

JPEG STEGANALYSIS BASED ON CLASSWISE NON-PRINCIPAL COMPONENTS ANALYSIS AND MULTI-DIRECTIONAL MARKOV MODEL

Guorong Xuan¹, Xia Cui¹, Yun Q. Shi², Wen Chen², Xuefeng Tong¹, Cong Huang¹

¹Dept. of Computer Science, Tongji University, Shanghai, China

²Dept. of Electrical and Computer Engineering, New Jersey Institute of Technology
Newark, New Jersey, USA

grxuan@public1.sta.net.cn, shi@njit.edu

ABSTRACT

This paper¹ presents a new steganalysis scheme to attack JPEG steganography. The 360 dimensional feature vectors sensitive to data embedding process are derived from multidirectional Markov models in the JPEG coefficients domain. The class-wise non-principal components analysis (CNPCA) is proposed to classify steganography in the high-dimensional feature vector space. The experimental results have demonstrated that the proposed scheme outperforms the existing steganalysis techniques in attacking modern JPEG steganographic schemes – F5, Outguess, MB1 and MB2.

1. INTRODUCTION

Steganalysis is the art and science to detect whether a given medium has hidden message in it. It is the counterpart of steganography. The steganalysis techniques generally can be classified into two categories: the specific steganalysis is designed to attack a particular steganographic scheme, and the universal steganalysis aims at detecting any existing steganographic techniques. In this paper, we present a new universal steganalyzer under pattern recognition framework.

There are several universal steganalyzers in the prior arts. Farid et al. [1] proposed a general steganalysis method based on image high order statistics derived from high-frequency wavelet subbands. In [2], Xuan et al. used statistical moments of characteristic functions of the test image, and all of its wavelet subbands as distinguishing features. The above steganalysis schemes are both based on the statistics related to the histogram of wavelet subbands. Histogram itself is known as the first order statistics which are not very effective in attacking modern JPEG steganography which makes efforts to keep the histogram remaining unchanged. In [3], Fridrich used some 2nd order

statistics such as co-occurrence in the JPEG coefficient domain as features, resulting in improved performance. Although there are only 23 features, the feature extraction is, however, computationally complex. In [4], Sullivan et al. for the first time introduced Markov chain into steganalyzer design. Although it is successful to some extent for the detection of spread spectrum data hiding, it does not perform well for attacking JPEG steganographic methods.

In this paper, we proposed a new universal steganalysis to further improve the performance of attacking JPEG steganography. This proposed approach extracts high-dimensional feature vectors from the probability transition matrixes associated with Markov models (referred to as Markov transition matrixes, even simply transition matrix for short in the rest of this paper), which are constructed from JPEG coefficients scanned along different directions. The class-wise non-principal component analysis (CNPCA) is employed in the classification [5]. The experiment conducted on the 1096 CorelDraw images [6] demonstrated its superior performance over all of the above-mentioned prior arts in attacking the most widely used JPEG steganographic techniques – Outguess, F5, MB1 and MB2.

The rest of this paper is organized as follows. In the next section, we describe the feature extraction from Markov models. Section 3 introduces the CNPCA classification method. In Section 4, the experimental results are given. Finally the conclusions are drawn in Section 5.

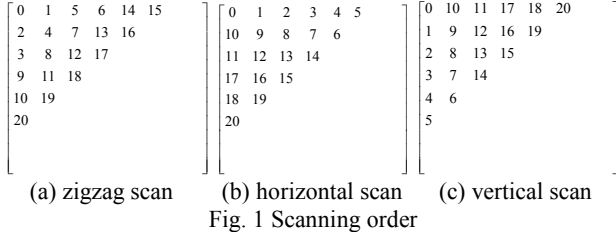
2. FEATURE EXTRACTION

From a given JPEG image, we can form a 2-D array consisting of all of JPEG quantized DCT coefficients (referred to as JPEG coefficients in this paper) from all of 8x8 blocks. Features extracted from these JPEG coefficients will enhance steganalysis ability. In this paper, different orders in scanning the JPEG coefficients are used, and resulting in different coefficient sequences, from which the correlation between the neighboring coefficients is exploited to construct respective Markov transition matrixes. Features for steganalysis are derived from these Markov transition matrixes.

¹ This research is supported partly by National Natural Science Foundation of China (NSFC) on the project (90304017).

2.1. Markov transition matrix

We model the 2-D JPEG coefficient array by using Markov model. According to the theory of random process, the Markov transition matrix can be used to characterize the Markov model. Three different scanning orders - zigzag, horizontal and vertical - are shown in Fig. 1(a), (b) and (c), respectively, where the numbers 0, 1, ..., 20 represent the sequence of the low-frequency coefficients. Unlike single direction scanning, multi-direction scanning can more effectively catch the change and thus provides better performance which is proven by the experimental study.



In the JPEG 8 x 8 DCT blocks, most of high frequency coefficients after quantization are zero, whereas low-frequency AC coefficients are often non-zero and utilized by JPEG steganography. Therefore, only low frequency coefficients are scanned to generate three coefficient sequences, each consisting of the DC coefficient and the first twenty low-frequency AC coefficients. The correlation between DC and its neighboring AC coefficients may reveal the possible manipulation of this AC coefficient. Therefore DC coefficients are included although JPEG steganography never touches DC coefficient. From these three coefficient sequences, the one-step Markov transition matrixes can be constructed. Specifically, the elements of transition matrixes reflect the transition probability between two immediately neighboring elements in the coefficient sequence. The $n-1$ ($n > 2$) transition matrixes can be constructed if we consider the transition between two elements separated by $(n-1)$ elements. In this work, we only consider the one-step transition in order to achieve a balance between good performance and affordable computational complexity.

Since the dynamic range of JPEG coefficient is large, the dimension of transition matrix is non-trivial. In order to reduce computational complexity, we propose to threshold the elements in the coefficient sequence. To do this, we select a threshold value T . If the value of an element is larger than T or smaller than $-T$, it will be represented by T or $-T$ accordingly. The threshold ensures that the values of all elements in the coefficient sequence fall into the range $\{-T, -T+1, \dots, -1, 0, 1, \dots, T-1, T\}$. The dimensionality of transition matrix will be $(2T+1) \times (2T+1)$. In this work, the threshold is selected to be 7, that is justified below.

Suppose that the total number of DCT blocks is n for a given image, with the scan ordering $i = 1, 2, 3$, the global

Markov transition matrix G_i can be obtained by averaging Markov transition matrix of each DCT block k , that is

$$G_i = \frac{1}{n} \sum_{k=1}^n G_{ik} \quad (1)$$

Considering the matrixes are symmetric since bidirectional Markov model is used here, we only adopt the elements of the upper triangle of the matrix, as shown in Fig. 2, to construct the feature vector. There are 120 (i.e. $15 \times 15 + 15$)/2 elements in this triangle. Therefore, we have 120 features extracted from each Markov transition matrix G_i . Totally, we have 360 (i.e., 3×120) features for a test image.

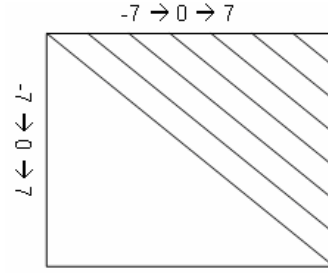


Fig. 2: Markov transition matrix (15×15)

2.2. Threshold selection

The threshold is a key parameter in our proposed scheme. The threshold should be chosen such that the information loss is minimized and the computational complexity is manageable. Most of AC coefficients have small values around the zero and is approximately Laplacian distributed. According to the statistics, we use threshold value of 7 in the experiment. Table 1 shows the percentage of JPEG coefficients falling into $[-T, T]$ for $T = 5, 6, 7, 8, 9$. It is calculated from a typical image, #16053, in CorelDraw image dataset. It is seen that most JPEG coefficients are falling into the selected threshold arrange, indicating that the information loss is negligible for these threshold values.

Table I: The percentage of AC coefficients in $[-T, T]$

$[-T, T]$	$[-5, 5]$	$[-6, 6]$	$[-7, 7]$	$[-8, 8]$	$[-9, 9]$
Percentage	97.3	98.1	98.6	98.9	99.2

3. CLASS-WISE NON-PRINCIPAL COMPONENT ANALYSIS (CNPCA)

The class-wise non-principal component analysis (CNPCA) [5] is to classify the samples based on the distances between the samples and the mean vectors of each class in the space spanned by the eigenvectors associated with the smallest eigenvalues of each class.

Let x denote the n -dimensional random vectors in the k th class, and assume that there are in total K different classes. When the eigenvalues of the covariance matrix generated from all of x are ranked from the largest to the

smallest in a non-increasing order, the corresponding eigenvector matrix can be expressed as:

$$\Phi_k = (\Phi_k)_{n \times n} = [\Phi_{rk}, \Psi_{rk}]_{n \times n} \quad (2)$$

where the n is the dimensionality; the r , ($r \leq n$), is the number of eigenvectors associated with the largest eigenvalues; the $(n-r)$ is the number of eigenvectors associated with the smallest eigenvalues; the $\Phi_k = (\Phi_k)_{n \times n}$ is the eigenvector matrix with all eigenvectors of the k^{th} class; the $\Phi_{rk} = (\Phi_{rk})_{n \times r}$ is the principal components matrix with all the r eigenvectors of the k^{th} class; the $\Psi_{rk} = (\Psi_{rk})_{n \times (n-r)}$ is the non-principal components matrix with all, $(n-r)$, remaining eigenvectors of the k^{th} class; the k^{th} class' non-principal components Ψ_{rk} and principal components Φ_{rk} are complementary to each other.

In CNPCA classification, given a test sample vector y , its Euclidean distance to the mean vector of the k^{th} class in the subspace spanned by the $(n-r)$ class non-principal components is adopted as the classification criterion, referred to as CNPCA distance. The CNPCA distance of the vector y to the k^{th} class is defined as:

$$D_{rk} = \|\Psi_{rk}^T (y - M_k)\| \quad (3)$$

where D_{rk} stands for the Euclidean distances between the sample y and the mean of the k^{th} class, M_k , in the $(n-r)$ dimensional CNPCA space, D_{rk} can be represented by the class-wise non-principal components matrix Ψ_{rk} . Obviously, there are two special cases. When $r=0$, CNPCA distance becomes the conventional Euclidean distance while when $r=n$, CNPCA distance equals to 0. Hence the case of $r>0$ and $r<n$ is usually used in CNPCA.

4. EXPERIMENTS

4.1. Classification performance

To evaluate the performance of the proposed approach, the 1096 BMP images with size of 768x512 in the CorelDraw [6] were used. To generate cover images, we compressed the BMP images with quality factor 75 by IJG JPEG compressor [7]. With the same quality factor, the stego-images are generated by implementing F5 [8], Outguess [9], MB1, and MB2 [10] with the BMP images. The length of embedded message is 1KB (1024 bytes), 2KB and 4KB, which correspond to 0.021bpp, 0.041bpp, and 0.083bpp, respectively, in our case. This procedure ensures that the difference between a stego image and its corresponding cover image is only caused by data embedding, avoiding the influence of JPEG double compression. For F5, MB1 and MB2, we generated 1096 stego images of which 896 were randomly selected for training and the remaining 200 for testing. For Outguess, the number of stego images are 1071,

1001 and 630 when we embed 1KB, 2KB and 4 KB, respectively. We randomly selected 200 for testing and the rest for testing. The experiment was run 10 times.

Table II: Detection accuracy

Payload		Xuan et al. [2]			Fridrich [3]			Proposed		
		tn	tp	t	tn	tp	t	tn	tp	t
F5	1KB	62	58	60	76	72	74	80	81	81
	2KB	64	65	65	86	87	87	91	92	92
	4KB	85	77	81	94	98	96	99	98	99
OG	1KB	61	41	51	91	87	89	98	99	99
	2KB	67	61	64	97	96	97	100	100	100
	4KB	73	79	76	98	97	98	100	100	100
MB1	1KB	59	60	59	66	65	66	91	90	91
	2KB	71	69	70	88	83	86	97	97	97
	4KB	83	81	82	91	88	90	99	99	99
MB2	1KB	65	55	60	64	61	62	94	94	94
	2KB	75	64	69	76	77	76	99	98	99
	4KB	85	83	84	88	80	84	98	99	99

(tn: true negative; tp: true positive; t = (tn+tp)/2)

4.2. Selection of dimensionality r

Table II illustrates the performance comparisons between the proposed scheme and the methods proposed by [2, 3]. For F5, $r=12$; for Outguess, MB1 and MB2, $r=29$. The detection rates shown in the table are the average of the 10 time runs. From this table, it can be seen that our proposed method shows the significantly improved classification performance compared with the prior arts [2,3].

The detection accuracy t is a function of the non-principal components dimensionality $(n-r)$. Since the feature number n is fixed, t actually varies with r as shown in Fig.3 and Fig.4. The experiments have showed that the best detection accuracy is achieved when r takes on 29 for MB1, MB2 and Outguess, and 12 for F5.

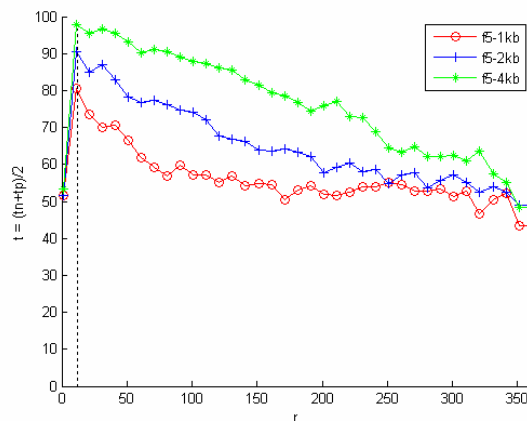


Fig.3: Detection rate t vs. dimensionality r of F5

5. CONCLUSIONS

In this paper, we proposed to extract features in JPEG coefficients domain. Since modern JPEG steganographic schemes such as Outguess, F5, MB1 and MB2 directly modify JPEG DCT coefficients to embed the secret message, the features are more efficient if they are extracted from the JPEG coefficients domain instead of from spatial domain. The experimental results in Table II have demonstrated that the detection accuracy of the proposed approach is 15%-30% higher than [2] which extracted features in spatial domain.

The coefficients are arranged in three different scanning sequential orders: the zigzag, horizontal and vertical orders, to fully exploit the correlation among neighboring JPEG coefficients.

The recently developed CNPCA is utilized in the classification. It has demonstrated good classification performance and computational efficiency.

The experiments have shown that the proposed approach outperforms the previous arts [2, 3, 11] in attacking JPEG steganography of F5, Outguess, MB1 and MB2 by a significant margin.

REFERENCES

- [1] S. Lyu and H. Farid, "Detecting hidden message using higher-order statistical models," *In: Proc. of the IEEE Int'l Conf. on Image Processing*, Vol II 905-908, New York, 2002.
- [2] G. Xuan, Y. Q. Shi, J. Gao, D. Zou, C. Yang, Z. Zhang, P. Chai, C. Chen, W. Chen, "Steganalysis based on multiple features formed by statistical moments of wavelet characteristic functions," *Information Hiding Workshop (IH2005)*, Barcelona, Spain, June 2005.
- [3] 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
- [4] K. Sullivan, U. Madhow, S. Chandrasekaran, and B. S. Manjunath, "Steganalysis of spread spectrum data hiding exploiting cover memory," *SPIE2005*, vol. 5681, pp38-46.
- [5] G. Xuan, P. Chai, Y. Q. Shi, X. Zhu, Q. Yao, C. Huang, D. Fu., "A novel pattern classification scheme: classwise non-principal component analysis (CNPCA)," *International Conference on Pattern Recognition (ICPR) 2006*.
- [6] <http://www.corel.com>
- [7] <http://www.ijg.org>
- [8] <http://www.rn.inf.tu-dresden.de/~westfeld/F5.html>
- [9] <http://www.outguess.org>
- [10] <http://redwood.ucdavis.edu/phil/papers/iwdw03.htm>
- [11] G. Xuan et al., "Steganalysis Using high-dimensional features derived from co-occurrence matrix and class-wise non-principal components analysis (CNPCA)," *International Workshop on Digital Watermarking (IWDW)*, November 2006, Jeju, Korea.

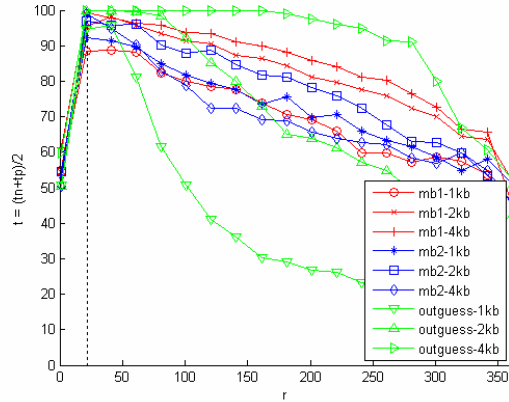


Fig.4: Detection rate t vs. dimensionality r of MB and Outguess

4.3. Stability study

To verify the stability of the detection accuracy, the above-mentioned test is carried out 10 times. The detection accuracy is quite stable as shown in Fig.5, where the number of horizontal axis is for different embedding schemes, and the vertical axis represents the detection accuracy.

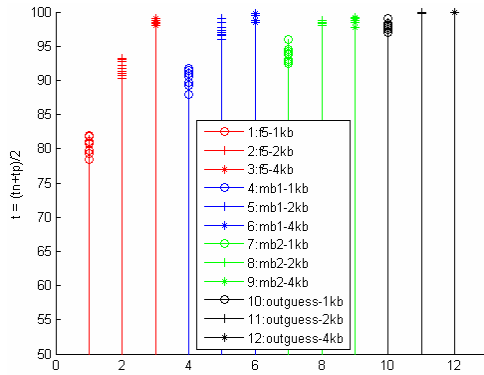


Fig.5: The stability performance

4.4. Computational complexity

Support Vector Machine (SVM) is a popular classification tool. In SVM classification, the features in the input space are projected into linearly separable higher (maybe infinite) dimensional space in which a separating hyperplane is found such that the margin is maximized. In our experiment, the proposed CNPCA can achieve the same classification performance as the SVM, while use less CPU time, as seen in Table III.

Table III: Comparison of computational complexity

	Image number	Test samples	time
CNPCA	1096	200	18.7s
SVM	1096	200	40.4s