Statistical Correlations and Machine Learning for Steganalysis

Qingzhong Liu[1], Andrew H. Sung[1], Bernardete M. Ribeiro[2]
[1]Department of Computer Science, New Mexico Tech, U.S.A
[2]Department of Informatics Engineering, University of Coimbra, Portugal
e-mail: [1]{liu, sung}@cs.nmt.edu, [2]bribeiro@dei.uc.pt

Abstract

In this paper, we present a scheme for steganalysis based on statistical correlations and machine learning. In general, digital images are highly correlated in the spatial domain and the wavelet domain; hiding data in images will affect the correlations. Different correlation features are chosen based on ANOVA (analysis of variance) in different steganographic systems. Several machine learning methods are applied to classify the extracted feature vectors. Experimental results indicate that our scheme in detecting the presence of hidden messages in several steganographic systems is highly effective.

1 Introduction

Steganography has recently received much attention due to its applications in copyright protection and covert communications. Unlike traditional watermark methods it does not affect the appearance of the image. With digital images (audios or videos) as carriers, detecting the presence of hidden messages poses significant challenges. Westfeld performed the blind steganalysis on the basis of statistical analysis of PoVs (pair of values). This method, so-called χ^2-statistical analysis [1], gave a successful result to a sequential LSB (Least Significant Bit) embedding steganography. Provos [2] extended this method by re-sampling the test interval and re-pairing values. Fridrich [3] introduced a RS steganalysis which is based on the partition of an image's pixels into three groups: Regular, Singular and Unusable and estimate the possible embedded message length of the LSB steganography. Farid and Lyu [4,5] described an approach to detect hidden messages in images that uses a wavelet-like decomposition to build higher-order statistical models of natural images. Support vector machines are then used to discriminate between untouched and adulterated images. In [6], Avcibas, et al. presented techniques for steganalysis of image based on image quality metrics. In [7], Rate-Distortion Curves are used for steganalysis.

On the side of steganography, Kawaguchi presented BPCS-steganography [8] to obtain a large hiding capacity; Westfeld proposed the F5 algorithm [9], which hides messages in the DCT (Discrete Cosine Transform) domain and can defeat χ^2-statistical analysis. Yu [10] proposed SES (Steganography Evading Statistical analyses) which can stand χ^2-statistical analysis and RS steganalysis. Meanwhile, many ordinary steganography tools can be downloaded from the Internet based on different hiding methods [11-14].

In this paper, we present a scheme for steganalysis based on statistical correlations and machine learning. In general, digital images are highly correlated in the spatial domain and the wavelet domain; hiding data will affect the statistics of images. Based on the correlation features extracted from images, ANOVA (analysis of variance) [15] is applied to choose the good measures and machine learning algorithms are applied to discriminate between untouched and adulterated images.

2 Correlation and Feature Extraction

2.1 Statistical Properties of Images

Several papers [16-19] described the statistical models of natural images such as probability models for images based on Markov Random Field models (MRFs) and the Gaussian Mixture Model (GMM). In general, natural images are highly correlated in adjacent pixels; as shown in the following.

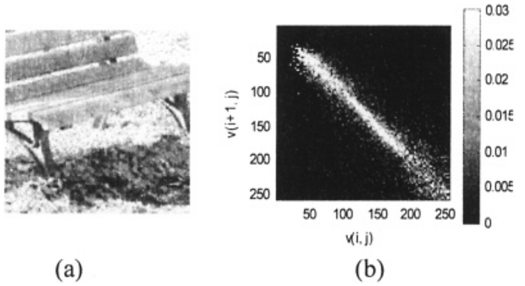

Fig. 1. A grayscale image (a) and the joint probability of adjacent pixels (b)

Figure 1 (a) is a grayscale ([0 255]) image with size $m \times n$. $v(i,j)$ denotes the grayscale value at point (i,j), $v(i+1,j)$ denotes the grayscale value at the adjacent point $(i+1,j)$. $(v(i,j), v(i+1,j))$ is the grayscale value pair at the two adjacent points. Figure 1 (b) shows the joint probability distribution $\rho(v(i,j), v(i+1,j))$, which indicates that the adjacent pixels are highly correlated.

2.2 Statistical Correlations for Feature Extraction

Throughout plenty of experiments, there are also correlations in intra-bit plane and inter-bit plane of images. $M_1(1{:}m,\ 1{:}n)$ denotes the binary bits of the Least Significant Bit Plane and $M_2(1{:}m,\ 1{:}n)$ denotes the binary bits of the second Least Significant Bit Plane. We present the correlation coefficients C1, C2 and C3 as follows:

$$C1 = cor\,(M_1, M_2) \quad (1)$$

$$C2 = cor\,(X_5, X_6) \quad (2)$$

$$C3 = cor\,(X_7, X_8) \quad (3)$$

Where

$$X_5 = M_1(1{:}m-1,\ 1{:}n),\quad X_6 = M_1(2{:}m,\ 1{:}n)$$

$$X_7 = M_1(1{:}m,\ 1{:}n-1),\quad X_8 = M_1(1{:}m,\ 2{:}n)$$

Besides the correlations in the spatial domain, we consider the autocorrelation of the probability density in the histogram. ρ_k denotes the probability density of the histogram at grayscale sample k (k = 0,1, …,N-1, for 8-bit grayscale image, N = 256).

$H = (\rho_0,\ \rho_1,\ \rho_2 \ldots \rho_{N-1})$ stands for the probability distribution in the histogram. H_e, H_o, H_{k1} and H_{k2} are defined as follows:

$$H_e = (\rho_0, \rho_2, \rho_4 \ldots \rho_{N-2})\ ,\qquad H_o = (\rho_1, \rho_3, \rho_5 \ldots \rho_{N-1});$$

$$H_{k1} = (\rho_0, \rho_1, \rho_2 \ldots \rho_{N-1-k}),\quad H_{k2} = (\rho_k, \rho_{k+1}, \rho_{k+2} \ldots \rho_{N-1}).$$

The autocorrelation coefficients C4 and C(k) are defined as follows:

$$C4 = cor\,(H_e, H_o) \quad (4)$$

$$C(k) = cor\,(H_{k1}, H_{k2}) \quad (5)$$

k is the lag distance in (5). Set k = 1, 2,

$$C5 = C(1) \quad (6)$$

$$C6 = C(2) \quad (7)$$

Meanwhile, wavelet decomposition is an analysis of scale- and location-dependence. There are high correlations in intra-subbands. cA, cH, cV and cD denote the approximate sub band, horizontal, vertical and diagonal detail sub bands with size $m' \times n'$, respectively. Define the following autocorrelation coefficients in the wavelet domain.

$$C7 = cor\,(cH_1, cH_2) \quad (8)$$

$$C8 = cor\,(cH_3, cH_4) \quad (9)$$

$$C9 = cor\,(cV_1, cV_2) \quad (10)$$

$$C10 = cor\,(cV_3, cV_4) \quad (11)$$

$$C11 = cor\,(cD_1, cD_2) \quad (12)$$

$$C12 = cor\,(cD_3, cD_4) \quad (13)$$

Where

$$cX_1 = cX(1{:}\ m'-1,\ 1{:}n'),\quad cX_2 = cX(2{:}\ m',\ 1{:}n'),$$

$$cX_3 = cX(1{:}\ m',\ 1{:}n'-1),\quad cX_4 = cX(1{:}\ m',\ 2{:}n'),$$

$$X \in \{H, V, D\}.$$

After extracting C1-C12 from the image, we apply ANOVA [15] to choose the good measures according to the steganographic system.

3 Experiments and Discussion

3.1 Experiments

Over 5000 images are taken from many different sources and cover several categories. Some are downloaded from http://www.freephoto.com and other websites. We store these images as 8-bit grayscales and hide messages in these carriers using the hiding methods SES [10], BPCS [9] and the hiding tools BMP Secrets [14], Invisible Secrets v4 [11] and Secure Engine 4.0 [12]. The corresponding hiding ratios are 12.5%, 25%, 25%, 12% and 12%, respectively. C1-C12 are extracted from the carriers and the steganograms, and ANOVA techniques are then applied to choose the good measures.

STPRtool and LS-SVMlab 1.5 are applied in our experiments. The core of the STPRtool comprises statistical pattern recognition algorithms [20] and the algorithms of LS-SVMlab 1.5 are described in [21]. Training sets are chosen at random and the remaining sets are tested. Classifiers are Fisher Linear Discriminant (FLD), Quadratic Classifier (QC), Support Vector Machines (SVM), Kernel Fisher Discriminant (KFD), and LSSVM [20-24]. RBF kernels are applied and the kernel parameters are 0.01 for SVM and KFD in STPRtool and 0.1 for LSSVM.

3.2 Results and Discussion

Table 1 lists the train accuracy and test accuracy for carriers and steganograms, using the five classifiers. The feature sets are {C1-C5} for SES, Invisible Secrets and Secure Engine; {C1, C4-C6} for BPCS and {C11, C12} for BMP Secrets. Table 1 shows kernel-based classifiers, SVM, KFD and LSSVM have better train accuracy than FLD and QC.

The ROC curves in Figures 2-6 indicate that the classification performance is best in the steganalysis of Invisible Secrets, followed by SES, BPCS and BMP Secrets; the classification performance in the steganalysis of Secure Engine is not as good as others. It is probably attributed to our lack of knowledge regarding its hiding methods, and so the feature set is likely not the best. Figure 6 also indicates that the kernel-based classifiers, KFD and LSSVM are not as good as FLD and QC in steganalysis of Secure Engine, although kernel-based classifiers have a better train accuracy (table 1).

We note that the classification accuracy is related to feature set, kernel parameter, image file format and image type (gray or color). The details are presented in the expanded version of this paper.

Table 1. Accuracy comparison of train and classification using **F**(FLD), **Q**(QC), **S**(SVM), **K**(KFD) and **L**(LSSVM). The first row for each classifier gives train accuracy for carriers; the second row gives train accuracy for steganograms; the third gives test accuracy for carriers and the last row gives test accuracy for steganograms.

	SES	BPCS	BMP Secrets	Invisible Secrets	Secure Engine
F	63%	82.5	87.8	65.8	61
	99	54	84.4	100	87
	58.5	81.6	85.6	66.8	57.4
	98.9	51.7	85.8	100	90.1
Q	69.4	86.9	86.2	73.2	73
	99	70	88.8	100	86
	67.8	85.7	86.3	71.9	64.4
	98.9	63.8	89.5	100	90.1
S	89	100	94.7	98.7	98
	97.6	99.6	94.8	98.9	97
	85.4	69.7	86.1	97.8	80.2
	87.5	96	88	97.3	65.4
K	91.6	100	96.3	98.7	100
	96.6	100	95.8	99.0	97
	85.8	66.5	85.7	97.6	82.2
	85.6	97.7	86.9	97.2	50.5
L	87	98.9	92.0	96.8	97
	97.4	96.0	90.3	97.6	97
	85.4	88	88	95.2	80.2
	93	88.6	88.9	96.8	73.3

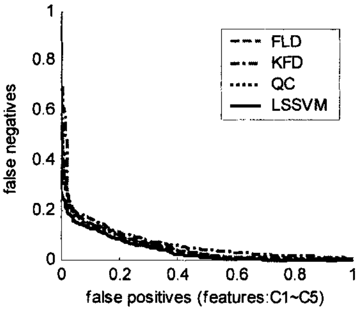

Fig. 2. ROC curves in steganalysis of SES

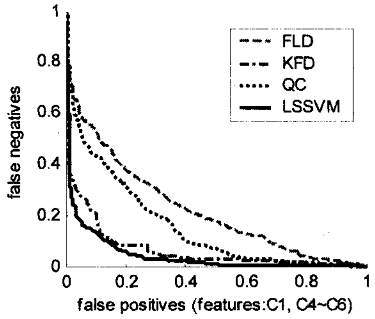

Fig. 3. ROC curves in steganalysis of BPCS

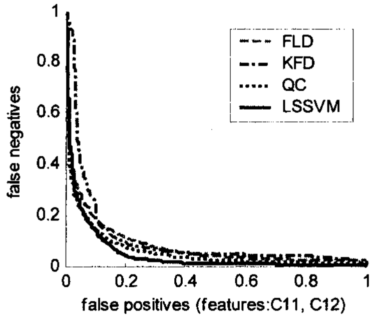

Fig. 4. ROC curves in steganalysis of BMP Secrets

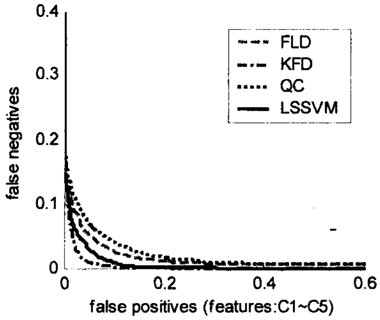

Fig. 5. ROC curves in steganalysis of Invisible Secrets

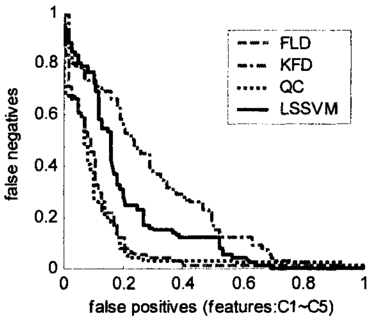

Fig. 6. ROC curves in steganalysis of Secure Engine

4 Conclusions

We presented a scheme for steganalysis based on statistical correlations and learning machine classifiers. Experimental results suggest that it can be applied successfully in the steganalysis of several steganographic systems. Overall, kernel-based classifiers give better train accuracy and test accuracy than the other classifiers, except in the steganalysis of images created using Secure Engine.

Steganalysis is a very challenging problem and, in our view, the successful development of a steganalytic tool will likely rely on multiple steganalytic algorithms and their independent decisions.

Acknowledgements

Partial support for this research received from ICASA (Institute for Complex Additive Systems Analysis, a division of New Mexico Tech), a DoD IASP Capacity Building grant, and an NSF SFS Capacity Building grant is gratefully acknowledged. We also thank S. Mukkamala, T. Xie, and J. Xu for numerous helpful discussions and the anonymous reviewers for their insightful comments and valuable suggestions. The collaborative work of the third author was performed during her sabbatical visit to New Mexico Tech in 2004.

References

[1] A. Westfeld, A. Pfitzmann (2000) Attacks on Steganographic Systems. LNCS 1768, Springer-Verlag, Berlin: 61-75

[2] N. Provos (2001) Defending against Statistical Steganalysis. Proceedings of the 10th USENIX Security Symposium: 323-335

[3] J. Fridrich, M. Goljan, and R. Du (2001) Detecting LSB steganography in color and gray-scale image, IEEE Multimedia: 22-28

[4] H. Farid (2002) Detecting hidden messages using higher-order statistical models. In ICIP 2002, Rochester, New York

[5] S. Lyu, H. Farid (2004) Steganalysis Using Color Wavelet Statistics and One-Class Support Vector Machines. Proc. SPIE: Security, Steganography, and Watermarking of Multimedia Contents VI, vol. 5306

[6] I. Avcibas, N. Memon and B. Sankur (2003) Steganalysis using image quality metrics. IEEE transactions on Image Processing, Vol. 12, No. 2

[7] M. U. Celik, G. Sharma, and A. M. Tekalp, (2004) Universal image steganalysis using rate-distortion curves. Proc. SPIE: Security, Steganography, and Watermarking of Multimedia Contents VI, vol. 5306

[8] E. Kawaguchi, R. O. Eason (1998) Principle and Applications of BPCS-Steganography. Proc. Of SPIE, Vol.3528: 464-473

[9] A. Westfeld (2001) High Capacity Despite Better Steganalysis (F5-A Steganographic Alogrithm). LNCS 2137, Springer-Verlag: 289-302

[10] J. Yu, J. Han, K. Lee, D. Ryu, S. Lee (2003) SES (Steganography Evading Statistical analyses). Pacific Rim Workshop on Digital Steganography 2003

[11] http://www.invisiblesecrets.com/

[12] http://securengine.isecurelabs.com/

[13] http://www.jjtc.com/stegoarchive/stego/

[14] http://www.pworlds.com/products/i_secrets.html

[15] Rencher A. C. (1995) Methods of Multivariate Analysis, John Wiley, New York

[16] J. Huang, D. Mumford (1999) Statistics of Natural Images and Models. Computer Vision and Pattern Recognition, Vol. 1

[17] A. Srivastava, A. Lee, E.P Simoncelli and S. Zhu (2003) on advances in statistical modeling of natural images, Journal of Mathematical Imaging and Vision 18(1): 17-33

[18] G. Winkler (1995) Image Analysis, Random Fields and Dynamic Monte Carlo Methods. Springer: Berlin

[19] M. J. Wainwright, E. P. Simoncelli (2000) Scale mixtures of Gaussians and the statistics of natural images. In: Advances in Neural Information Processing Systems, S. A. Solla, T. K. Leen and K.-R. Muller (Eds.): 855–861

[20] M. I. Schlesinger and V. Hlavac (2002) Ten lectures on statistical and structural pattern recognition. Kluwer Academic Publishers

[21] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle (2002) Least Squares Support Vector Machines, World Scientific, Singapore

[22] R. O. Duda, P. E. Hart, D. G. Stork (2001) Pattern Classification. John Wiley & Sons, 2nd edition

[23] B. Scholkopf, A. J. Smola (2002) Learning with Kernels, MIT Press

[24] S. Raudys (2001) Statistical and Neural Classifiers: An Integrated Approach to Design, Springer-Verlag