Signal Processing 88 (2008) 2138–2157

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Review

A review on blind detection for image steganography

Xiang-Yang Luo [a,b,*], Dao-Shun Wang [b], Ping Wang [a], Fen-Lin Liu [a]

[a] *Institute of Information Science and Technology, 450002 Zhengzhou, PR China*
[b] *Department of Computer Science and Technology, Tsinghua University, 100084 Beijing, PR China*

ARTICLE INFO

Article history:
Received 28 August 2007
Received in revised form
24 March 2008
Accepted 26 March 2008
Available online 1 April 2008

keywords:
Steganography
Steganalysis
Blind detection
Digital image
Review

ABSTRACT

Blind steganalysis techniques detect the existence of secret messages embedded in digital media when the steganography embedding algorithm is unknown. This paper presents a survey of blind steganalysis methods for digital images. First, a principle framework is described for image blind steganalysis, which includes four parts: image pretreatment, feature extraction, classifier selection and design, and classification. We then classify the existing blind detection methods into two categories according to the development of feature extraction and classifier design. For the first category, we survey the principles of six kinds of typical feature extraction methods, describe briefly the algorithms of features extraction of these methods, and compare the performances of some typical feature extraction algorithms by employing the Bhattacharyya distance. For the second category, the development of classifier design, we make a survey on various classification algorithms used in existing blind detection methods, and detail the algorithms behind several classifiers based on multivariate regression analysis, OC-SVM, ANN, CIS and Hyper-geometric structure. Finally, some open problems in this field are discussed, and some interesting directions that may be worth researching in the future are indicated.

Contents

[*] Corresponding author at: Institute of Information Science and Technology, 450002 Zhengzhou, PR China. Tel.: +86 13810496126.
E-mail address: luoxy@theory.cs.tsinghua.edu.cn (X.-Y. Luo).

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.03.016

1. Introduction

Steganography is an art of hiding communication by embedding messages into an innocuous-looking cover medium such as digital image, video, audio and so on, while steganalysis focus on revealing the presence of the secret messages and extract them. Generally speaking, if an algorithm can judge whether a given image contains a secret message or not, the steganographic system is considered broken by this algorithm [1]. Hence, the first aim of steganalysis is detecting the presence of secret messages. Usually, steganalysis methods fall broadly into one of two categories: steganalysis for specific embedding or universal blind steganalysis. The former can be called as specific steganalysis, and it can reveal secret message or even estimate the embedding ratio with the knowledge of the steganographic algorithm, just like RS [2], SPA [3], DIH [4], and LSM [5] algorithms can detect the spatial LSB steganography reliably, and the algorithms of Fridrich and Pevny et al. [6–9] can determine the presence of secret message for some steganography methods that embed message in the DCT domain of the image. But the steganalysis for specific embedding is hardly practical because it is actually difficult for steganalyzers to know what steganography method was used in images. While the latter, universal blind steganalysis, can detect the secret message independent of the embedding algorithm, it is more attractive in many practical applications. Usually, it is likely that steganalysis methods that target a specific embedding method can give more accurate and reliable results than any universal blind steganalysis. Nevertheless, universal blind approaches are very important because of their flexibility and ability to be quickly adjusted to new or completely unknown steganographic methods [1].

Universal blind steganalysis is a meta-detection method in the sense that it can be adjusted, after training on original and stego images, to detect any steganographic method regardless of the embedding domain [1]. In existing literatures on non-specific steganalysis, there are also two kinds of steganalysis methods. One detects original and stego images using original images as the training set and extracting features to classify images, without the help of features of stego images. Strictly speaking, because these methods do not depend on the condition that we have known certain hiding methods used in images, we regard them as actual blind detection methods. The other kind of steganalysis method detects original and stego images by combining the original images and stego images as the training set, where the stego images are obtained using multiple steganography methods embedding messages into original images. This kind of steganalysis method assumes that it is possible to use some special hiding methods in images, but the analyzers do not know which special hiding methods are used. Hence, we can call this kind steganalysis the "half-blind" detection method. It is worth pointing out that the classifier obtained from half-blind detection methods has a certain generalizing capability. Namely, it is possible for these classifiers to detect some new and unknown steganography methods. For example, Avcibas et al. [10] made two cross-validation experiments to show the generalizing capability of the detection algorithm. In one of the cross-validation experiment, the steganalyzer trained on images embedded by Digimarc [11], and tested on images embedded by PGS [12] and Cox et al.'s spread spectrum (SS) method [13]. In another experiment, the steganalyzer trained on images embedded with Steganos [14] and S-tools [15], and tested on images embedded with Jsteg [16] for cross-validation purposes. Results showed that the classifiers are still able to classify when the tested images come from an embedding technique unknown to the steganalyzer, which indicates that the half-blind steganalysis method has a generalizing capability of capturing the general intrinsic characteristics of steganographic techniques. Hence, the kind of half-blind steganalysis is an important part of blind detection researches. In this paper, we also regard this kind of steganalysis as blind detection methods, similar to most existing references.

Since the first blind steganalysis method was presented by Avcibas et al. [17] in 2000, ever-increasing attention has been recently paid to blind steganalysis, and various techniques have been developed to detect stego images blindly in recent years. Among them, some techniques focused on features extraction of images, such as the features extraction algorithms based on image quality metrics [10,17–19], high-order probability density function (PDF) statistics moments of the decomposition subbands coefficients [20–30], center of mass (COM) of histogram characteristic functions (HCFs) [31], statistical moments of characteristic function (CF) of subband histograms [32–37], statistics of empirical or co-occurrence matrix (Co-M) [6,38–41], and merging features of spatial and DCT domains [42,43]. Some others techniques were absorbed in the selection and design of the classifier, such as the classifier based on multivariate regression analysis [10,17,19], One-Class Support Vector Machine (OC-SVM) [23,24,30,45,46], artificial neural network (ANN) [42,43,47], computational immune system (CIS) [26,27] and hyper-geometric structure [28].

This paper reviews the blind steganalytical methods proposed in existing literatures. The organization of the paper is as follows. In the next section, we describe the principle framework of blind image steganalysis and classify the existing blind detection methods into two categories according to the development of feature extraction and classifier design used in existing blind steganalysis methods. In Section 3, we survey the development of feature extraction of image blind steganalysis techniques. Moreover, the classification capabilities of typical feature extraction algorithms are compared using the Bhattacharyya distance [44]. In Section 4, to develop the classifier design, we survey the various classification algorithms used in existing blind detection methods, and especially describe five typical kinds. Some open problems of image blind steganalysis and some interesting directions that may be worth future research are discussed in Section 5. Finally, we conclude our paper in Section 6.

2. Structure of blind image steganalysis

Image blind detection for steganography is actually similar to pattern classification, which centers around two-class classification. Blind detection aims at classifying given images into two categories: original (or cover) and stego images. Some existing blind image steganalysis methods first extract some features from images, then select or design a classifier, and train the classifier using the features extracted from training image sets, and lastly, classify the features. These methods are from Refs. [10,17,42]. Some others detection methods which make image processing operations before extracting feature, such as the methods in Refs. [21,31,34,38]. In Ref. [25], the authors described a general structure of blind steganalysis, which consists of three main stages: (1) stego signal estimation; (2) feature extraction; and (3) classification. In addition, after extracting features, a feature preprocessing process may be able to enhance the efficiency of classification and its accuracy, such as in the methods in Refs. [25,28,37]. Unfortunately, to date, there is no detailed framework to describe how to detect images steganography blindly. Here, we provide a more rounded framework of blind steganalysis tentatively, which consists of the following major parts:

(1) *Image pretreatment*: Take some operations for the considered images before feature extracting, such as converting RGB image into grayscale, cropping, JPEG compression, DCT or DWT transformation and so on, to improve the classification performance.
(2) *Feature extraction*: Extract informative features, namely, select feature must be sensitive to embedding or modification. We should selected features and construct the feature vector with low dimension, which will decrease the computational complexity of training and classification.
(3) *Classifier selection and design*: Select or design appropriate classifiers on the basis of extracted features, adopt a large set of images (the classes of images are known) to train classifiers, and obtain some important parameters of classifiers, which will be utilized for the following classification.
(4) *Classification*: Exploit the deduced classifier in (3) to discriminate the given images, and classify them into two categories: stego and original images.

According to the steps mentioned above, we can present the structure of blind steganalysis in Fig. 1. In fact, we can consider the step of image pretreatment as a part of the process of feature extraction. In the following two sections, we will make a survey on the development of feature extraction and the development of classifier design, respectively.

3. Development of feature extraction of blind image steganalysis

In this section, we will give a survey on the principles of feature extraction approaches used in existing blind detection methods, describe briefly typical algorithms of features extraction and then compare the performances of some typical feature extraction algorithms using the Bhattacharyya distance.

3.1. Principles of features extraction for blind detection

According to the principles, we categorize the existing features extraction of image blind steganalysis methods to six categories, including features extraction based on image quality metrics, higher-order PDF moments of subband coefficients, COM of the HCFs, CF statistical moments of subband histograms, statistical analysis of empirical or Co-M and merging features from multi-domains.

Principle 1: A general underlying idea behind steganography is to create a stego image that is perceptually identical but statistically different from the cover signal. Therefore, this statistical difference can be exploited for detection with the aid of image quality measure (IQM).

Principle 2: The PDF statistics moments from multi-scale decomposition subband coefficient of image will change after embedding secret messages. As a result, statistics PDF moments of subband coefficients can be exploited for detection.

Principle 3: The COM of the HCF of the image will decrease after data embedding, and this phenomenon can facilitate steganalysis.

Principle 4: Data embedding will smoothen out the peaky distributions of wavelet subband coefficients. Then, the statistical moments of CF are extracted from wavelet subband histograms as a feature vector for steganalysis.

Principle 5: Message embedding will change the statistical property of the empirical or Co-Ms of images, and these can be extracted as features.

Principle 6: According to energy conservation, alterations resulting from data embedding in one domain

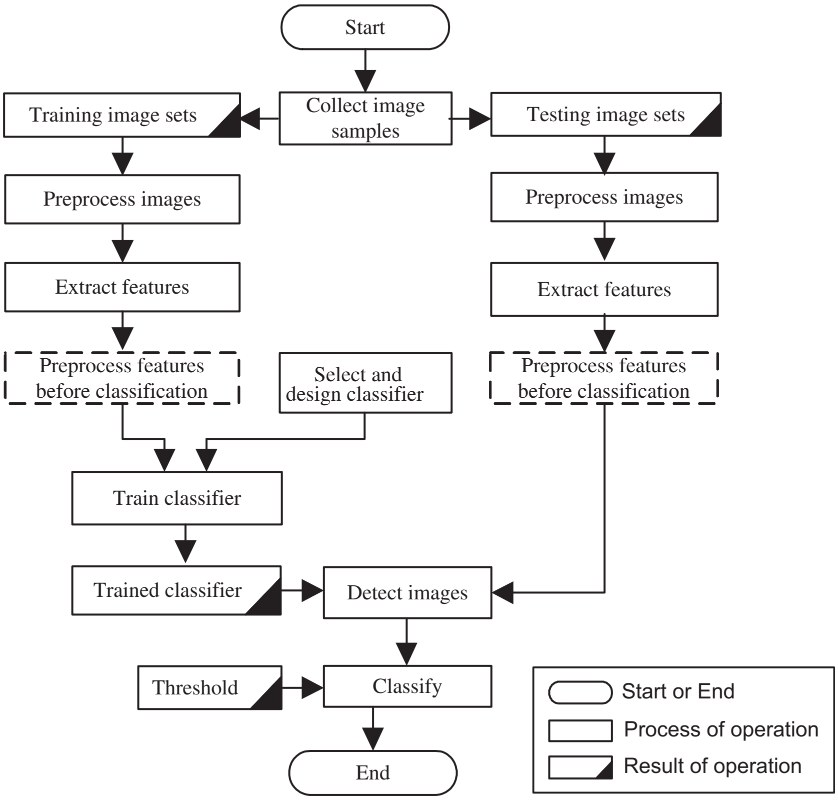

Fig. 1. Framework of blind steganalysis.

will be reflected in the other domains [42]. Therefore, a combination of features in different domains may be more promising.

3.2. *Typical algorithms of features extraction*

Corresponding to the principles mentioned above, there are many algorithms of features extraction. We describe briefly some typical feature extraction algorithms of each category in the following subsections.

3.2.1. *Image quality metrics*

Avcibas et al. [17] presented the image steganalysis technique to judge the existence of watermarks with IQM and multivariate regression analysis. In this paper, Avcibas et al. used some sophisticated image quality metrics as the feature set to distinguish between watermarked and unwatermarked images.

A good IQM should be accurate, consistent and monotonic in predicting quality. Avcibas et al. conducted a statistical analysis on the sensitivity and consistency behavior of objective IQMs in [18]. The measures had been categorized into pixel difference-, correlation-, edge-, spectrum-, context- and HVS-based measures. Their consistency and sensitivity to coding as well as additive noise and blur were investigated by analysis of variance (ANOVA). ANOVA can help us distinguish measures that were most consistent and accurate via the effects of watermarking and blurring. In Ref. [18], 26 measures had been investigated to predict compression, blur and noise artifacts. It was found that measures based on HVS, phase spectrum and multiresolution mean square error were the most discriminative to coding artifacts.

The choice of IQMs in Ref. [17] referred to Ref. [18], and, with respect to the discriminative power, selected subset of IQMs as follows: (1) mean square error; (2) multiresolution distance measure; (3) structural content; (4) cross correlation; (5) weighted spectral distance; (6) median block weighted spectral distance; (7) normalized absolute HVS error; (8) mean Square HVS error; and (9) gradient measure. For an individual image, whose size is $N \times N$, the computing methods of these first 8 measures are given in Ref. [17], as shown in Table 1. Denote the multispectral components of an image at pixel positions i and j, and in band k as $C_k(i, j)$, where $k = 1,\ldots,K$. The boldface symbols $\mathbf{C}(i, j)$ and $\hat{\mathbf{C}}(i,j)$ indicate the multispectral pixel vectors at position (i, j). In Table 1, all quantities with a caret correspond to distorted versions of the same original image. The parameters of mathematical expressions and others IQMs' computing methods can refer to Ref. [18].

Avcibas et al. [19] also discussed the principle of using multiple quality measures, selected the same subset of IQMs in Ref. [17] and extended this technique to further distinguish between specific watermarking techniques.

Table 1
Computing methods of 9 measures in [17]

No.	Quality measure	Computing method
1	Mean square error	$D1=\frac{1}{K}\frac{1}{N^2}\sum_{i,j=0}^{N-1}\lVert\mathbf{C}(i,j)-\hat{\mathbf{C}}(i,j)\rVert^2$
2	Multiresolution distance measure	$D2=\frac{1}{K}\sum_{k=1}^{K}\sum_{r=1}^{R}d_r^k$
3	Structural content	$D3=\frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i,j=0}^{N-1}C_k(i,j)^2}{\sum_{i,j=0}^{N-1}\hat{C}_k(i,j)^2}$
4	Cross correlation	$D4=\frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i,j=0}^{N-1}C_k(i,j)\hat{C}_k(i,j)}{\sum_{i,j=0}^{N-1}C_k(i,j)^2}$
5	Weighted spectral distance	$D5=\frac{1}{N^2}\left(\lambda\sum_{u,v=1}^{N}\lvert\varphi(u,v)-\hat{\varphi}(u,v)\rvert^2+(1-\lambda)\sum_{u,v=1}^{N}\lvert M(u,v)-\hat{M}(u,v)\rvert^2\right)$
6	Median block weighted spectral distance	$J^l=\lambda\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{u,v=1}^{b}(\lvert\Gamma_k^l(u,v)\rvert-\lvert\hat{\Gamma}_k^l(u,v)\rvert)^2\right)^{1/2}+(1-\lambda)\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{u,v=1}^{b}(\lvert\phi_k^l(u,v)\rvert-\lvert\hat{\phi}_k^l(u,v)\rvert)^2\right)^{1/2}$ $D6=\operatorname*{Median}_{l=1,\ldots,L}J^l$
7	Normalized absolute error (HVS)	$D7=\frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i,j=0}^{N-1}\lvert U\{C_k(i,j)\}-U\{\hat{C}_k(i,j)\}\rvert}{\sum_{i,j=0}^{N-1}\lvert U\{C_k(i,j)\}\rvert}$
8	HVS-based L2	$D8=\frac{1}{K}\sum_{k=1}^{K}\frac{\sum_{i,j=0}^{N-1}[U\{C_k(i,j)\}-U\{\hat{C}_k(i,j)\}]^2}{\sum_{i,j=0}^{N-1}[U\{C_k(i,j)\}]^2}$

A total of 48 images were generated: 12 cover images and 36 watermarked images with Digimarc, Cox et al.'s SS and PGS. Results of simulation showed that by using these IQMs as features, one can indeed distinguish accurately among cover images, Digimarc-, Cox's- and PGS-watermarked images.

In Ref. [10], IQMs continued to be the subject of intensive research based on the experimentations of Avcibas et al. [17–19]. Avcibas et al. presented the steganalysis technique for image that had been potentially embedded by steganographic algorithms, within both the passive warden and active warden frameworks. They assumed that steganographic schemes would yield statistical evidence that can be exploited for detection with the aid of IQMs and multivariate regression analysis. Similarly, ANOVA was also used to identify good IQMs, and the multivariate regression technique was adopted to build the classifier between cover images and stego images. There are 10 kinds of selected IQMs by ANOVA, including mean absolute error M_1, mean square error M_2, Czekznowski Correlation measure M_3, angle mean M_4, image fidelity M_5, cross correlation M_6, spectral magnitude distance M_7, median block spectral phase distance M_8, median block weighted spectral distance M_9 and normalized mean square HVS error M_{10}. The feature sets with respect to active warden, passive warden and combined framework are as follows:

(1) Feature set for active warden framework: $\Psi=\{M_1,M_2,M_3,M_5,M_6,M_7,M_{10}\}$.
(2) Feature set for passive warden framework: $\Psi=\{M_4,M_8,M_9,M_{10}\}$.
(3) Feature set for pooled active and passive warden framework: $\Psi=\{M_1,M_2,M_3,M_7,M_8,M_9,M_{10}\}$.

In experiments with well-known and commercially available watermarking and steganographic techniques, such as Digimarc, Cox's and PGS steganography, Steganos [14], S-tools [15] and Jsteg [16] steganography, results indicated that the selected IQMs can make a classification of marked and non-marked images reliably. The selection of IQMs decides the accuracy of detection; however, the choice of IQMs in existing references are experiential. In practice, it is hard to choose the optimum one due to the existing large numbers of metrics standard. In addition, seletion of multiple measures will increase the implement complexity of feature extraction.

3.2.2. *Higher-order PDF moments of subband coefficients*

Farid [20] firstly proposed a steganalysis framework with the help of this kind of higher-order statistics of subband coefficients. He started to decompose an image with separable quadrature mirror filters (QMFs) [48]. As illustrated in Fig. 2, this decomposition splits the frequency space into multiple scales and orientations (refer to Ref. [20] or [48]). This is accomplished by applying separable low-pass and high-pass filters along the image axes and generating a vertical, horizontal, diagonal and low-pass subband. Then subsequent scales can be created by recursively decomposing the low-pass subband. The elements of multiscale decomposition are shown in Fig. 3. Fig. 3(b) and (c) show the result and structure of multiscale decomposition of image Lena (see Fig. 3(a)) with a scale of 3.

Given n-level decomposition of an image, the vertical, horizontal and diagonal subbands at scale $i=1,\ldots,n$ can be denoted by $V_i(x,y)$, $H_i(x,y)$ and $D_i(x,y)$. In the next step, extract the mean, variance, skewness and kurtosis of the subband coefficients at each orientation and each scale as the first set of feature vector. These statistics characterize the basic coefficient distributions. For example, in Ref. [26], for each clean image, a three-level pyramid based on the Daubechies (7,9) biorthogonal filter was constructed. For each subband, higher-order moments, including mean, variance, skewness and kurtosis, were computed; thus a 36-D feature vector for each image was collected. Then these 36 features were reduced to the top three (coefficient mean for each detail subband at scale 1, viz. subband V_1, H_1 and D_1) or the top six (includes top three, coefficient skewness for the D_1 subband, and coefficient mean for V_2 and D_2 subbands) prior to CIS classifier generation.

The second set of statistics was based on the errors in a linear predictor of coefficient magnitude. Although wavelet decomposition minimizes the correlations among coefficients, the little weak correlation across space, orientation and scale still exists. The linear predictor errors of coefficient magnitude can eliminate this kind of weak correlation further, and a linear predictor for the magnitude of these coefficients in a subset of all possible neighbors can be given by

$$\begin{aligned} V_i(x,y) = {} & w_1V_i(x-1,y) + w_2V_i(x+1,y) + w_3V_i(x,y-1) \\ & + w_4V_i(x,y+1) + w_5V_{i+1}\left(\frac{x}{2},\frac{y}{2}\right) + w_6D_i(x,y) \\ & + w_7D_{i+1}\left(\frac{x}{2},\frac{y}{2}\right) \end{aligned} \tag{1}$$

where w_k ($1\leqslant k\leqslant 7$) denoted the scalar weighting values. This linear relationship was expressed more compactly in the matrix form as $\vec{V} = Q\vec{w}$, where the column vector $\vec{w} = (w_1 \cdots w_7)^{\mathrm{T}}$, the vector $\vec{V}$ contained the coefficient magnitudes of $V_i(x,y)$ strung out into a column vector, and the columns of the matrix Q contain the neighboring coefficient magnitudes as specified in Eq. (1), also strung out into a column vector. The coefficients are determined by minimizing the quadratic error function:

$$E(\vec{w}) = [\vec{V} - Q\vec{w}]^2 \tag{2}$$

This error function is minimized by differentiating with respect to $\vec{w}$:

$$\frac{\mathrm{d}E(\vec{w})}{\mathrm{d}\vec{w}} = 2Q^{\mathrm{T}}[\vec{V} - Q\vec{w}] \tag{3}$$

set the result equal to zero, and solve $\vec{w}$ to obtain

$$\vec{w} = (Q^{\mathrm{T}}Q)^{-1}Q^{\mathrm{T}}V \tag{4}$$

Then, the log error of the linear predictor can be shown by

$$\vec{E}_v = \log_2(\vec{V}) - \log_2(Q\vec{w}). \tag{5}$$

Thus, four moments of the log error of three high-frequency subband coefficients at each scale were collected as total 12 features. For n-level decomposition, we can obtain $12n$ statistic moments of the log error. Combining these $12n$ coefficient statistics, $24n$ statistics can be collected to form a feature vector, which can be used to differentiate cover and stego images.

In Ref. [20], a 3-level, three-orientation QMF pyramid was constructed for each image, from which 72 features were collected. Note that the choice of decomposition type may have a significant impact on the quality of the estimation. The authors showed the efficacy of these

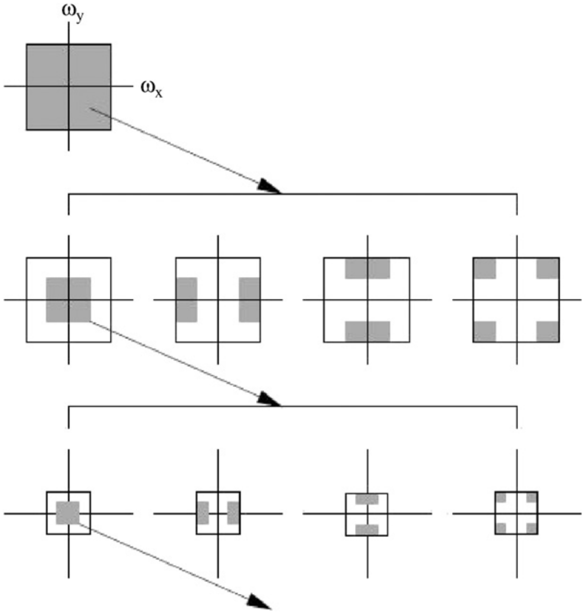

Fig. 2. Image decomposition based on QMF [20].

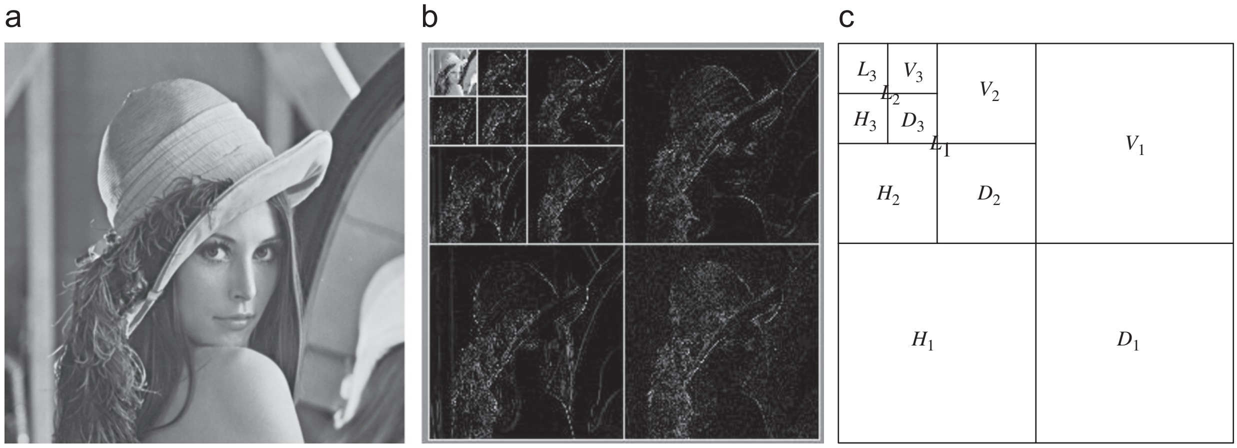

Fig. 3. Three-scale standard wavelet decomposition of image: 3(a) image Lena.bmp; 3(b) result of decomposing; 3(c) structure of decomposing.

extracted features on Jsteg, Ezstego [49] and Outguess [50]. The classifier was based on the Fisher linear discriminant (FLD, see [51,52]). Results of experiment showed that these features have a good performance at discriminating cover and stego images generated by these three steganograpic methods. Later, the 72-length feature vector and the FLD classifier used in Ref. [20] were also used in Ref. [21]. And messages were embedded into images using Jsteg, Ezstego, Outguess and LSB. Results of simulation showed that it is possible to detect stego images produced by Jsteg, Ezstego and Outguess with reasonable accuracies, but LSB with low accuracy.

The method in Ref. [22] differed from Ref. [20] with regard to the classifier and the feature extraction of RGB image. In addition, F5 [53] was also tested in experiments. This steganalysis method performs well in RGB images, because $3 \times 72 = 216$ features were extracted from three-color channels. The addition of color statistics had a considerable improvement in detection accuracy. Analogously, Lyu and Farid [23] extracted features not from the grayscale image but from the RGB image. Correspondingly, the linear predictor was also modified due to the correlation among three-color channels. Thus, for n-level decomposition, a $72n$-D feature vector consisted of $36n$ coefficient statistics, and $36n$ error statistics ($12n$ for per color channel) can be obtained.

In the final version, Lyu and Farid [24] summarized all the above researches and included phase statistics as the other set of feature vector. This paper began with the choice of image decomposition using basis functions that were localized in the spatial, orientation and scale directions. Because both the pixel- and Fourier-based representations do not offer compromise between space and frequency, wavelet decomposition and a local angular harmonic decomposition (LAHD) can localize image structure in both space and frequency. Hence, Farid built an n-level, 3-orientation QMF pyramid for each color channel, and computed $72n$ statistics by the method of Lyu and Farid [23]. Then, another set of statistics was phase statistics extracted from LAHD.

In addition, Holotyak et al. [25] also proposed a new blind steganalysis approach based on higher-order statistical features of wavelet subband coefficients. The first step is to estimate the stego image, and the purpose is to remove or at least minimize the impact of the cover image and obtain a more sensitive signal to embedding. They started with the first-level image wavelet decomposition, which is different from Farid et al.'s methods where higher decomposition levels are used. Moreover, wavelet decomposition was based on orthonormal db8 basis. They showed that stego images can be most accurately estimated in the first-level image decomposition. Obviously, the decomposition only produces 4 subbands, can reduce the dimensionality of feature vector and simplify the construction of the classifier. From the point of estimation of stego image, Holotyak et al. [25] pointed out that the prediction errors used in Farid et al.'s methods also come from some heuristic estimator of the stego signal. In addition, the principal component analysis (PCA) was introduced to reduce the dimensionality of the feature space. Results of experiments based on ±1 steganography [75] and LSB embedding indicated that this feature extraction method has reliable detection results not only for never-compressed images but also for the compressed JPEG images. In addition, an improved version of this method was presented in Ref. [29]. The features are higher-order absolute moments calculated from the wavelet domain, called wavelet absolute moments (WAM).

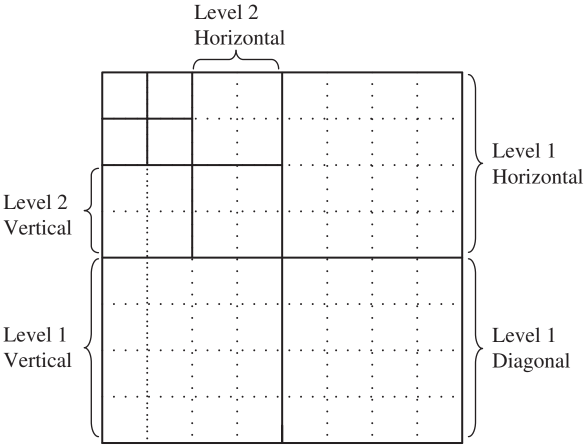

Fig. 4. Rearranged, the DCT transforms in a wavelet decomposition structure [30].

Especially, Rodriguez et al. [30] presented an anomaly detection method using simple one-class classification. They pointed out that the main concern in detecting steganographic with machine learning is in training specific embedding procedures to determine whether the method has been used to generate a stego image, and this will lead to a possible flaw in the detection system when the learned model of stego is faced with a new stego method which does not the existing model. There are three basic components for this paper: (1) a novel DCT multilevel decomposition with wavelet structure was constructed, (2) a new set of feature vectors was extracted, and (3) two classification methods, OC-SVM [45] and Parzen-windows [52], were adopted to solve the anomalous detection of steganography.

The features focus on the energy band of the DCT coefficients of the image. For the standard DCT used in JPEG compression does not generate the multilevel energy bands that wavelet decomposition creates. In order to extract the various energy bands, the DCT transforms are rearranged in a wavelet decomposition structure (see Fig. 4). This structure is created by using the 8×8 pixel blocks from the JPEG compression technique. Rearranging the coefficients of the DCT splits the frequency spectrum into uniform-spaced bands containing vertical, horizontal and diagonal energy. This structure captures the energy better than the normal DCT as well as some wavelet decompositions. Higher-order statistics and predicted log errors are then calculated from the decomposition and are used as features for classification. In this paper, a three-level DCT analysis is performed on each image, and statistics are calculated from the resulting coefficients. Thus, for each image, 120 features can be gathered.

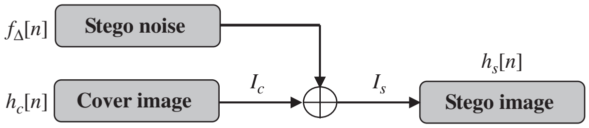

Fig. 5. Additive noise steganography model [31], where <Object Deletion Spot> corresponds to the embedding algorithm.

Although many papers focused on extracting PDF moments of subband coefficients as features, there are also some flaws. The matching between the training samples and test samples strongly affects the result of detection. For example, when the quality factor of JPEG image of the training set is different from that of the test set, the detection accuracy is non-ideal. The embedding ratio also affects the classification; when the embedding ratio is lower, the performance is poorer. In addition, too many features are extracted by some methods, and this will affect the efficiency of classification.

3.2.3. *COM of histogram characteristic functions*

Harmsen and Pearlman [31] began with the fact that data hiding was modeled as additive noise in Fig. 5. The stego noise probability mass function (PMF) was the distribution of the additive noise, defined as

$$f_\Delta[n] \triangleq p(I_s - I_c = n) \tag{6}$$

where I_s and I_c were the pixel values after and before embedding, respectively, $f_\Delta[n]$ is the probability that a pixel will be altered by n, $n \in [-255, 255]$. Let the histogram of stego image and cover image be h_s and h_c, respectively. In a hiding system where the additive noise is independent of the cover image, the histogram of stego image equals the convolution of the stego noise PMF and the cover image histogram, and can be described as follows:

$$h_s[n] = h_c[n] \circledast f_\Delta[n] \tag{7}$$

Namely, given a hiding scheme in the form of $f_\Delta[n]$ as well as the knowledge of $h_c[n]$, the histogram of the stego message would be known, where $n = 0, \ldots, N-1$, and N is the largest intensity possible in the image. For example, $N = 2^8 = 256$ in an 8-bit grayscale image. Considering DFT of the PMFs involved, the CFs were defined as follows:

$$\tilde{f}_\Delta[k] = \mathrm{DFT}(f_\Delta[n])$$
$$\tilde{h}_c[k] = \mathrm{DFT}(h_c[n])$$
$$\tilde{h}_s[k] = \mathrm{DFT}(h_s[n])(8)$$

Particularly, DFT of a histogram will be referred to as HCF. The formulation (8) in the DFT domain was rewritten as

$$\tilde{h}_s[k] = \tilde{f}_\Delta[k]\tilde{h}_c[k] \tag{9}$$

This shows that data embedding will alter the HCF. Then, the COM of HCF can be expressed by

$$\mathrm{COM}(\tilde{h}[k]) = \frac{\sum_{k \in K} |k\tilde{h}[k]|}{\sum_{i \in K} |\tilde{h}[i]|} \tag{10}$$

where $K = \{0, \ldots, (N/2) - 1\}$ and N was the DFT length. The COM gave general information about the energy distribution in HCF. Data embedding will result in a decrease in COM, namely,

$$\mathrm{COM}(\tilde{h}_s[k]) \leqslant \mathrm{COM}(\tilde{h}_c[k]) \tag{11}$$

which were used as features input of a bivariate classifier to differentiate cover from stego images. Finally, a blind detection scheme was built by used-only statistics from original images. By calculating the Mahalanobis distance from a test COM to the training distribution, a threshold was used in classification.

To verify the efficacy of this scheme, 24 images from the Kodak digital camera were used, and these images were 24-bit, 768 × 512 pixels and lossless true-color images stored in PNG format. Two detection systems were built, one against known steganography schemes (SSIS [54] steganography was used to experiment) and the against unknown schemes, in which SSIS, DCT [55] and LSB steganographic methods were adopted. Although the paper showed that this method worked well, there still existed some flaws, for example, there were only three simple steganographic methods and a few training and test images (less than 24 images) used in experiments, and only full embedding was considered in experiments so that its performance was desirable, but it may suffer in the case of lower embedding or some other steganographic methods.

3.2.4. *CF moments of subband histograms*

In Ref. [32], on each image was performed a 2-level Haar wavelet decomposition, and some features were extracted from eight subbands denoted by LL_1, HL_1, LH_1, HH_1, LL_2, HL_2, LH_2 and HH_2, and the image itself by LL_0. For each subband, the n-order CF moments are defined as follows:

$$M_n = \frac{\sum_{j=1}^{N} f_j^n A_j}{\sum_{j=1}^{N} A_j} \tag{12}$$

The first- and second-order moments were adopted as features in Ref. [32]. Here, A_j was the amplitude of the jth frequency component f_j, and N is the total number of points in the horizontal axis of the histogram. Totally, an 18-D feature vector was extracted for steganalysis. In addition, Bayes classifier and 1096 CorelDraw [56] image database were introduced in this paper. In experiments, 100 images were randomly selected for training, and the remaining 996 images were used for test. For the non-blind Cox's SS, the correct detection rate was 79%; for the blind Piva et al.'s SS, it reached 88%; and for the LSB-like method, it amounted to 91%.

In Ref. [36], the statistical moments of CFs from the test image, the prediction-error image and their wavelet subbands are combined and selected as features. Here, the prediction-error image is used to erase the image content, and the prediction algorithm is expressed as follows:

$$\hat{x} = \begin{cases} \max(a,b) & c \leqslant \min(a,b) \\ \min(a,b) & c \geqslant \max(a,b) \\ a+b-c & \text{otherwise} \end{cases} \tag{13}$$

where a, b and c are contexts of the pixel x under consideration, and $\hat{x}$ is the prediction value of x.

The BP neural network was used as the classifier. Steganography methods, including Cox et al.'s non-blind SS [13] ($\alpha = 0.1$), Piva et al.'s blind SS [57], Huang and Shi's block-DCT-based SS [58], a generic QIM (Quantization Index Modulation) [59] (0.1 bpp) and a generic LSB (0.3 bpp), were used in experiments. This detection method demonstrates a significant performance improvement over those of Farid [21] and Harmsen and Pearlman [31]. Results of experiments showed that prediction-error images can enhance the changes caused by data hiding though reducing the effect caused by the diversity of natural images. The combined features extraction steganalysis approach holds promise for blind and practically powerful steganalysis.

Theoretical analysis had pointed out that the defined nth statistical moment of a wavelet CF was related to the nth derivative of the corresponding wavelet histogram in Ref. [34]. Therefore, these moments will be sensitive to data embedding. For each image, Xuan et al. built a three-level pyramid using Haar wavelet, extracted 39 CF moments from wavelet histogram as features and selected the Bayes classifier for steganalysis. The experiments about five above-mentioned steganography methods were made and the results demonstrated this method's more powerful classification capability. Besides, this paper applied the Bhattacharyya distance into reducing the 39-D feature vectors to the 3-D feature vectors, and the corresponding average detection rate as 87.0%, which is also better than the detection rates of Refs. [21,31], except for Jsteg steganography.

For JPEG image steganography, Chen et al. [35] extracted statistical CF moments derived from both image pixel array and JPEG coefficient array as features. In addition to the first-order histogram, the second-order histogram was considered. Results of experiments showed that this detection method based on these features outperforms in general the methods of Refs. [6,21,36] in detecting Outguess, F5 and Model-based steganography [60].

Recently, Wang et al. [37] investigated the feature extraction problem of image steganalysis from the following three angles: (1) Image subband decomposing. Farid et al.'s [22] image representation includes wavelet subband coefficients and their cross-subband prediction errors. Wang et al. discovered that decomposing the diagonal subband on the first scale and combining the resulting detail subbands with Farid et al.'s representation is beneficial. The image decomposing is shown in Fig. 6. (2) Choice of features. This paper considered both empirical PDF and CF statistic moments as a feature. A reasonable embedding model in the wavelet domain takes the form of a generalized Gaussian cover signal plus independent Gaussian embedding noise. Under this model, the authors proved that the empirical CF moments of subband histograms were more sensitive to embedding and were better features than empirical PDF moments of subband coefficients. However, for the prediction-error subband, unlike features of wavelet coefficient subband, the empirical PDF moments outperform the empirical CF moments. It is a pity that this conclusion is not proved in theory, and is only an experimental conclusion. (3) Feature evaluation and selection. They applied feature dimensionality reduction techniques from the pattern recognition and machine-learning literature [61] to image steganalysis, and improved the classification performance. It is well known that all features are not equivalence to classification, and too many features are undesirable in terms of classification performance due to which one cannot reliably learn the statistics of too many features when given a limited training set. Hence, the evaluation and selection of features are a significant.

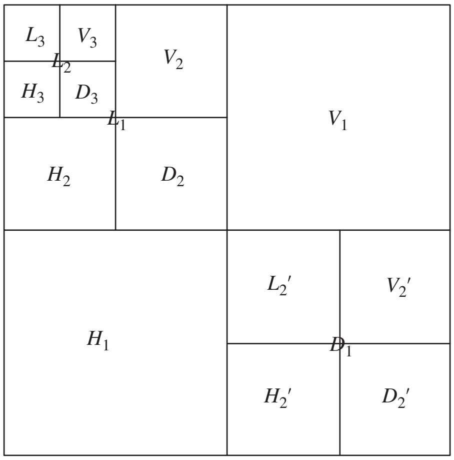

Fig. 6. Three-scale standard wavelet decomposition and an extra level of decomposition on the first-scale diagonal subband D1 (refer to [37]).

3.2.5. *Statistical analysis of empirical or co-occurrence matrix*

In Ref. [38], the empirical matrix (EM) (or the Co-M) was regarded as a raw representation of the statistical characteristics of images. For a given grayscale image I with N gray levels, the $N \times N$ EM $M_{r,\theta}$ of image is defined as

$$M_{r,\theta}(i,j) = P\Big(I(x_1,y_1) = i, I(x_2,y_2) = j \,\Big|\, \begin{matrix} x_2 = x_1 + r\cos\theta \\ y_2 = y_1 + r\sin\theta \end{matrix}\Big) \tag{14}$$

where $I(x,y)$ means the value of pixel at location (x,y), r is the step, θ is the direction, and P represents the normalized probability. When $r = 1$, $\theta = 0$, it can be seen that the EM is highly concentrated in the diagonal line of $i = j$, the reason being that the neighboring pixels are highly correlated and tend to have the same or close gray value. The concentration effect of EM will be weakened after data hiding. Chen et al. [38] projected the EM along the diagonal line, and generated the 1-D projection histogram $h_{r,\theta}$:

$$h_{r,\theta}(k) = \sum_i M_{r,\theta}(i, i+k) \tag{15}$$

where $1-N \leqslant k \leqslant N-1$, $k \in Z$, and the length of $h_{r,\theta}$ is $2N-1$. It is noted that $h_{r,\theta}$ will be flattened after data hiding, and the multiple-order moments of $h_{r,\theta}$ are then used as features to detect the data hiding. The nth order moments of $h_{r,\theta}$ are defined as follows:

$$mh_{r,\theta}^{n} = \left| \frac{\sum_{k=1-N}^{N-1} k^{n} h_{r,\theta}(k)}{\sum_{k=1-N}^{N-1} h_{r,\theta}(k)} \right| \tag{16}$$

Assuming that noise introduced by data hiding is additive and Gaussian distributed, the CF of $h_{r,\theta}$ is DFT of $h_{r,\theta}$, and marked by $F_{\gamma,\theta}$, then the multi-order moments of $F_{\gamma,\theta}$ can be defined as

$$mF_{r,\theta}^{n} = \frac{\sum_{j=1}^{(L/2)} f_j^{n} |F_{r,\theta}(f_j)|}{\sum_{j=1}^{(L/2)} |F_{r,\theta}(f_j)|} \tag{17}$$

where $F_{\gamma,\theta}(f_j)$ is the component of $F_{\gamma,\theta}$ at frequency f_j.

In experiments, three directions and three steps were expressed as

$$\{(r,\theta)|r = 1,2,3;\ \ \theta = 0, \tfrac{\pi}{2}\} \cup \{(r,\theta) \\ \left| r = \sqrt{2}, 2\sqrt{2}, 3\sqrt{2}; \theta = \tfrac{\pi}{4} \right\}. \tag{18}$$

There were totally 9 Ems; the projection histogram corresponding to EM was generated and the first three order moments of each histogram were calculated. Additionally, the CF of each histogram was generated and the first three order moments of each CF were calculated, and thus obtain a 54-D feature vector. Another 54-D feature vector was also extracted from the prediction-error image similar to Ref. [36]. A combined 108-D feature vector and SVM [45] were utilized as classification.

In experiments, the image database comprised all the 1349 images in CorelDraw version 11 CD#4, and the stego images generated by six steganography schemes: Cox et al.'s non-blind SS, Huang et al.'s 8 × 8 DCT block SS, Piva et al.'s blind SS, generic LSB, Lie et al.'s adaptive LSB and the generic QIM method. To show the efficacy of the above method, a very small amount of data was embedded in an image. Results showed that the average detection rate of stego images is 98%, and the correct detection rate was more effective than those of Refs. [36,42] when embedding a small amount of data.

In Ref. [6], Co-Ms of the neighboring JPEG coefficients are used to extract features. In Ref. [39], the Markov chain was firstly used for steganalysis, the algorithm of this paper scanned the whole image horizontally row-by-row and then calculated the empirical transition matrix. Xuan et al. [40] indicated that this is essentially something similar to the Co-M. Since the dimensionality of Co-M is extremely high (e.g., $256 \times 256 = 65{,}536$ for an 8-bit gray-level image), not all elements of the matrix can be used as features. Sullivan [39] only selected several largest probabilities along the main diagonal together with their neighbors, and then randomly selected some other probabilities along the main diagonal as features, resulting in a 129-D feature vector. This technique, though designated to attack spread spectrum data hiding, provides some insights that motivated some other methods in attacking JPEG steganography.

For example, Xuan et al. [40] proposed another blind steganalysis scheme using high-dimensional features derived from Co-M. In addition to working on the gray-level Co-Ms for attacking steganographic methods in the spatial domain, this paper also worked on the Co-Ms associated with the JPEG coefficient domain to attacking JPEG steganographic techniques, such as Outguess, F5 and model-based steganography. Considering the 2-D nature images, in either case, this paper considers the vertical, main-diagonal and minor-diagonal directions in addition to the horizontal direction when generating Co-Ms. Fu et al. [41] also proposed a steganalysis scheme to effectively attack JPEG steganographic methods. In this scheme, Markov empirical transition matrices are proposed to capture both intra-block and inter-block dependencies between block-DCT coefficients in the JPEG image. Since the hidden messages are independent of cover data, the embedding process usually decreases the dependencies existing in cover data. Therefore, the second-order statistics used in this scheme can capture such kind of changes. In Refs. [40,41], to solve the classification problem in high-dimensional space, a class-wise non-principal component analysis (CNPCA) and a threshold technique are applied, respectively.

3.2.6. *Merging features from multidomains*

In Ref. [42], the authors focused on a feature extraction technique based on merging two statistical properties in the spatial and DCT domains, and determined the existence of hidden messages in images.

- *Spatial domain feature: gradient energy*
  The gradient energy of the image would increase after data embedding, and the spatial feature f_1 of an image based on gradient energy can be computed in the following steps
  (1) Calculate the vertical gradient energy GE_V by

  $$GE_V = \frac{1}{N_H N_W} \sum_x \sum_y (I(x,y) - I(x,y-1))^2 \tag{19}$$

  where N_H and N_W represent the width and the height of the region of interest, respectively, and $I(x, y)$ denotes the pixel value at (x, y).
  (2) Calculate the horizontal gradient energy GE_H by

  $$GE_H = \frac{1}{N_H N_W} \sum_x \sum_y (I(x,y) - I(x-1,y))^2 \tag{20}$$

  (3) Calculate total gradient energy as f_1:

  $$f_1 = GE_V + GE_H \tag{21}$$

- *DCT domain feature: statistical variance of the Laplacian parameter*
  Every four adjacent blocks of 8 × 8 pixels were grouped into a macroblock (MB) for parameter estimation, and then all DCT coefficients except for the DC components in an MB can be modeled as Laplacian distribution. Then, the feature f_2 can be calculated as follows
  (1) For MB U_i in an image, model all AC coefficients t as Laplacian distribution, and then the PDF is

  $$p(t) = \frac{\lambda}{2} e^{-\lambda |t|} \tag{22}$$

  (2) Estimate parameter λ by maximum likelihood estimation. The parameter λ_i can be estimated as

Table 2
Feature extraction methods in the existing blind steganalysis methods

	1	2	3	4	5	6
Extracted feature	Image quality metrics	PDF moments of subbands coefficients	Mass's center of HCF of subbands coefficients	CF moments of subbands histograms	Statistics of EM and Co-M	Merging spatial and DCT domain features
Main papers	[10,17,19]	[20–30]	[31]	[32–37]	[6,38–41]	[42,43]
Representative works	[10]	[22]	[31]	[34]	[38]	[42]
Dimensions of feature vectors	10	72	1	39	54	2

follows:

$$\hat{\lambda}_i = \frac{K}{\sum_{j=1}^{k} |t_j|} \tag{23}$$

where $K = 4 \times (64-1) = 252$. Let the result be denoted as $\hat{\lambda} = \left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_L\right)$, where L was the number of MBs in an image.

(3) Calculate the mean $\bar{\lambda}$ and variance $\mathrm{Var}(\lambda)$ of the estimated parameter vector.

$$\bar{\lambda} = \frac{1}{L}\sum_{i=1}^{L} \hat{\lambda}_i, \quad \mathrm{Var}(\lambda) = \frac{1}{L-1}\sum_{i=1}^{L}\left(\hat{\lambda}_i - \bar{\lambda}\right)^2 \tag{24}$$

Finally, $f_2 = \mathrm{Var}(\lambda)$.

Later, Lin et al. [43] presented another steganalysis scheme, which was also based on gradient energy and statistical variance of the coefficient distribution in MB, to identify the data-hiding domain. The 16 × 16 pixel block was regarded as an MB, and the Sobel operator was also adopted to calculate the gradient energy, and then MBs with lower gradient values were chosen as candidate MBs. For the spatial domain, one of the features is the average of gradient of all candidate MBs, denoted by f_1^p, while the other is the average of energies of the middle and high bands in each block, marked by f_2^p. For the DCT domain, AC coefficients in MB can be modeled as Laplacian distributed. The parameter estimated via maximum likelihood forms a feature f_1^D; furthermore, the average energies of the middle-and high-frequency subbands in selected MBs can be the feature f_2^D. For the DWT domain, high-frequency subbands can be divided into several MBs, and all coefficient histograms can be obtained. The average standard deviations of the histogram in each MB form is f_1^W. Besides, the average gradients of the vertical and horizontal directions can be calculated as f_2^W. In Ref. [43], a multilayer neural network classifier was exploited to detect whether a test image is embedded into a secret message or not. A minimum likelihood ratio test was adopted to solve the hiding domain classification problem. The result demonstrated that this method could achieve right hiding domain discrimination among spatial, DCT and DWT domains.

We take the logical next step toward detecting the image by merging features from three domains; spatial, DCT and DWT domains, respectively. Results of a series of experiment validated the performance of the proposed method for 8 kinds of steganography methods, including 4 kinds of BMP image steganography: LSB, PMK [75], LTSB [76,77] and S-Tools, and 4 kinds of JPEG image steganography: Jsteg, F5, Outguess and JPHide. The results show that the merging can make more reliable detection for these typical steganography methods than Ref. [42]. However, there are also some flaws in these methods. This is only a simple combining of features and not an analysis of the relation of consistent and afoul among these features. Besides, when more domains are used to extract feature, more excellent-performance is uncertain. The problem as to what domains and statistical features should be selected to merge also needs further research. It is noticeable that the merging of features should be restricted to a selection of just the best features and also to a merging of the features to create better ones.

3.3. Evaluation of the discrimination capability of features

Here, we summarize briefly typical blind steganalysis methods in Table 2, and use the Bhattacharyya distances to evaluate the usefulness of features in discriminating between classes.

The Bhattacharyya distances can be calculated by

$$B(p_0, p_1) = -\log \int_{\chi} \sqrt{p_0(x)p_1(x)}\,\mathrm{d}x \tag{25}$$

where x is a feature vector, χ is the feature space,and $p_0(x)$ and $p_1(x)$ are the feature PDFs under Class 0 and Class 1, respectively. The larger the $B(p_0,\ p_1)$ for a feature, the better the suitability of that feature for classification. Always $B(p_0,\ p_1)\geqslant 0$; only when $p_0 = p_1$, $B(p_0,\ p_1) = 0$, the feature is useless. In practice, p_0 and p_1 are often unavailable and, instead, we use their histogram estimates from training features and compute the empirical Bhattacharyya distance. Consider that the dimensions of various extracted feature vectors are different, we calculate the average Bhattacharyya distances $B(p_0,\ p_1)/\mathrm{Dim}$ to evaluate the usefulness of feature vectors, where Dim means the dimension of feature vectors.

In order to examine the Bhattacharyya distances of various features extracted by the representative methods of Table 2, we perform a series of experiments under the platform Matlab 7.0. Original images in our experiments include 2008 images from the image library of NRCS [81] and 48 usually standard images (Lena, Baboo, Peppers, etc.). Half of the original images (1028 images) are BMP images and the others are JPEG images (image quality: 75), and the types of size of this original image set include

768 × 512 (634 images), 640 × 480 (84 images), 512 × 512 (25 images) and 496 × 328 (285 images). Besides, contents of these images are selected widely, including nature scenery, artificial buildings, human portraits and animal photographs. Seven categories of typical stego images are generated for each original image using the following seven typical data-hiding schemes, which include three kinds of steganography methods for BMP images and four kinds of steganography methods for JPEG images.

#1: Generic LSB (1 bpp).
#2: PMK [75] (1 bpp, the value of k is 1).
#3: Cox el al.'s SS [13] (10,000 bits).
#4: Jsteg [16] (image with 256 × 256 pixels).
#5: F5 [53] (image with 200 × 200 pixels).
#6: JPHide [62] (image with 64 × 64 pixels).
#7: Model-based steganography [60] (with maximal embedding capability).

To make our tests persuasive, we customized the data-hiding methods to embed various sizes of message, which are shown in brackets. For each stego method, the average Bhattacharyya distance of feature vectors extracted by each representative detection method is shown in Fig. 7, which shows that the average Bhattacharyya distances of feature vectors of [34] are almost bigger for all seven kinds of steganography. This indicates that, in the mass, the discrimination capability of features by extracting CF moments of wavelet subbands coefficients is more or less the most powerful.

4. Development of classifier choosing and design

Besides many features extraction methods that are presented in the existing literatures of blind steganalysis, various classifiers are selected or designed to classify the extracted features. In this section, we try to survey the development of classifiers in blind steganalysis methods.

4.1. Survey of classifiers in blind steganalysis methods

The classifiers used in existing blind steganalysis references almost include all categories of classical classifiers, such as FLD, Bayes, ANN and SVM. Moreover, some papers selected and improved the typical classifiers (such as OC-SVM), and proposed some new classification methods for steganalysis; for example, the classifiers based on multivariate regression analysis, CIS and Hyper-geometric. The classifiers used in the existing blind detection methods are shown in Table 3, where the italic content means the classifier has the best classification accuracy among all classifiers adopted in the corresponding paper.

4.2. Some specific classifiers used in existing blind steganalysis methods

In the following subsections, we will briefly introduce the classifiers based on multivariate regression analysis, OC-SVM, ANN, CIS and Hyper-geometric.

4.2.1. Classifier based on multivariate regression analysis

In Refs. [10,17,19], some appropriate IQMs were selected as the feature set, and then they built an optimal classifier between watermarked and unwatermarked images using the multivariate regression analysis technique.

In the process of classifier design, the normalized IQM scores were regressed to −1 and 1, which depend on whether an image contained a watermark or not. In the regression model, each decision label $y_i(i = 1,\dots,N)$ in a sample of N images was expressed as a linear function of IQM scores, plus a random error $\varepsilon_i(i = 1,\dots,N)$,

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_q x_{1q} + \varepsilon_1 \\ y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_q x_{2q} + \varepsilon_2 \\ &\vdots \\ y_N &= \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_q x_{Nq} + \varepsilon_N \end{aligned} \tag{26}$$

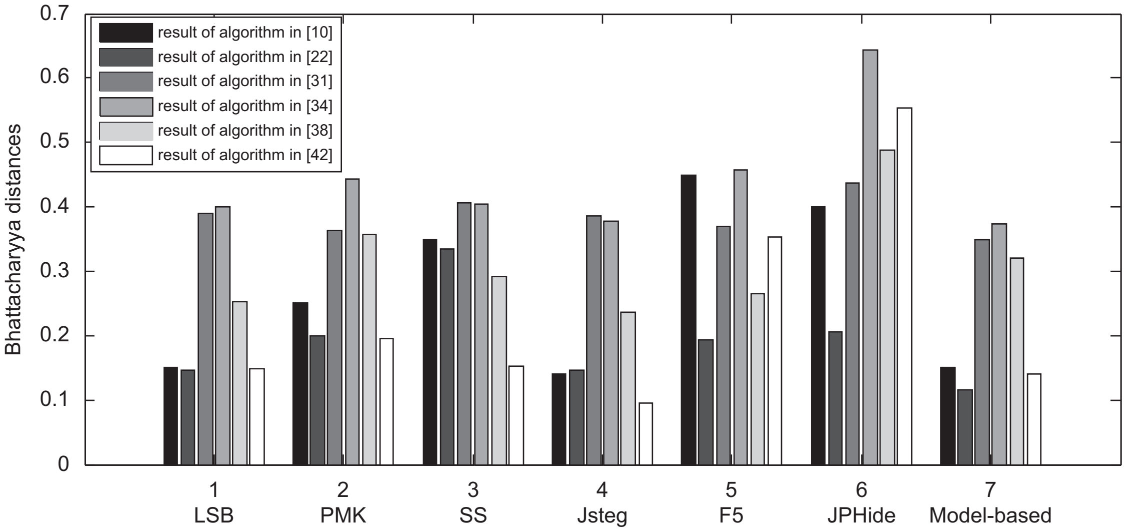

Fig. 7. Bhattacharyya distances of representative blind detection algorithms for seven typical steganographic methods (LSB, PMK, SS, Jsteg, F5, JPHide and model-based).

Table 3
Classifiers in the existing blind steganalysis methods

Category	Extracted feature	Paper	Classifier
1	Image quality metrics	[10]	Multivariate regression analysis
		[17]	Multivariate regression analysis
		[19]	Multivariate regression analysis
2	PDF moments of image subbands coefficients	[20]	FLD
		[21]	FLD
		[22]	Linear separable SVM, Linear non-separable SVM, *Non-linear SVM*
		[23]	OC-SVM
		[24]	Linear SVM, *Nonlinear SVM*, OC-SVM
		[25]	FLD
		[26]	CIS
		[27]	CIS
		[28]	Hyper-geometric
		[29]	FLD
		[30]	OC-SVV, *Parzen-window*
3	Mass's center of HCF of subbands coefficients	[31]	Bayes
4	CF moments of subband histograms	[32]	Bayes
		[33]	Bayes
		[34]	Bayes
		[35]	SVM
		[36]	Bayes, *ANN*
		[37]	FLD
5	Statistics of EM and Co-M	[6]	FLD
		[38]	SVM
		[39]	SVM
		[40]	*CNPCA*, PCA+Bayes
		[41]	SVM
6	Merging features (spatial and DCT domain features)	[42]	ANN
		[43]	Bayes, *ANN*

where $x_{ij}(i = 1,\ldots,N, j = 1,\ldots,q)$ represent IQM scores in the ith image and the jth quality measure, q means the total number of quality measures. The complete statement of the standard linear model is

$$y = X_{Nxq}\beta + \varepsilon \tag{27}$$

where the data matrix X_{Nxq} has rank q, and ε is a zero-mean Gaussian noise. The corresponding optimal minimum mean square error (MMSE) linear predictor β can be obtained by

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}y) \tag{28}$$

Given a test image, q IQM scores can be obtained using the image. Then, these scores can be regressed to the output. That is

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_q x_q \tag{29}$$

If the output $\hat{y}$ exceeded the threshold Thd, the image will be considered as a stego image, otherwise as a cover image.

Three watermarking techniques, Digimarc, Cox's watermarking technique and PGS, were adopted to examine the performance of this detection algorithm. Two types of training and testing were performed: (1) The individual steganalysis of each watermarking algorithm was trained by 12 images and tested on 10 images, and the detection rates for Digimarc, PGS and Cox's were 85%, 75% and 100%, respectively. (2) The mixed steganalysis of three watermarking algorithms were trained on 36 images and tested on 30 images, and the detection rate was 85%. These results show that the detection algorithm can reliably detect these three kinds of watermarking techniques. Expanded experimental results indicate that the detection algorithm can still perform a blind classification when the tested images come from an embedding technique unknown to the steganalyzer; in other words, this means it has a generalizing capability of capturing the general intrinsic characteristics of watermarking and steganographic techniques.

4.2.2. *Classifier based on OC-SVM and Parzen-Window*

Farid and Lyu [22] employed three classifiers; namely linear separable SVM, linear non-separable SVM and non-linear SVM, to classify the extracted features, and compared the detection results. They found that the detection accuracy by linear non-separable SVM was almost the same as that by FLD, but non-linear SVM can drastically improve the detection results. While the linear SVM and non-linear SVM techniques afforded good classification accuracy, they required training with both cover and stego images.

Since there are numerous stego techniques that might need to be trained, it would be advantageous to build a classifier from only the more easily obtained cover images. Shown in Fig. 8(a) is a toy 2-D example where a non-linear SVM was trained on black dots (cover) and white squares (stego), and where the dashed line corresponds to the separating surface. In this same figure, the gray squares correspond to previously unseen images from a different stego program. Notice that without explicit training on the gray squares, the classifier is

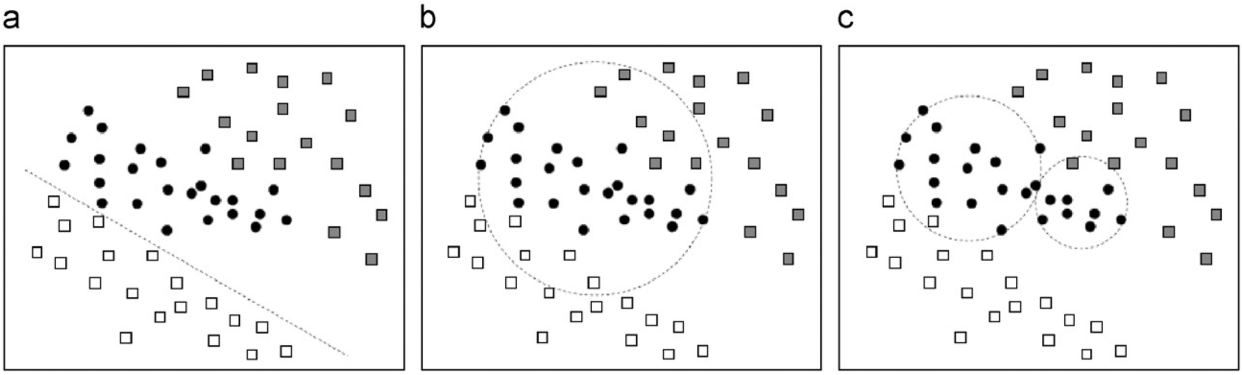

Fig. 8. Toy examples of (a) two-class SVM, (b) one-class SVM with one hyper-sphere, and (c) one-class SVM with two hyper-spheres (refer to [23]).

unable to correctly classify them. To contend with this problem, Farid et al. employed OC-SVM in Refs. [23,24].

In each case of Fig. 8, the dotted line or circle represents the classifier. The two-class SVM is trained on the black dots (cover) and white squares (stego)—notice that the gray squares (also stego) will be incorrectly classified as they were not included in the training. The OC-SVMs are trained on only the black dots—notice that in these cases the classifier is better able to generalize as both the white and gray squares generally fall outside the support of the bounding circle(s).

An OC-SVM is trained on data from only one class by computing a bounding hyper-sphere (in the projected high-dimensional space) that encompasses as much of the training data as possible, while minimizing its volume. For example, Fig. 8(b) shows an OC-SVM trained on the black dots. Note that, unlike the two-class SVM shown in panel (a), this classifier is able to classify, reasonably well, both types of stego images (white and gray squares). The implement details of OC-SVM refer to Ref. [23], which used LibSVM [46] to realize various SVM classifiers.

The OC-SVM is an ‘‘actual blind'' classification technique and makes the training easier and faster. In Refs. [23,24], linear SVM, non-linear SVM and OC-SVM were used to discriminate between cover and stego image. A database of 40,000 color images under JPEG compression was used with a quality of 90, and the central 256×256 region of each image was regarded as objects in the steganalysis. Secret messages were embedded using Jsteg, Outguess, F5, JPHide [62] and Steghide [63]. Sets of experiments were also made on images with different quality factors. Results showed these classifiers did not generalize well to new JPEG quality factors. Furthermore, the classifier was also trained on TIFF stego images with LSB embedding and GIF stego images with Ezstego embedding. The detection accuracy was 72.3% and 64.4%, respectively, in the case of the maximum capacity. Extensive experiments show that non-linear SVM significantly outperforms linear SVM, whereas OC-SVM affords a simple training but degrades in detection accuracy.

Like the OC-SVM, Parzen-windows [30,52] can also be used to classify images based on only the feature sets of cover images, and also is an actual blind classifier. Rodriguez et al. [30] adopted this technique to compare with OC-SVM and indicated that the performance of Parzen-windows outperforms that of OC-SVM. Parzen-windows can be used as a blind classifier, but are often not. The additional step required is to identify a cutoff threshold between the trained and untrained classes.

4.2.3. *Classifier based on artificial neural network*

In Ref. [36], an ANN, the feed-forward neural network [47], was used as the classifier. It is expected that the powerful learning capability possessed by the neural network outperforms the linear classifiers. The numbers of neuron of the hidden and the output layers are 4 and 1, respectively, and all neurons in the hidden layer use the tan-sigmoid function. For the one-neuron output layer, all three activation functions (linear, log-sigmoid and tan-sigmoid) were tested in the simulation. In experiments, they found that in the training stage, the outputs of log-sigmoid and tan-sigmoid neurons have a larger mean square error than the linear neuron output, and in the testing stage, the linear neuron output provides a higher classification rate than the non-linear outputs. The authors analyzed the reason of this phenomenon and constructed a reasonable structure of ANN, which is composed of four tan-sigmoid neuron hidden layers and one linear neuron output layer. The scope of the output value is [0.0, 1.0] (Lie and Lin [42] indicated this scope is (0.0, 1.0), we think this should be [0.0, 1.0] because the value 0 and 1 often appear in the output results). When the output value is close enough to 1.0, the classifier indicates the existence of hidden data in an image. On the contrary, if the output value is near 0.0, the classifier reveals the considered image is an original image.

In Ref. [42], after extracting the spatial and DCT domains features, a three-layer feed-forward neural network was introduced and implemented to classify the original and stego images. The input layer contains two neurons, which accept the extracted feature vector (f_1, f_2), the output layer contains one neuron to indicate the output result, and the hidden layer is composed of several neurons (the number of neurons has not been provided in Ref. [42]) to memorize the training sample set. The log-sigmoid function $g(x) = 1/(1+\exp(-x))$ was adopted due its good properties in continuity and smoothness. This paper considered the over-fitting problem in the training phase, and proposed a method to decrease the effect of this problem. They adopted the Q-fold cross-validation method [52]. That is, the training set is randomly divided into Q disjoint sets of equal size n_s/Q, where n_s is the total number of samples. The $(Q-1)$ sets arbitrarily selected from Q disjoint sets are exploited to train the neural classifier and the remaining one is used to estimate the generalization error. The neural classifier is then trained Q times, each time with a different set as the validation set, and stops training at a minimum of errors on the validation set. The estimated classification error with respect to the original training set is then calculated as the mean of the Q errors.

Seven popular steganographic methods, including DCT domain embedding [58,64,65], spatial domain embedding [66–68] and wavelet domain embedding [69], were chosen to evaluate the proposed method. They sampled 132 images with 256×256 pixels as seeds to generate a total of 2088 images. Two-thirds of the image set were randomly chosen as the training set, and the remaining images as the validation set. The steganalysis system worked well for LSB-like and spreadspectrum-like steganography, and the average detection rate of cover image and stego image amounted to 80%, but suffered for wavelet domain embedding. Moreover, although it had a higher stego image detection rate (about 90%), the cover image detection rate was only about 70%.

4.2.4. *Classifier based on computational immune system*

This kind of classifier was represented in Refs. [26,27], and these papers developed CIS classifiers that were constructed using a genetic algorithm (GA) and that distinguish between clean and stego images using statistics gathered from wavelet decomposition.

A CIS is a two-class classification algorithm based on a simplified model of the human biologic immune system (BIS) [70]. The main classification mechanism of the BIS is a set of antibodies. The proper working of a BIS calls for antibodies to detect only the presence of anomalous matter (infections, cancerous cells, etc.) and trigger a defensive response. New antibodies must therefore go through a screening process, called negative selection, in which those that match against the body's own biologic uniqueness are eliminated. If such self-matching antibodies were released into circulation, they would trigger false intrusion alarms against the body's own tissues and cause the immune system to attack the body it is meant to defend. The system of antibodies is thus trained to recognize two classes: self and non-self. Instances of non-self-trigger an immune response while instances of self are ignored. The antibody creation process is blind because it trains only on instances of the class self. The CIS uses this model to create antibodies, or classification mechanisms, that ignore clean images (or self) and trigger on anything else (stego images or non-self).

The initial CIS classifiers are randomly generated and then compared to self-data. In the computational environment, self can be thought of as allowable activity and non-self can be thought of as prohibited or anomalous activity. The classifiers that match the self-data are eliminated through a process called *negative selection*. An optional process called *affinity maturation* is used to improve the quality of the classifiers before they are deployed against suspect data. When affinity maturation is complete, the classifiers are considered mature and ready for deployment against suspect data. The effectiveness of a CIS classifier lies in the matching technique used. When classifiers are compared to suspect data, an *inexact matching* function is used to ensure that the classifiers are not required to completely match the non-self, thus making them general detectors.

A GA was used in this research as the mechanism that performs affinity maturation due to the large problem space. The underlying structure of the classifiers in all stages of evolution is based on statistics gathered from wavelet decomposition of images. Negative selection eliminates classifiers that match clean and is usually performed in conjunction with the initial generation. Chromosomes represent solutions to the particular problem and consist of genes, which represent the features of a particular solution. Then, natural selection determines which solutions should be carried over into the next generation. Finally a, classifier has evolved through crossover, mutation and so on, in the generic algorithm context.

The steganalysis process of the algorithm in Ref. [26] includes the following steps:

(1) Create the clean and stego image databases, make wavelet decomposition for each image and obtain statistics of wavelet coefficients.
(2) Evolution of classifiers based on a subset of the clean wavelet coefficient statistics. The first CIS step is creating a random and initial population of classifiers, then subject the population to negative selection. For negative selection, the immature classifiers were compared to known self-points. If at least one known self-point was within the volume defined by the classifier's feature values, the classifier was discarded and a replacement was generated. This process continued until the entire initial population of classifiers did not encapsulate any known self-points.
(3) Affinity maturation using a GA to improve the randomly created classifiers by maximizing their volume or minimizing the classifier-self-distance, and then obtain the mature classifier.
(4) Test suspect image based on the mature classifier using the inexact matching function. The inexact matching function classified a suspect image point as non-self if all of the suspect image features are within the respective feature range of the mature classifier.

This method can apply to 8-bit BMP, JPEG and GIF image files. In Ref. [26], the clean images were taken with a digital camera to avoid the possibility that images downloaded from the Internet had already been stego images. The clean images were embedded messages by Ezstego, Outguess and Jsteg steganography. Results of the experiment show that the detection ratios are good if the embedding rate is higher and they significantly decrease when the secret data shrunk.

4.2.5. Classifier based on hyper-dimensional geometric

The CIS model can also be interpreted from a three-dimensional geometric standpoint. Each image instance in the CIS is defined by three features (although limited testing was done using six features) derived from image wavelet coefficient statistics. Each image instance can be represented as a point in a three-dimensional feature space. The CIS uses stochastic search techniques to create "antibodies" in the form of 3D boxes enclosing portions of the feature space. Any antibody that matches with self (i.e., a box that encloses a clean image point) is discarded. Antibodies that survive the negative selection process are retained and are used to produce a new generation of antibodies. This training process is repeated until the non-self-feature space is well-enclosed by boxes. A test image is declared to be clean if its point in the feature space is not enclosed by any of the antibody boxes. In summary, the CIS uses simple geometric constructs (boxes) to enclose the non-self-space of a class (steganography-free images) in three dimensions.

McBride et al. [28] considered there were some more powerful, versatile geometric classifiers for the following reasons as follows. First, creating a class model by enclosing the self-space rather than the non-self-space could result in a more compact class model, especially when the non-self-space is significantly larger than the self-space (or even infinite), and facilitate the construction of several co-existing class models for a hybrid signature/anomaly-based classifier rather than merely the anomaly classification of the CIS. Second, geometric structures are more versatile than boxes and can be employed to capture a wider range of complexities in a class feature space. Also, creating a geometric model deterministically rather than with a stochastic process allows for more consistent results. Third, increasing the dimensionality of the feature space to arbitrary dimensions allows for more powerful classification models generalizable to arbitrarily complex feature spaces. McBride et al. [28] focused on developing a new blind detection method, and used hyper-dimensional geometric constructs to create a blind detection model of a clean image without referencing other classes.

The instances of the training and testing sets are mapped to *d*-vectors representing points in *d*-space. Then, a class model is created by geometrically enclosing the set of class training points. A test point is declared to be a member of the class if the geometric class model encloses it. Three different geometric enclosures, including polytope, hyper-sphere and hyper-ellipsoid, are exploited to model original JPEG images so as to discriminate plain images and stego images in Ref. [28].

- *Polytope*
  A *d*-polytope is a closed geometric construct bounded by the intersection of a finite set of hyper-planes, or half-spaces, in *d* dimensions. It is the generalized form of a point, line, polygon and polyhedron in zero, one, two and three dimensions, respectively, but it is also defined for arbitrarily higher dimensions. As the number of dimensions rises, the polytope structure becomes increasingly complex and unintuitive. Generally speaking, the convex polytope is more appropriate as a classifier.
- *Hyper-sphere*
  The second classifier makes use of the generalized circle, or hyper-sphere. A *d*-sphere is a hyper-sphere in *d* dimensions that is defined simply by a center point and a radius. This construct is significantly less complex than the convex polytope, which makes it less computationally expensive and also less flexible in the kinds of shapes that it can enclose.
- *Hyper-ellipsoid*
  The third classifier employs hyper-ellipsoids. A hyper-ellipsoid in *d* dimensions is represented by three parameters that define its size (*s*: a scalar value), location (μ: a *d*-vector specifying the center point) and shape (Σ: a $d \times d$ matrix). Any point, *x*, on the ellipsoid boundary satisfies Eq. (30):

$$(x-\mu)^{\mathrm{T}}\sum\nolimits^{-1}(x-\mu)=s \tag{30}$$

  Besides, *K*-means clustering algorithm was used for dividing points into *k* disjunctive sets or clusters [71].

The authors showed a number of conditions under which a blindly created hyper-geometric class model performs well, including numeric features, discriminating features, training diversity, tight fit, granularity and noise tolerance. Good performance requires features that can be discriminated well individually or collectively between classes. Determining collective performance requires hyper-dimensional analysis. However, individual performance can be estimated before applying the features to the hyper-geometric classifier. Such estimation can be beneficial in situations that call for feature prioritization or reduction. When discriminating between classes A and B, individual feature performance can be given by computing the *J* score of feature. The *J* score is an implementation of the Fisher criterion from the FLD [52]:

$$J=\frac{(\mu_{\mathrm{A}}-\mu_{\mathrm{B}})^2}{\sigma_{\mathrm{A}}^2+\sigma_{\mathrm{B}}^2} \tag{31}$$

where μ_* and σ_*^2 are the mean and variance, respectively. A high *J* score indicates that a feature individually discriminated well between classes A and B. On the other hand, a low *J* score does not necessarily guarantee poor collective performance when the individual feature combines with other features.

A database of 1100 images was converted to grayscale. And stego images were embedded with 100%, 50%, 25% and 12.5% of the maximum steganography capacity. Compared to CIS, the hyper-geometric classifier had a slightly better false-positive rate while achieving significantly better Jsteg and Outguess detection. As expected, classification accuracy degrades as the embedded percentage shrank. The hyper-sphere and hyper-ellipsoid classifiers fail to detect F5 even at the maximum embedding rates.

4.3. Brief summary

For a steganalysis algorithm, classifier plays a significant role in the ability of the algorithm to generalize. In the domain of blind steganalysis, the feature extraction results in a decision space that the classifier then must identify a separating hyper-plane between two classes. The hyper-plane is an assumption that we are enforcing onto the space that has generalizing capabilities. An example of this in shown in Fig. 8 where the different bias assumptions (line vs. circle) have different generalizing outcomes, i.e. the linear separator classifies the gray squares as black circles, and the circles classify some of the empty squares and gray squares as black circles. This occurs even though these are not trained components.

Additionally, the accuracy of this generalization on the testing set depends on the sampling of the decision space during training. It is often assumed that the sampling of the decision space is either uniform or tailored to focus on specific areas needed to improve accuracy and reduce variance. What these "half-blind" classifiers are doing is assuming that the bias that the learning algorithm is applying to the space is strong enough and that the distribution in the feature space by using a sampling of steganography techniques made for the training set is well-enough distributed so that the trained classifier can then detect other items. This has always been the goal of machine learning, to take existing data and generalize from it to correctly classify new instances.

From the elements of various classifiers and research results of existing literatures of blind detection, we can make some conclusion as follows:

(1) The ANN performs better in steganalysis than the Bayes classifier due to its powerful learning capability. The multilayer neural network classifier outperforms the Bayes classifier, since it supplies a non-linear decision boundary.
(2) SVM has comparable performance to ANN, but the former has powerful learning capability and more efficient calculation.
(3) Among the linear separable SVM, linear non-separable SVM and non-linear SVM, the third affords the most flexible classification scheme and the best classification accuracy.
(4) Among the linear SVM, non-linear SVM, OC-SVM, the non-LINEAR SVM gives a clear advantage over the linear SVM, while the OC-SVM results in only a modest degradation in detection accuracy, when affording a simpler training stage.
(5) A classifier based on Parzen-window outperforms the OC-SVM in classification accuracy more or less.

5. Open problems and some interesting topics for research

Much progress has been made in blind steganalysis; however, new sophisticated steganographic methods will obviously require more refined detection methods. In the future, we will likely see more works addressing these important issues. We summarize previous works and conclude some open problems about the blind detection of image steganography, which are as follows:

(1) *How to decrease the influence of image contents for detection performance?*
Steganalysis methods always based on the statistical features of images, for example, image wavelet coefficients in high-frequency subbands of cover image, are well modeled by generalized Gaussian distributions (GGDs) [72]. Actually, because the contents of images are daedal, the statistical features are not completely consistent, and the description of statistical features will affect the accuracies of detection methods. It is an obvious phenomenon that we can obtain higher detection accuracy if the category of image contents in the training set is the same as that of the test set. Hence, we need to research the statistical features of various images, consider some classification techniques of image contents and design the detection methods based on the contents and the statistical features of images.
(2) *How to evaluate and select feature reliably?*
Various methods discussed above extract different numbers of image features, such as 72 features in Ref. [20], 432 features in Ref. [24], 39 features in Ref. [34], 108 features in Ref. [38] and so on. It is well known that the more the features extracted, the more the time cost of the classifier. And using too many features is undesirable in terms of classification performance due to the problem of dimensionality [52]: one cannot reliably learn the statistics of too many features given a limited training set. In addition, all features are not equally valuable to the classifier system, some features will contribute more to the detection while some others may decrease the accuracy of classification, namely maybe some consistent and afoul relations exist in the extracted features. Hence, we should make a feature evaluation, and consider the consistent and afoul relations among all extracted features, then select them before classification.
(3) *How to choose and design classifiers according to the characteristic of extracted features?*
Classifiers based on FLD, Bayes, SVM, ANN and others have been introduced into the current blind detection methods. It is well known that different classifiers with the same feature vectors may lead to different detection results. Thus, does a feature vector have relation to classifiers? Currently no paper discusses this problem. Undoubtedly that it is an important problem to be solved to choose and design the optimum classifier according to the specific features.
(4) *How to detect image in the case of low embedding ratio?*
Obviously, existing methods are far from being applied in reality. One of the main reasons is that the detection accuracies of the existing methods are not enough, especially for the case of low embedding ratio. Martín et al. [73] experimentally investigated the problem whether stego images, bearing a secret message, are statistically "natural" or not. For this

purpose, they used recent results of steganalysis methods on the statistics of natural images and investigated the effect of some popular steganography techniques. They found that these fundamental statistics of natural images are, in fact, generally altered by the hidden "non-natural" information. Frequently, the change caused by data embedding is consistently biased in a given direction. However, for the stego image that is considered as the natural image, the change falls within the intrinsic variability of the statistics, and thus does not provide reliable detection, unless knowledge of the data-hiding process is taken into account. When the embedding ratio is low, how to detect the existence of the secret message reliably is a difficult problem. For some steganalysis methods against special steganography, such as the LSB embedding method, the detection problem of low embedding ratios have been tackled by Agaian et al. [74] and Fridrich et al. [2]. Similarly, for the blind steganalysis, the problem of low embedding ratios should also be considered and discussed.

In fact, steganalysis and steganography is a cat and mouse game, and the analyzers will always be chasing the steganography developers. For example, some typical steganography methods, such as LSB, Jsteg and Outguess, can be detected by existing steganalysis methods reliably, such as the COM [31] and EM [38]. However, some new hiding method, such as YASS [82], can counter the attack of these steganalysis methods. Nowadays, image blind steganalysis is still challenging in many aspects, and we highlight some future research directions as follows.

(1) *Looking for new methods of image feature extraction*
Extract more informative features to detect the existence of secret messages embedded with most kinds of steganography methods. Although a number of features have been found out, they are not effective enough to have desirable accuracy for most embedding schemes. Some research results of the subject of image processing and texture analysis may provide some help to the feature extraction of blind steganalysis.

(2) *Identifying the embedding domain and algorithm*
Existing algorithms mainly focus on detecting the existence of secret messages, but some paid more attention to identifying the data-hiding domain and the type of steganographic algorithm. Recently, Rodriguez et al. have inspired this work, and have made some creative research on the identification of the embedding domain and algorithm. They discussed the problem of the multiclass classification of steganalysis for JPEG images in Refs. [78–80]. In future work, this problem may be able to take more efforts and improve the reliability of identification.

(3) *Identifying the image modified by steganography or normally processing operation*
Usually most images transmitted in the open channel, such as Internet, suffer from some normal processing operations, such as images splicing, stretching, smoothing, sharpening, erosion, dilation and so on. These operations always destroy the statistical characteristics of natural images, and lead to wrong classification results when we use steganalysis methods to detect them. Therefore, how to distinguish the image modified by normal image processing operation or steganography is a new challenge for steganalyzers.

6. Conclusions

In this paper, we gave an overview of the blind detection methods for image steganography. To begin with, we described the principle framework of image blind steganalysis, which includes four parts, namely, image pretreatment, feature extraction, classifier selection and design, and classification. Then we classified the existing blind detection algorithms into two categories according to the fact that the main contribution of the algorithm is feature extraction or classifier design. For the development of feature extraction of blind steganalysis, we classified the existing various methods to six categories, described briefly their principles and introduced their detailed algorithms. Then we compared the performances of seven kinds of representative methods by employing the Bhattacharyya distance between feature vectors of classes, and showed that, in the mass, the discrimination capability of features based on the CF moments of wavelet subbands coefficients is more or less the most powerful. To develop classifier selection and design blind steganalysis, we surveyed various classification algorithms used in existing blind detection methods, and especially described some classifiers based on multivariate regression analysis, OC-SVM, ANN, CIS and the hyper-geometric structure. At last, we concluded and discussed some important problems in this field, and indicated some interesting directions that may be worth researching in the future.

Steganography is the art of secret communication, and its purpose is to hide the presence of information, using digital media as covers. The main purpose of steganalysis is to discover the existence of the secret message among the covers. Through some steganalysis based on statistics methods, one can make a judgment; however, one point must be remembered, for the class of natural images considered, if the embedding ratio is so small that the change falls within the intrinsic variability of the statistics, a reliable detection is very difficult, unless knowledge of the data-hiding process is taken into account. In fact, steganalysis and steganography is just like a cat and mouse game, and the steganalyzers will always be chasing the steganography developers.

Acknowledgments

The authors wish to thank the anonymous reviewers for their insightful and invaluable suggestions and comments. This work was supported by the National Natural Science Foundation of China (Grant No. 90204013 and 60673082), National High Technology Research and

Development Program of China ("863" Program, Grant No. 2006AA10Z409), Foundation and Frontier Technology Research Program of Henan Province (Grant No. 082300410150), and the Science-Technology Project of Henan Province of China (Grant No. 0623021500).

References

[1] J. Fridrich, M. Goljan, Practical steganalysis of digital images—state of the art, in: Proceedings of the SPIE, Security and Watermarking of Multimedia Contents IV, vol. 4675, 2002, pp. 1–13.

[2] J. Fridrich, M. Goljan, R. Du, Reliable detection of LSB steganography in color and grayscale images, in: Proceedings of ACM Workshop Multimedia Security, 2001, pp. 27–30.

[3] S. Dumitrescu, X. Wu, Z. Wang, Detection of LSB steganography via sample pair analysis, IEEE Trans. Signal Process. 51 (7) (2003) 1995–2007.

[4] T. Zhang, X.J. Ping, A new approach to reliable detection of LSB steganography in natural image, Signal Process. 83 (10) (2003) 545–548.

[5] P.Z. Lu, X.Y. Luo, Q.Y. Tang, S. Li, An improved sample pairs method for detection of LSB embedding, in: Proceedings of 6th Information Hiding Workshop, Lecture Notes in Computer Science, vol. 3200, 2004, pp. 116–127.

[6] J. Fridrich, Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes, in: Proceedings of sixth Information Hiding Workshop, Lecture Notes in Computer Science, vol. 3200, 2004, pp. 67–81.

[7] T. Pevny, J. Fridrich, Toward multi-class blind steganalyzer for JPEG images, in: Proceedings of International Workshop on Digital Watermarking, Lecture Notes in Computer Science, vol. 3710, Springer, Berlin, 2005, pp. 39–53.

[8] J. Fridrich, M. Goljan, D. Hogea, Attacking the outguess, in: Proceedings of ACM Workshop Multimedia Security, 2002, pp. 3–6.

[9] J. Fridrich, M. Goljan, D. Hogea, Steganalysis of JPEG images: breaking the F5 algorithm, in: Proceedings of fifth International Workshop on Information Hiding, Lecture Notes in Computer Science, vol. 2578, Springer, Berlin, 2002, pp. 310–323.

[10] I. Avcibas, N. Memon, B. Sankur, Steganalysis using image quality metrics, IEEE Trans. Image Process. 12 (2) (2003) 221–229.

[11] PictureMarc Embed Watermark, v 1.00.45, Copyright 1996, Digimarc Corporation.

[12] M. Kutter, F. Jordan, JK-PGS (Pretty Good Signature). 1998, [Online]. Available: 〈http://www.epfl.ch/~kutter/watermarking/JK_PGS.html〉.

[13] I.J. Cox, J. Kilian, T. Leighton, T. Shamoon, Secure spread spectrum watermarking for multimedia, IEEE Trans. Image Process. 6 (12) (1997) 1673–1687.

[14] Steganos II Security Suite. [Online]. Available: 〈http://www.steganos.com/english/steganos/download.htm〉.

[15] A. Brown, S-tools version 4.0. [Online]. Available: 〈http://members.tripod.com/steganography/stego/s-tools4.html〉.

[16] J. Korejwa, Jsteg shell 2.0. [Online]. Available: 〈http://www.tiac.net/users/korejwa/steg.htm〉.

[17] I. Avcibas, N. Memon, B. Sankur, Steganalysis of watermarking techniques using image quality metrics, in: Proceedings of the SPIE, Security and Watermarking of Multimedia Contents II, vol. 4314, 2000, pp. 523–531.

[18] I. Avcibas, B. Sankur, K. Sayood, Statistical evaluation of image quality measures, J. Electron. Imaging 11 (2) (2002) 206–223.

[19] I. Avcibas, N. Memon, B. Sankur, Steganalysis based on image quality metrics, in: Proceedings of the fourth IEEE Workshop on Multimedia Signal Processing, 2001, pp. 517–522.

[20] H. Farid, Detecting steganographic messages in digital images, Technical Report TR2001-412, Dartmouth College, Hanover, NH, 2001.

[21] H. Farid, Detecting hidden messages using higher-order statistical models, in: Proceedings of IEEE International Conference on Image processing, vol. 2, 2002, pp. 905–908.

[22] H. Farid, S. Lyu, Detecting hidden messages using higher-order statistics and support vector machines, in: Proceedings of fifth International Information Hiding Workshop. Lecture Notes in Computer Science, vol. 2578, Springer, Berlin, 2002, pp. 340–354.

[23] S. Lyu, H. Farid, Steganalysis using color wavelet statistics and one-class support vector machines, in: Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI, vol. 5306, 2004, pp. 35–45.

[24] S. Lyu, H. Farid, Steganalysis using higher-order image statistics, IEEE Trans. Inf. Forensics Secur. 1 (1) (2006) 111–119.

[25] T. Holotyak, J. Fridrich, S. Voloshynovskiy, Blind statistical steganalysis of additive steganography using wavelet higher order statistics, in: Proceedings of ninth IFIP TC-6 TC-11 Conference on Communications and Multimedia Security, Lecture Notes in Computer Science, vol. 3677, Springer, Berlin, 2005, pp. 273–274.

[26] J.T. Jackson, G.H. Gunsch, R.L. Claypoole Jr., G.B. Lamont, Blind steganography detection using a computational immune system: a work in progress, Int. J. Digit. Evidence 4 (1) (2003) 1–19.

[27] J. T. Jackson, G. H. Gunsch, R. L. Claypoole, Jr., G. B. Lamont, Wavelet-based steganalysis using a computational immune system approach, in: Proceedings of the SPIE, Visual Communications and Image Processing, vol. 5150, 2003, pp. 1884–1894.

[28] B.T. McBride, G.L. Peterson, S.C. Gustafson, A new blind method for detecting novel steganography, Digit. Invest. 2 (1) (2005) 50–70.

[29] M. Goljan, J. Fridrich, T. Holotyak. New blind steganalysis and its implications, in: Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VIII, vol. 6072, 2006, pp. 1–13.

[30] B.M. Rodriguez, G.L. Peterson, S.S. Agaian, Steganography anomaly detection using simple one-class classification, in: Proceedings of the SPIE, Mobile Multimedia/Image Processing for Military and Security Applications, vol. 6579, 2007, pp. 65790E.1–65790E.9.

[31] J.J. Harmsen, W.A. Pearlman, Steganalysis of additive noise modelable information hiding, in: Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents V, vol. 5020, 2003, pp. 131–142.

[32] Y.Q. Shi, G.R. Xuan, C.Y. Yang, J.J. Gao, Z.P. Zhang, P.Q. Chai, D.K. Zou, C.H. Chen, W. Chen, Effective steganalysis based on statistical moments of wavelet characteristic function, in: Proceedings of IEEE International Conference on Information Technology: Coding and Computing, 2005, pp. 768–773.

[33] G.R. Xuan, J.J. Gao, Y.Q. Shi, D.K. Zou, Image steganalysis based on statistical moments of wavelet subband histograms in DFT domain, in: Proceedings of IEEE International Workshop on Multimedia Signal Processing, 2005, pp. 1–4.

[34] G.R. Xuan, Y.Q. Shi, J.J. Gao, D.K. Zou, C.Y. Yang, Z.P. Zhang, P.Q. Chai, C.H. Chen, W. Chen, Steganalysis based on multiple features formed by statistical moments of wavelet characteristic functions, in: Proceedings of seventh International Information Hiding Workshop, Lecture Notes in Computer Science, vol. 3727, Springer, Berlin, 2005, pp. 262–277.

[35] C.H. Chen, Y.Q. Shi, W. Chen, G.R. Xuan, Statistical moments based universal steganalysis using JPEG 2-D array and 2-D characteristic function, in: Proceedings of IEEE International Conference on Image Processing, 2006, pp. 105–108.

[36] Y.Q. Shi, G.R. Xuan, D.K. Zou, Image steganalysis based on moments of characteristic functions using wavelet decomposition, prediction-error image, and neural network, in: Proceedings of IEEE International Conference on Multimedia and Expo, 2005, pp. 269–272.

[37] Y. Wang, P. Moulin, Optimized feature extraction for learning-based image steganalysis, IEEE Trans. Inf. Forensics Secur. 2 (1) (2007) 31–45.

[38] X.C. Chen, Y.H. Wang, T.N. Tan, L. Guo, Blind image steganalysis based on statistical analysis of empirical matrix, in: Proceedings of 18th International Conference on Pattern Recognition, vol. 3, 2006, pp. 1107–1110.

[39] K. Sullivan, U. Madhow, S. Chandrasekaran, B.S. Manjunath, Steganalysis of spread spectrum data hiding exploiting cover memory, in: Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VII, vol. 5681, 2005, pp. 38–46.

[40] G.R. Xuan, Y.Q. Shi, C. Huang, D.D. Fu, X.M. Zhu, P.Q. Chai, J.J. Gao, Steganalysis using high-dimensional features derived from co-occurrence matrix and class-wise non-principal components analysis (CNPCA), in: Proceedings of International Workshop on Digital Watermarking, Lecture Notes in Computer Science, vol. 4283, Springer, Berlin, 2006, pp. 49–60.

[41] D.D. Fu, Y.Q. Shi, D.K. Zou, G.R. Xuan, JPEG steganalysis using empirical transition matrix in block DCT domain, in: Proceedings of IEEE International Workshop on Multimedia Signal Processing, 2006, pp. 310–313.

[42] W.N. Lie, G.S. Lin, A feature-based classification technique for blind image steganalysis, IEEE Trans. Multimedia 7 (6) (2005) 1007–1020.

[43] G.S. Lin, C.H. Yeh, C.C. Jay Kuo, Data hiding domain classification for blind image steganalysis, in: Proceedings of IEEE International Conference on Multimedia and Expo, 2004, pp. 907–910.

[44] E. Choi, C. Lee, Features extraction based on the Bhattacharyya distance, Pattern Recogn. 36 (8) (2003) 1703–1709.

[45] C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining Knowl. Disc 2 (1998) 121–167.

[46] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines, 2001 [Online]. Available: 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉

[47] C.M. Bishop, Neural Network for Pattern Recognition, Oxford, New York, 1995.

[48] P.P. Vaidyanathan, Quadrature mirror filter banks, M-band extensions and perfect reconstruction techniques, IEEE ASSP Mag. (1987) 4–20.

[49] R. Machado, EZStego. [Online]. Available: 〈http://www.ezstego.com〉.

[50] N. Provos, Outguess. Software available at 〈www.outguess.org〉.

[51] R. Fisher, The use of multiple measures in taxonomic problems, Ann. Eugenics 7 (1936) 179–188.

[52] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley, New York, 2001.

[53] A. Westfeld, F5. [Online]. Available: 〈http://www.wrn.inf.tu-dresden.de/westfeld/f5〉.

[54] L.M. Marvel, C.G. Boncelet Jr., C.T. Retter, Spread spectrum image steganography, IEEE Trans. Image Process. 8 (8) (1999) 1075–1083.

[55] F. Alturki, R. Mersereau, A novel approach for increasing security and data embedding capacity in images for data hiding applications, in: Proceedings of IEEE International Conference on Information Technology: Coding and Computing, 1997, pp. 228–233.

[56] CorelDraw Software. [Online]. Available: 〈http://www.corel.com〉

[57] A. Piva, M. Barni, F. Bartolini, V. Cappellini, DCT-based watermark recovering without resorting to the uncorrupted original image, in: Proceedings of IEEE International Conference on Image Processing, vol. 1, 1997, pp. 520.

[58] J.W. Huang, Y.Q. Shi, An adaptive image watermarking scheme based on visual masking, IEE Electron. Lett. 34 (8) (1998) 748–750.

[59] B. Chen, G.W. Wornell. Digital watermarking and information embedding using dither modulation, in: Proceedings of IEEE international workshop on multimedia signal processing, 1998, pp. 273–278.

[60] P. Sallee, Model-based steganography, in: Proceedings of International Workshop on Digital Watermarking, Lecture Notes in Computer Science, vol. 2939, Springer, Berlin, 2003, pp. 154–167.

[61] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (2) (1997) 153–158.

[62] A. Latham, JPEG Hide-and-Seek. [Online]. Available: 〈linux01.gwdg.de/alatham/stego〉.

[63] S. Hetzl, Steghide. [Online]. Available: 〈steghide.sourceforge.net〉.

[64] T. Ogihara, D. Nakamura, N. Yokoya, Data embedding into pictorial with less distortion using discrete cosine transform, in: Proceedings of International Conference on Pattern Recognition, 1996, pp. 675–679.

[65] W.N. Lie, G.S. Lin, C.L. Wu, Robust image watermarking on the DCT domain, in: Proceedings of IEEE International Symposium on Circuits and Systems, 2000, pp. 1228–1231.

[66] N. Nikolaidis, I. Pitas, Robust image watermarking in the spatial domain, Signal Process. 66 (3) (1998) 385–403.

[67] W. Lie, L. Chang, Data hiding in images with adaptive numbers of least significant bits based on human visual system, in: Proceedings of IEEE International Conference on Image Processing, 1999, pp. 286–290.

[68] W. Bender, D. Gruhl, N. Morimot, A. Lu, Techniques for data hiding, IBM Syst. J. 35 (3) (1996) 313–336.

[69] Y.S. Kim, O.H. Kwon, R.H. Park, Wavelet based watermarking method for digital images using the human visual system, Electron. Lett. 35 (6) (1999) 466–468.

[70] S. Forrest, S. Hofmeyr, A. Somayaji, Computer immunology, Commun. ACM 40 (10) (1996) 88–96.

[71] C. Wong, C. Chen, S. Yeh, K-means-based fuzzy classifier design, in: Proceedings of ninth IEEE International Conference on Fuzzy Systems, vol.1, 2000, pp. 48–52.

[72] S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693.

[73] A. Martín, G. Sapiro, G. Seroussi, Is image steganography natural?, IEEE Trans. Image Process. 14 (12) (2005) 2040–2050.

[74] S. S. Agaian, B. M. Rodriguez, G. Dietrich, Steganalysis using modified pixel comparison and complexity, in: Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI, vol. 5306, 2004, pp. 46–57.

[75] J. Fridrich, D. Soukal, M. Goljan, Maximum likelihood estimation of length of secret message embedded using +-K steganography in spatial domain, in: Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VII, vol. 5681, 2005, pp. 595–606.

[76] A.D. Ker, Steganalysis of embedding in two least-significant bits, IEEE Trans. Inf. Forensics Secur. 2 (1) (2007) 46–54.

[77] X.Y. Luo, C.F. Yang, D.S. Wang, F.L. Liu, LTSB steganalysis based on quartic equation, LNCS Trans. Data Hid. Multimedia Secur. (II) (2007) 68–90.

[78] T. Pevny, J. Fridrich, Towards multi-class blind steganalyzer for JPEG images, in: Proceedings of International Workshop on Digital Watermarking, Lecture Notes in Computer Science, vol. 3710, Springer, Berlin, 2005, pp. 39–53.

[79] T. Pevny, J. Fridrich, Multi-class blind steganalysis for JPEG images, in: Proceedings of SPIE, Security, Steganography, and Watermarking of Multimedia Contents VIII, vol. 6072, 2006, pp. 60720O. 1–60720O.13.

[80] B. M. Rodriguez, G. L. Peterson, S. S. Agaian, Multi-class classification averaging fusion for detecting steganography, in: Proceedings of IEEE International Conference on System of Systems Engineering, 2007, pp. 1–5.

[81] NRCS Photo Gallery. [Online]. Available: 〈http://photogallery.nrcs.usda.gov〉

[82] K. Solanki, A. Sarkar, B.S. Manjunath, YASS: Yet another steganographic scheme that resists blind steganalysis, in: Proceedings of 9th International Information Hiding Workshop, Lecture Notes in Computer Science, vol. 4567, 2007, pp. 16–31.