coefficients pair in a row (same block) as shown in Figure 3. N_b and N_c are the height and width of the 2-D array in Figure 3. Thus N_b is actually the total number of the 8×8 blocks and N_c is the number of the low frequency coefficients we kept in the 2-D array, i.e. $N_c = L$.

$$\delta(x) = \begin{cases} 1 & if \quad x = true \\ 0 & otherwise \end{cases}$$

Since the magnitude of the block-DCT coefficients has a quite large range, in order to reduce the whole feature dimensionality, we resort to a thresholding technique. That is, we select a threshold T to clip the element values in the 2-D array in Figure 3. In other words, If an element whose value is larger than T, it will be represented by T. Therefore, $m, n \in \{0, 1,\ldots, T\}$ and the dimensionality of the transition matrix will be reduced to $(T+1)\times(T+1)$. ***e)*** To capture the dependencies between inter-block coefficients in each column, we calculate the vertical Markov empirical transition matrix of the 2-D array in Figure 3. Its element is given by:

$$p_v(|C_{i+1,j}| = n \mid |C_{i,j}| = m) = \frac{\sum_{i=1}^{N_b-1}\sum_{j=1}^{N_c-1}\delta(|C_{i,j}| = m, |C_{i+1,j}| = n)}{\sum_{i=1}^{N_b-1}\sum_{j=1}^{N_c-1}\delta(|C_{i,j}| = m)}$$

where $|C_{i,j}|$ and $|C_{i+1,j}|$are absolute neighboring DCT coefficients pair in a column (same frequency or mode) as shown in Figure 3. The other parameters have same definitions as in the formula of Step d. Similar to Step d, we set threshold when calculate the transition matrix to reduce the feature dimensionality.

Setting a threshold in Steps d and e does not lose much information because the DCT coefficients follow the generalized Laplacian distribution which has a very large peak around zero. Therefore, most DCT coefficients are small. As an example, we calculate the percentage of the absolute DCT coefficients which are below a given threshold T for the popular Lena image with Q-factor 80. The results are shown in Table I. As seen, when we set the threshold *T=9, 97.07%* coefficients fall into the range we are considering. Therefore, using an appropriate threshold just loss trivial information

while reduce the computational complexity dramatically. T=9 is used in all of our experiments.

TABLE I. DISTRIBUTION OF ABSOLUTE COEFFICIENTS FOR LENA IMAGE

T	6	7	8	9
Below *T*	96.26%	96.59%	96.85%	97.07%

III. EXPERIMENTS

Once the proposed features have been generated, we need to evaluate the effectiveness of these features. In our experiments, we adopt support vector machine (SVM) as the classifier. The second order polynomial kernel is used in the experiments. The Matlab SVM code is obtained from LIBSVM [10].

Our test image dataset consists of 7560 JPEG images with quality factors ranging from 70 to 90. Among them, 2500 images were taken by members of our research group in different places at different time with different digital cameras; the other 5060 images were downloaded from the Internet. Each image was cropped (central portion) to the dimension of either 768×512 or 512×768. In this way, a reasonably wide diversity of the image database is achieved.

TABLE II.PERFORMANCE COMPARISON USING SVM CLASSIFICATION WITH POLYNOMIAL KERNEL (IN THE UNIT OF %;OG STANDS FOR OUTGUESS. TN STANDS FOR TRUE NEGATIVE RATE, TP FOR TRUE POSITIVE RATE, AND AR FOR DETECTION ACCURACY,AR=(TN+TP)/2)

	bpc	**MW [4]**			**MC [5]**			**JF [6]**			**Proposed**		
		TN	**TP**	**AR**	**TN**	**TP**	**AR**	**TN**	**TP**	**AR**	**TN**	**TP**	**AR**
OG	0.05	59.0	57.6	58.3	55.6	58.5	57.0	49.8	75.4	62.6	72.9	74.2	73.5
OG	0.1	70.0	63.5	66.8	61.4	66.3	63.9	68.9	83.3	76.1	86.6	89.3	87.9
OG	0.2	81.9	75.3	78.6	72.4	77.5	75.0	90.0	93.6	91.8	96.2	96.2	96.2
F5	0.05	55.6	45.9	50.8	57.9	45.0	51.5	46.1	61.0	53.6	56.6	57.0	56.8
F5	0.1	55.5	48.4	52.0	54.6	54.6	54.6	58.4	63.3	60.8	66.2	67.9	67.1
F5	0.2	55.7	55.3	55.5	59.5	63.3	61.4	77.4	77.2	77.3	83.5	85.7	84.6
F5	0.4	62.7	65.0	63.9	71.5	77.1	74.3	92.6	93.0	92.8	96.0	97.3	96.7
MB1	0.05	48.5	53.2	50.8	57.0	49.2	53.1	39.7	66.9	53.3	72.7	75.1	73.9
MB1	0.1	51.9	52.3	52.1	57.6	56.6	57.1	45.6	70.1	57.9	87.8	88.9	88.4
MB1	0.2	52.3	56.7	54.5	63.2	66.7	65.0	58.3	77.5	67.9	97.1	97.0	97.1
MB1	0.4	55.3	63.6	59.4	74.2	80.0	77.1	82.9	86.8	84.8	99.5	99.7	99.6

For each image in the image database, we have prepared stego-images generated by each of the JPEG steganographic techniques discussed in previous sections. The embedded message length is represented by bpc (bits per non-zero DCT AC coefficients). In the preparation of our test dataset, we notice that OutGuess and F5 re-compress the JPEG image before the embedding process. In order to minimize the influence of the JPEG re-compression, we use the re-compressed images without data embedding as the cover image dataset and the re-compressed images with data embedding as the stego-image dataset for classification. In this way, we guarantee the differences between the cover image and the stego-image are solely caused by the steganography itself.

In the classification process, we randomly selected half of original images and the corresponding half of stego-images for training and the remaining half pairs of the cover images and stego-images for testing the trained classifier. For comparison reason, we have also implemented the steganalysis schemes proposed by Farid [4], Shi et al. [5] and Fridrich [6] (denoted by MW, MC and JF, respectively). Then we apply them to the same set of images and the same steganographic methods. The same training and testing procedures are used. All the results are listed in Table II. The experimental results reported here are the averages of the 20 times of tests. As in Table II, the proposed steganalysis scheme outperforms the existing methods by a significant margin under all embedding bit rates for all steganographic techniques, thus demonstrating the effectiveness of the proposed second order statistics features.

IV. CONCLUSIONS

We have presented an effective steganalysis scheme in attacking today's sophisticated JPEG steganographic schemes. A second order statistics, Markov empirical transition matrix, has been adopted to capture the changes of the dependency between block-DCT coefficients in both intrablock and interblock sense. The experimental results have demonstrated that the proposed scheme outperforms the state-of-the-arts in detecting the modern steganographic methods for JPEG images: OutGuess, F5, and MB1.

REFERENCES

[1] N. Provos, "Defending against statistical steganalysis," 10th USENIX Security Symposium, Washington DC, USA, 2001.

[2] A. Westfeld, "F5 a steganographic algorithm: High capacity despite better steganalysis," 4th International Workshop on Information Hiding, Pittsburgh, PA, USA, 2001.

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

[4] H. Farid, "Detecting hidden messages using higher-order statistical models", International Conference on Image Processing, Rochester, NY, USA, 2002.

[5] Y. Q. Shi, G. Xuan, D. Zou, J. Gao, C. Yang, Z. Zhang, P. Chai, W. Chen, C. Chen, "Steganalysis based on moments of characteristic functions using wavelet decomposition, prediction-error image, and neural network," International Conference on Multimedia and Expo, Amsterdam, Netherlands, 2005.

[6] J. Fridrich, "Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes," 6th Information Hiding Workshop, Toronto, ON, Canada, 2004.

[7] K. Sullivan, U. Madhow, S. Chandrasekaran, and B. S. Manjunath, "Steganalysis of Spread Spectrum Data Hiding Exploiting Cover Memory", the International Society for Optical Engineering, Electronic Imaging, San Jose, CA, USA, 2005.

[8] R. Bohme and A. Westfeld, "Breaking Cauchy model-based JPEG steganography with first order statistics", ESORICS 2004, LNCS 3193, pp. 125-140, 2004.

[9] C. Tu and T. D. Tran, "Context based entropy coding of block transform coefficients for image compression", IEEE Transaction on Image Processing , vol. 11, No. 11, November 2002.

[10] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines", 2001. (http://www.csie.ntu.edu.tw/~cjlin/libsvm)