DETECTION OF WATERMARKING METHODS USING STEGANALYSIS

Gokhan GUL[1], Fatih KURUGOLLU[2]

[1]Faculty of Engineering, University of Kiel, Kiel, Germany

[2]School of Electronics, Electrical Eng. and Computer Science, Queen's University Belfast, UK

ABSTRACT

This paper presents a Singular Value Decomposition (SVD) based steganalysis technique to determine the watermarking method used to embed a watermark in an image. The detection is carried out in three steps. First, the proposed technique determines whether an image under consideration contains a watermark. If a watermark is detected, the embedding domain is revealed. Finally, the exact watermarking algorithm is named. The idea behind the method is that when the image is watermarked, relative and strict linear dependencies of rows/columns will differ from the original image and this can be modeled by the analysis of SVD. By using SVD, several features for the classification of the original and watermarked images are defined. The classification operation including both a linear and a SVM classifier is performed with a feature selection algorithm, which serves to reduce the number of features and to increase the detection performance. The performance of the proposed technique is promising and simulation results indicate that the chosen features can reliably detect the watermarking domain as well as the watermarking method.

Index Terms — Watermarking, Steganalysis, Classification

1. INTRODUCTION

Watermarking is an emerging technology which aims to hide some message into the digital media to protect the copyright ownership or to authenticate the content [1]. The main requirement of watermarking is to resist the attacks targeting to destroy the watermark or the watermarking protocol. Watermarking attacks can be divided into four main categories: removal attacks, geometric attacks, cryptographic attacks and protocol attacks [2]. The aim of removal attacks is to expel the watermark from the image completely. Denoising, lossy compression, quantization, remodulation, collusion and averaging are considered in this category. Geometric attacks intend to invalidate the synchronization between the watermark embedder and the watermark detector through spatial and temporal alterations of watermarked data. Cryptographic attacks target to find the secret key used in embedding by means of brute force searches. Another attack in this context, called Oracle attack, can be devised to find a version of the watermarked data, which cannot be detected by the watermark detector if the watermark detector is available to the adversary. Finally, the goal of the protocol attacks is to create confusion in the concept of the watermarking application. Copy attack is considered in this category. It predicts the watermark message from a watermarked media and copies this message into a target asset that is not supposed to be watermarked. Therefore, the watermark detector will give a false positive error in the target asset by indicating the existence of the watermark. Then no one can claim a detected watermark really embedded into the data under consideration. The details of these attacks can be found in [2]. Among these attacks, cryptographic and protocol based ones are interesting since they involve extracting the watermark message from the embedded data. In this process, if the watermarking algorithm or at least the embedding domain is known one can design a targeted attack to reveal the watermark message. Therefore, it is of great interest to reveal the embedding algorithm or the embedding domain from the watermarked data.

In this work, we address this problem and present a Singular Value Decomposition (SVD) based steganalysis technique to determine the watermarking algorithm used to watermark the image. The rest of the paper is organized as follows. In the next section, SVD is summarized and the proof of a preposition declared the relationship between number of linearly dependent rows, number of linearly dependent columns and number of zeros in the singular value vector of a square matrix. The features used to discriminate the embedding domains and the methods are explained in Section 3. The experiments and the results are presented in Section 4. Finally, the conclusions are drawn in Section 5.

2. SINGULAR VALUE DECOMPOSITION

SVD is an important tool to factorize matrices leading several applications in image processing. A matrix, A, can be expressed with the product of two orthonormal matrices, U and V, and a diagonal matrix, S ($A = USV^T$). The diagonal elements of matrix S, called singular values, provide characteristic of the given matrix. These elements are organized into a vector, called singular value vector Sv.

Proposition: For a given square matrix A, if the number of linearly dependent rows is i, the number of linearly dependent columns is j and the number of zeros in the singular value vector of A is k then;

$$k = \max(i, j) \quad (1)$$

Proof: Due to the definition of singular value decomposition U and V are the orthonormal matrices and of

full rank. Thus, the rank of matrix A equals to the number of non-zero elements of the diagonal matrix S (Eq. 2). The rank of a matrix is the minimum number of linearly independent rows or columns (Eq. 3). Substituting the Eq. 2 in the Eq. 3, for the Eq. 1 to hold, it is sufficient to show that the Eq. 4 holds. From the given proposition, we know that $m = n$ and thus the Eq. 4 can be reduced to the Eq. 5. For the Eq. 5, there are two conditions; a- $m \geq i \geq j$ and b- $m \geq j \geq i$. Since the Eq. 5 holds for both conditions, the given proposal holds as well.

$$R = \min(m,n) - k \quad (2)$$

$$R = \min(m-i, n-j) \quad (3)$$

$$\max(i,j) = \min(m,n) - \min(m-i, n-j) \quad (4)$$

$$m = \max(i,j) + \min(m-i, m-j) \quad (5)$$

3. SVD AND WAVELET BASED FEATURES

3.1. SVD based features

In order to obtain Sv based vectors, images are first divided into sub-blocks of sizes W×W (W=3, 4,…20). Singular values, obtained by applying SVD to each block, are then normalized with the sum of singular values to reduce the effects of different energy levels in different images. Image blocks are overlapped proportionally to the block size to be able to take into account the correlations within and among the image blocks. Therefore, from W=3 to 12 no overlapping, from W=13 to 15, 50% overlapping and from W=16 to 20, 75% overlapping strategy is accepted. SVD based features are defined as follows:

Features of Type-1: These features are the means of the number of zeros at index i in Sv vectors of sub-blocks. Let B be the integer number of W×W sized sub-blocks according to the overlapping strategy. Type-1 features are defined as:

$$f_W^{(1)}(i) = \frac{1}{B}\sum\nolimits_B \delta(Sv(i)), \quad W=3,\ldots,20 \text{ and } i = 1,\ldots,W \quad (6)$$

where $\delta(k) = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$

Features of Type-2: These features are the means of the singular values at index i in Sv vectors of sub-blocks.

$$f_W^{(2)}(i) = \frac{1}{B}\sum\nolimits_B Sv(i), \quad W=3,\ldots,20 \text{ and } i = 1,\ldots,W. \quad (7)$$

Features of Type-3: Type-3 features are defined as the mean and the variance of type-1 and type-2 features varying the window size. This type of features can be given as in the following equations.

$$f_{mean}^{(3,n)}(i) = \begin{cases} \dfrac{1}{N-2}\displaystyle\sum_{w=3}^{N} f_w^n(i) & \text{if} \quad n+1 \geq i \geq n \\ \dfrac{1}{N-i+1}\displaystyle\sum_{w=i}^{N} f_w^n(w-i+4) & \text{if} \quad N-1 \geq i \geq n+2 \end{cases} \quad (8)$$

$$f_{var}^{(3,n)}(i) = \begin{cases} \dfrac{1}{N-2}\displaystyle\sum_{w=3}^{N} (f_w^n(i)-\mu)^2 & \text{if} \quad n+1 \geq i \geq n \\ \dfrac{1}{N-i+1}\displaystyle\sum_{w=i}^{N} (f_w^n(w-i+4)-\mu)^2 & \text{if} \quad N-1 \geq i \geq n+2 \end{cases} \quad (9)$$

where n=1,2 indicates the feature type, μ is the mean of non-normalized Type-1 and Type-2 features by the number of windows and N is set to 20 as the maximum window size. Type-1 and Type-2 features are the same features treated in [3]. Type-3 features differ only with evaluating the energy and the linear dependency components separately. The basis of Type-1 features depends on the Eq. 1. According to the proposition, overall strict linear dependency of a given matrix can be determined by the sum of the number of "0"s at a fixed index of Sv over sub-blocks. Additionally, by adding the singular values at a certain index up overall relative linear dependency of a matrix can be modeled by the use of Type-2 features.

3.2. Higher order statistical features

Lyu and Farid [4] proposed a very effective steganalysis method based on higher order statistical (HOS) features. These features have also been used effectively in computer generated image detection [5] and in blind identification of cellular cameras [6]. Wavelet decomposition employed by quadrature mirror filters (QMFs) which decomposes the frequency space into multiple scales and orientations. Given this image decomposition, the statistical model is composed of the mean, variance, skewness and kurtosis of the sub-band coefficients at each orientation and each scale. These statistics characterize the basic coefficient distributions. The second set of statistics is based on the errors in an optimal linear predictor of coefficient magnitude. Thus over three level wavelet decomposition and three different orientations, four different statistics as mean, variance, skewness and kurtosis and two different domains, namely wavelet and optimum linear prediction errors, result in total 72 features. The advantages and disadvantages of the proposed Singular value based method (SVBM) over Lyu and Farid's method (LFM) are compared in the Table 1.

Table 1. Comparison of LFM and SVBM based features

	LFM	SVBM
Various features obtained by	Sub-Bands	Sub-windows
Transformation Coefficients	QMFs	Eigen matrices
Statistics	First four moments	First two moments
Total number of features	72	470
Selected number of features	Less than 10% of image samples	Less than 10% of image samples

4. EXPERIMENTAL RESULTS

4.1. Watermarking methods

In this work, three embedding domains are considered: Spatial (Spt.), DCT and DWT. In each domain, the following algorithms have chosen. Spatial: Bruyndonckx et al. (Bry), Kutter et al. (Kut), Sebe et al. (Seb), Lee et al. DCT: Cox et al., Barni et al. (Bar), Koch et al. (Koc), Hsu and Wu (Hsu). DWT: Kim and Shik (Kim), Wang et al. (Wan), Xia et al., Zhu et al. Details of these algorithms can be found in [7][8][9][10].

4.2. Image set

200 randomly chosen images from the personal collection of P. Greenspun (http://philip.greenspun.com) have been used

in the experiments. Central portions of images are then cropped to 512*512 to conform to the watermarking algorithms. Image set is then subjected to embedding operations for all watermarking algorithms with a fixed watermark message length of 32000 (0.12bpp) except for Bruyndonckx et al., Sebe et al. and Koch et al. algorithms. For these algorithms message length of 1000 is used due to the algorithmic restrictions. Most of watermarking methods have a parameter to provide a trade-off between robustness and imperceptibility. According to [11], a watermarked image whose PSNR value is over 38 dB is acceptable for the imperceptibility requirement. Therefore, for all embedding algorithms we justified the PSNR value of the reference image as 39 dB by adjusting the trade-off parameter.

4.2. Classification and Feature Selection

Classification over two-class cases is performed with a Fisher linear discriminant classifier following the feature selection process with the Sequential Forward Feature Selection (SFFS) algorithm [12]. Nevertheless, the SFFS takes one-way path for the feature selection and can be improved with pre-determining the correct path. For this purpose, we designed a Pre-Feature-Selection (PFS) algorithm, which can be explained briefly as follows:

- Find the ANOVA statistics of the features and order the features in ascending order respected to p-values, which indicate that, the probability of finding in reality that there is no difference between the means [13].
- Choose from the set of N features the pair of features yielding the best classification result starting with the one having the lowest p value.
- Add the most significant feature from the remaining ones until there is no performance increase.
- Do the first two steps starting with the next feature having higher p-value till the user defined number of feature vectors is not exceeded.
- Feed the SFFS with the first feature of the most significant feature vector.
- Determine the feature vector between two feature vectors which gives the optimum performance as the selected feature set.

User defined value is set to 30 and the maximum number of features in the feature vector is set to 40. Because according to [14] the number of features should be less than 10% of the training image set when the feature selection is performed in order to claim that classifier generalizes but not memorizes. For multi-class classification, SVM classifier is adopted. Optimum SVM classifier parameters are determined before the classification and feature selection operations for SVBM and LFM. We note that only those features selected, which correspond to at most 10% of the number of images in the training phase of the related watermarking algorithm, are considered for all experiments.

4.3. Experiments

In the first step, the linear classifier has been trained with the features derived from 100 original and 100 watermarked images and it has been tested on the images from the rest of the set. Table 2 shows the comparative results for Spatial, DCT and DWT domain watermarking algorithms. Detection performance of SVBM is higher than that of LFM for Spatial and DCT domain in general. For DWT domain, LFM performed higher detection performances than SVBM since it directly uses the features derived from DWT coefficients.

To realize the second stage, namely multi-class classification, we adopted a two-level hierarchical model in the classification. Let $P_D(C_i)$ denotes the detection probability of C_i^{th} class and $P_D(\text{domain})$ indicates the detection probability of the domain where the watermarking operation is performed, we can express this probability by means of conditional probabilities as $P_D(C_i) = P_D(C_i \mid \text{domain}_j) P_D(\text{domain}_j)$. Spatial, DCT and DWT domain data sets are formed with 50 images from each watermarking algorithm within the related domain. Then the images in the data sets are permuted in order to train and test the SVM classifier over all four algorithms for each domain. Again half of the images are used for the training and the rest for the testing. Table 3 shows how reliably SVBM and LFM can detect the embedding domain. Diagonal of the given two confusion matrices indicates each $P_D(\text{domain}_j)$. Average detection performances are 81.25% and 72.75% for SVBM and LFM, respectively.

Given the embedding domain, detection performance of individual watermarking algorithms is depicted in the Table 4. Overall performance of SVBM and LFM are 94.75% and 96% for spatial domain algorithms, respectively, while both are 96.75% for DCT domain ones. For the DWT domain methods performance of SVBM and LFM are 61% and 76%.

Based on both $P_D(C_i \mid \text{domain}_j)$ and $P_D(\text{domain}_j)$ detection probabilities of both classifiers overall detection performances of SVBM and LFM are given in Table5. Overall SVBM outperforms LFM on Spatial and DCT domain watermarking algorithms. However, LFM performs better for wavelet domain algorithms. In the last stage, to improve the detection results both approaches are merged assuming that merged classifier is informed with the detection accuracies of SVBM and LFM obtained in the pervious stages. This increases overall performance as depicted in Table 5. The merged classifier have achieved more than 80% detection accuracy for Spatial and DCT based methods while its performance is limited to about 55% in DWT based ones. Therefore, it can be concluded that DWT based methods are more robust against the steganalysis and provide more secure watermarking against targeted attacks to reveal the watermark message

5. CONCLUSION

In this paper we have proposed a SVD based method in order to determine the method used to watermark the images. Experimental results showed that proposed method not only can reliably detect the presence of a watermark but also can determine the watermarking domain and the watermarking method. For the watermark domain detection, performance of the proposed method is better than that of LFM for Spatial and DCT domain watermarking algorithms, while LFM provides slightly better performance for DWT domain methods. To determine the particular watermarking method as well as the embedding domain, a multi-class SVM classifier using two-level hierarchical model is devised. In this classifier the results obtained from SVBM and LFM are merged to improve the detection rate. Then this merged method has achieved more than 80% detection accuracy for Spatial and DCT based methods while its performance is about 55% for DWT ones. This shows that DWT based methods are more robust against this kind of steganalysis.

6. REFERENCES

[1] I. Cox, M.Miller, J. Bloom, *Digital Watermarking: Principles and Practice*, Morgan Kaufmann, 2001.

[2] S. Voloshynovskiy et al., "Attacks on digital watermarks: classification, estimation based attacks, and benchmarks", *IEEE Communication Magazine*, vol. 39, no. 8, pp.118-126, 2001.

[3] G. Gul, A.Emir Dirik, I. Avcibas, "Steganalytic Features for JPEG Compression Based Perturbed Quantization", *IEEE Signal Processing Letters*, vol.14, pp. 205-208, 2007.

[4] S. Lyu and H. Farid, "Detecting Hidden Messages Using Higher-Order Statistics and Support Vector Machines," *Lecture Notes in Computer Science*, vol. 2578., pp. 340-354, Springer-Verlag, 2002.

[5] S. Lyu and H. Farid, "How realistic is photorealistic?", *IEEE Transactions on Signal Processing*, vol. 53, , pp. 845–850, 2005.

[6] O. Celiktutan, I. Avcıbas, and B. Sankur, "Blind identification of cellular phone cameras", *Proceedings of SPIE*, vol. 6505, January 2007.

[7] P. Meerwald, *Digital Image Watermarking in the Wavelet Transform Domain,* MSc Thesis, University of Salzburg, Austria, 2001.

[8] O. Bruyndonckx, J.J.Quisquater and B. M. Macq, "Spatial method for copyright labeling of digital images", *IEEE Workshop on Nonlinear Signal and Image Processing*, Thessaloniki, Greece, pp. 456 - 459, 1995.

[9] M. Kutter, F. Jordan and F. Bossen, "Digital Signature of Color Images using Amplitude Modulation", *Proc. SPIE Storage and Retrieval for Image and Video Databases*, vol. 2952, pp. 518-526, 1997

[10] F. Sebe, J. Domingo-Ferrer, and J. Herrera, "Spatial-Domain Image Watermarking Robust against Compression, Filtering, Cropping, and Scaling", *Lecturer Notes in Computer Science*, vol. 1975, pp. 44-53, Springer-Verlag, 2000.

[11] S. Katzenbeisser, and A. P. Petitcolas, *Information Hiding Techniques for Steganography and Digital Watermarking*, Artech House Inc., 2000

[12] Pudil, P., J. Novovicova, and J. Kittler, "Floating search methods in feature selection", *Pattern Recognition Letters*, vol. 15, pp. 1119-1125, 1994.

[13] X. Rencher, Methods of Multivariate Data Analysis, Wiley, New York, 1995.

[14] A.K. Jain, R.P.W. Duin and M. Jianchang, "Statistical pattern recognition: a review," *IEEE Trans. On PAMI*, vol. 22, no 1, pp. 4– 37, 2000.

Table 2. Watermark detection performance according to watermarking methods using Fisher Linear Discriminant Classifier.

		LFM			SVBM		
		False	Miss	Acc.	False	Miss	Acc.
Spt.	Bruyn	1	15	92%	0	0	100%
	Kutter	0	0	100%	0	0	100%
	Sebe	1	0	100%	0	5	98%
	Lee	11	4	93%	5	3	96%
DCT	Cox	14	17	85%	11	1	94%
	Koch	20	12	84%	0	0	100%
	Barni	0	0	100%	0	0	100%
	Hsu	3	5	96%	3	1	96%
DWT	Kim	5	11	92%	23	17	80%
	Wang	31	25	72%	4	2	97%
	Xia	8	11	91%	9	3	94%
	Zhu	6	7	94%	17	10	87%

Table 3. Embedding domain detection performance using SVM classifier

	LFM				SVBM			
	Cov.	Spt.	DCT	DWT	Cov.	Spt.	DCT	DWT
Cov.	72	4	9	15	79	1	5	15
Spt.	10	82	3	5	7	87	4	2
DCT	24	4	67	5	8	3	84	5
DWT	27	1	2	70	19	2	4	75

Table 4. Method detection performance using SVM classifier

Spt.	LFM				SVBM			
	Bry	Kut	Seb	Lee	Bry	Kut	Seb	Lee
Bry	98	0	1	1	95	1	3	1
Kut	0	93	5	2	2	98	0	0
Seb	0	1	98	1	2	3	95	0
Lee	1	1	3	95	6	3	0	91
DCT	LFM				SVBM			
	Cox	Koc	Bar	Hsu	Cox	Koc	Bar	Hsu
Cox	95	4	0	1	94	0	5	1
Koc	1	97	0	2	1	99	0	1
Bar	0	0	100	0	2	0	98	0
Hsu	4	1	0	95	4	0	0	96
DWT	LFM				SVBM			
	Kim	Wan	Xia	Zhu	Kim	Wan	Xia	Zhu
Kim	68	14	9	9	65	4	22	9
Wan	12	75	8	5	5	57	1	37
Xia	10	7	78	5	38	5	49	8
Zhu	7	4	6	83	8	18	1	73

Table 5. Overall Detection Performances of SVBM and LFM using SVM.

		LFM	SVBM	MERGED
	Cover	72%	79%	79%
Spt.	Bruyn.	80.4%	82.7%	85.3%
	Kutter	76.3%	85.3%	85.3%
	Sebe	80.4%	82.7%	85.3%
	Lee	77.9%	79.2%	82.3%
DCT	Cox	63.7%	79.0%	79.8%
	Koch	65.0%	83.2%	83.2%
	Barni	67.0%	82.3%	84.0%
	Hsu	63.7%	80.6%	80.6%
DWT	Kim	47.6%	48.8%	51.0%
	Wang	52.5%	42.8%	56.3%
	Xia	54.6%	36.8%	58.5%
	Zhu	58.1%	54.8%	62.3%