Multimed Tools Appl (2014) 71:497–515
DOI 10.1007/s11042-013-1522-9

Steganalysis based on distribution characters of stego-images in reduced dimension space

Guoming Chen · Qiang Chen · Dong Zhang · Duanning Zhou

Published online: 7 June 2013

Abstract In this paper we propose an Improved Kernel Linear Discriminant Analysis algorithm to analyze the distribution differences between cover images and stego-images in the reduced dimensional space. We observe that the hidden information, the information hidden in the cover images, of stego-images are clustered in a plane while all other information of cover images are scattered more evenly in the whole space and have no other clusters. Based on this fact, we develop a steganalysis scheme to discriminate stego-images from innocent images. The experiment results show the effectiveness of the propose approach.

Keywords Dimension reduction · Data distribution · Steganalysis

1 Introduction

Steganography is a technique for covert communication by embedding secret information into a cover medium such as digital images. A cover image becomes a stego-image when secret information is embedded into the cover image. As a counterpart

G. Chen (✉) · Q. Chen
Guangdong University of Education, Guangzhou, Guangdong 510310, China
e-mail: isscgm@mail.sysu.edu.cn

Q. Chen
e-mail: cq_c@gdei.edu.cn

D. Zhang
Sun Yat-sen University, Guangzhou, Guangdong 510275, China
e-mail: zhangd@mail.sysu.edu.cn

D. Zhou
Eastern Washington University, Spokane, WA 99202, USA
e-mail: dzhou@ewu.edu

of steganography, steganalysis [11, 14] is the method to detect the existence of stego-images. There are two major kinds of steganalysis: specific steganalysis that can detect a specific steganography, and blind(universal) steganalysis that can detect the existence of hidden messages without knowing details of steganography algorithms. Finding a suitable way to analyze the features of the stego-images is critical for an efficient blind steganalysis method. The Probability Density Function (PDF) moment and Characteristic Function (CF) moment are two typical statistic features frequently used in blind steganalysis algorithms.

Steganalysis can be categorized as either active or passive. Estimation of the length of embedded messages in stego-images is important to active steganalysis. Active steganalysis methods are powerful in length estimation such as in regular singular (RS) and sample pairs analysis (SPA) steganalysis schemes, but they become invalid in frequency domain. Passive steganalysis methods can discriminate stego-images from suspicious images in both spatial and frequency domains, such as in Lyu and Fraid's [6] steganalysis scheme, but they cannot estimate the length of the hidden messages.

Lou et al. [5] proposed an active steganalysis algorithm which analyzes the characteristics of histogram changes during the data embedding procedure to discriminate cover images from stego-images. Their research [5] found that the original histogram's peak will disappear and becomes concave after data embedding. This phenomenon is called "pair effect". Ding and Ping [9] proposed a steganalysis method based on the analysis of the pulse positions of histograms of cover and stego speech. They found that although the influence of steganographic embedding differs for different cover signals, the trend is that the pulse positions of the histogram becomes smoother after consistent embedding. Steganographic methods randomize the pulse positions distribution, therefore the pulse positions of the histogram for stego signal is smoother than that of the cover signal. Xuan et al. [15] proposed a novel steganalysis scheme which uses the adequate information of co-occurrence matrix to capture the changes before and after data embedding. The energy differences between the gray-level co-occurrence matrix of the original cover image and the stego-image are expected to be able to capture the changes caused by the data embedding.

Dimension reduction is a hotspot in machine learning and data mining. Traditional statistical approaches have difficulties in directly modeling data in high dimensional spaces. Dimension reduction techniques play an important role in alleviating the difficulty of high dimensional problems. Dimension reduction techniques can be categorized as linear or nonlinear methods. Linear methods are limited to discovering the structure of data lying on or near a linear subspace of the high dimensional input space. The most widely used linear dimensional reduction methods include the classic Principal Component Analysis (PCA) [8] and Linear Discriminant Analysis (LDA) [1, 18]. These methods have been applied to a wide range of signal processing problems such as feature transformation and signal analysis. A low dimensional submanifold may have a highly nonlinear structure that linear methods could fail to handle. Recently, a number of manifold learning (also referred to as nonlinear dimensionality reduction) algorithms such as ISOMAP, LLE, Laplacian Eigenmap, etc. have been proposed to overcome the limitations of linear methods. These methods have been successfully applied to a number of benchmark manifold problems and have also been proved useful in several pattern recognition applications. In the past few years, the kernel trick has also been widely applied to extend linear dimension reduction algorithms to nonlinear ones by a kernel mapping function. Linear discriminant analysis (LDA) seeks to reduce dimensionality while preserving as much of the class discriminatory information as possible. The limitation of LDA is that if the distributions are significantly non-Gaussian, the LDA projections may not preserve complex structure in the data needed for classification. Recently, the Gaussian scale mixture (GSM) [7] has been proposed to model the natural images in wavelet domain. Laplacian Eigenmaps find a low-dimensional data representation by preserving local properties of the manifold. In Laplacian Eigenmaps, the local properties are based on the pairwise distances among near neighbors. Laplacian Eigenmaps compute a low dimensional representation of the data in which the distances between a data point and its k nearest neighbors are minimized. Laplacian Eigenmaps can mostly preserve the distances among nearest neighbors while maximizing the distances among points that are not the nearest neighbors. Despite the success of the LDA algorithm in many real world applications, it still has some drawbacks in efficiency. For example, it cannot keep the intrinsic geometry property of data in most cases and has limited efficiency in classifying sample data. In order to effectively exploit favorable attributes of both LDA and Laplacian and avoid their unfavorable ones, we try to solve the steganalysis problem in the graph embedding framework. Graph embedding framework [17] can be used to develop new dimension reduction algorithms to overcome the limitations of LDA. Another important aspect of dimension reduction is that if the high dimension is reduced properly, it will show great values in pattern recognition and data visualization.

The contribution of this paper are as follows: (1) We propose a passive steganalysis scheme to analyze the characteristics of distribution changing of the dimension reduction space to discriminate the cover images from stego-images. (2) We develop a new dimension reduction algorithm, i.e., Improved Kernel Linear Discriminant Analysis (IKLDA), to extract the hidden information of stego-images and find that they are clustered in a plane while cover images are scattered more evenly and have no other clusters. The rest of the paper is organized as follows: Section 2 introduces the IKLDA algorithms; Section 3 presents our experimental results; and Section 4 concludes the paper.

2 Algorithm

2.1 LDA algorithm and its improvement

LDA is a supervised method. It searches the project axes on which the data points of different classes are as far from each other as possible while the data points of the same class are as close to each other as possible. Suppose we have a set of samples $x_1,x_2,\ldots x_l \in R^d$, which belongs to C classes, where d is the original dimension. Regarding supervised learning problem, let $z\in(1,2,\ldots,C)$ be a class label, where C is the number of classes as mentioned above. Let X be the matrix representation of the whole sample set, i.e., each sample is treated as a column of X. Let $Y\in R^r(1\leq r\leq d)$ be the projection of X, where r is the dimension of the lower dimensional space. We first consider the linear dimension reduction methods. We define a $d \times r$ transformation matrix α such that the low dimensional data representation Y is given by: $Y = \alpha^T X$.

The goal of LDA is to look for a transformation matrix α to characterize the intra class compactness and the inter class separability, i.e., to find α_{lda} such that

$$\alpha_{\text{lda}} = argmax \frac{\alpha^T S_b \alpha}{\alpha^T S_w \alpha} \tag{1}$$

$$S_b = \sum_{k=1}^{c} l_k (u^k - u)(u^k - u)^T \tag{2}$$

$$S_w = \sum_{k=1}^{c} \left(\sum_{i=1}^{l_k} ((x_i^k - u^k)(x_i^k - u^k)^T) \right) \tag{3}$$

$$S_t = \sum_{i=1}^{l} (x_i - u)(x_i - u)^T \tag{4}$$

$$S_t = S_b + S_w$$

where u is the total sample mean vector, l_k is the number of samples in the k-th class, u^k is the average vector of the k-th class and x_i^k is the i-th sample in the k-th class, S_w is called intra class scatter, a metric to measure intra-distances and S_b is called inter class scatter, a metric to measure the inter-distances. So, the purpose of LDA is to find a transformation matrix α_{lda} to maximize the linear separability of data points belonging to the different classes while minimize the linear compactness of data points within the same classes.

The maximization problem of obtaining α_{lda} in (1) can be converted to solve the following generalized eigenproblem:

$$S_b \phi = \lambda S_w \phi$$

Let $\{\phi_k\}_{k=1}^{d}$ be the eigenvectors of the generalized eigen problem corresponding to the eigenvalue λ_k, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. Then, the eigenvectors $[\phi_1, \phi_2 \ldots \phi_d]$ form the columns of the linear transformation matrix α_{lda} and Y can be computed by mapping X onto the linear basis matrix α_{lda}. That is,

$$Y = \alpha_{\text{lda}}^{\text{T}} X$$

In order to improve LDA on the ground of graph embedding framework, intra class scatter S_w can be transformed to pairwise expressions. As u^k is the average vector of the k-th class, i.e.,$u^k = \frac{1}{l_k} \sum_{j=1}^{l_k} x_j^k$, $(u^k)^T = \frac{1}{l_k} \sum_{i=1}^{l_k} x_i^k$, from (3), we obtain:

$$\begin{aligned}
S_w &= \sum_{k=1}^{c} \left(\sum_{i=1}^{l_k} \left((x_i^k - u^k)(x_i^k - u^k)^T \right) \right) = \sum_{k=1}^{c} \sum_{i=1}^{l_k} x_i^k (x_i^k)^T - \sum_{k=1}^{c} \frac{1}{l_k} \sum_{i,j=1}^{l_k} x_i^k (x_j^k)^T \\
&= \frac{1}{l_k} \sum_{k=1}^{c} \sum_{i,j=1}^{l_k} x_i^k (x_i^k)^T - \frac{1}{l_k} \sum_{k=1}^{c} \sum_{i,j=1}^{l_k} x_i^k (x_j^k)^T \\
&= \frac{1}{2 \times l_k} \sum_{i,j=1}^{l} \left(x_i (x_i)^T + x_j (x_j)^T - x_i (x_j)^T - x_j (x_i)^T \right) \\
&= \frac{1}{2 \times l_k} \sum_{i,j=1}^{l} (x_i - x_j)(x_i - x_j)^T
\end{aligned} \tag{5}$$

For the same reason, inter class scatter S_b has also a pairwise expression. As μ is the total sample mean vector, i.e., $\mu = \frac{1}{l}\sum_{j=1}^{l} x_j$, $\mu^T = \frac{1}{l}\sum_{i=1}^{l} x_i$. From (2), we have:

$$\begin{aligned} S_b &= S_t - S_w \\ &= \sum_{i=1}^{l}(x_i-\mu)(x_i-\mu)^T - S_w \\ &= \sum_{i=1}^{l} x_i {x_i}^T - \frac{1}{l}\sum_{i,j=1}^{l} x_i {x_j}^T - S_w \\ &= \frac{1}{l}\sum_{i,j=1}^{l} x_i {x_i}^T - \sum_{i,j=1}^{l} x_i {x_j}^T - S_w \\ &= \frac{1}{2\times l}\sum_{i,j=1}^{l}(x_i-x_j)(x_i-x_j)^T - \frac{1}{2\times l_k}\sum_{i,j=1}^{l}(x_i-x_j)(x_i-x_j)^T \end{aligned} \tag{6}$$

In the process of deduction of (5) and (6), a connotative condition that x_i and x_j belong to the same class has been added, because neighbor points are mostly possible in the same class labels. In graph embedding framework, we combine the advantage of Laplacian method that nearby points remain nearby and far apart points remain far apart in dimension reduction and the advantage of LDA method that concentrating the points of intra-classes and repulsing the points of inter-classes. To this end, we employ the matrix representation of the graph G which takes the samples $x_i (i = 1\ldots l)$ as its vertices. We denote by W=($W_{i,j}$) the representation matrix of G, i.e., $W_{i,j}$ =1 if x_i and x_j are neighbors; otherwise, $W_{i,j} = 0$. That is,

$$W_{i,j} = \begin{cases} 1 & if x_i \in N_k(x_j)\ or\ x_j \in N_k(x_i) \\ 0 & else \end{cases} \tag{7}$$

where $N_k(x_i)$ denotes the set of k nearest neighbors of x_i. Equation (7) can be upgraded to Euclidean distance based on Gaussian distribution. $W_{i,j}$ is defined by:

$$W_{i,j} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|^2}{\delta_i \delta_j}\right) & if x_i \in N_k(x_j)\ or\ x_j \in N_k(x_i) \\ 0 & else \end{cases} \tag{8}$$

where $\delta_i = \|x_i - x_i(k)\|$, in which $x_i(k)$ is the k-th nearest neighbors of x_i. The diagonal matrix D and the Laplacian matrix L of a graph G can be defined as:

$$L = D - W;\ D_{ii} = \sum_{i \neq j} W_{i,j} \tag{9}$$

The idea of the algorithm of graph embedding framework is to find an appropriate low dimensional representation which can keep the neighborhood property of the vertices of the graph G. Let Y=[y$_1$,y$_2$,...y$_{l_k}$]^T be the low dimensional space to be found, where y_i is the low dimensional representation of vertex x_i. The data points in embedding space should have the same geometric properties as that of the original data. For example, the k-nearset points of y_i in the embedding space should corresponding to the k-nearset points of x_i. As consequence, the nearby points in the high dimensional space are also projected to nearby points in the low dimensional representation. In fact, a rough low dimensional representation can be obtained by:

$$Y = \underset{y^T By = d}{\operatorname{argmin}} \sum_{i \neq j} \|y_i - y_j\|^2 W_{i,j} = \underset{y^T By = d}{\operatorname{argmin}} y^T L y \tag{10}$$

where d is a constant and B is a constraint matrix to avoid multi-solutions. To obtain a better result, we need to find a more suitable projection matrix, still denoted by w, which satisfies Y= x^Tw. w can be obtained by optimizing the following equation.

$$\begin{aligned} W = & \underset{w^T x B x^T w = d}{\operatorname{argmin}} \sum_{i \neq j} \|w^T x_i - w^T x_j\|^2 W_{i,j} \\ = & \underset{w^T x B x^T w = d}{\operatorname{argmin}} w^T x L x^T w \end{aligned} \tag{11}$$

where x is the matrix of taking the samples $x_i (i = 1 \ldots l)$ as its columns as mentioned above.

2.2 Kernel LDA algorithm

The algorithms mentioned above are linear methods which are usually not good for the classification problems with nonlinearly distributed data. Therefore, it is necessary to introduce a new kernel trick to handle data with nonlinear distributions. In machine learning, the use of the kernel functions has been introduced to find a close-to-optimal projection based on different sample distributions. The kernel matrix implicitly maps the data into a nonlinear feature space. The choice of the kernel is crucial to incorporate a priori knowledge in application. Each kernel can be expressed as: k(x,y)=$< \phi(x), \phi(y) >$, in which $< \phi(x), \phi(y) >$ is the scalar product,where $\phi(x)$ is a dimensional elevating mapping. We call the dimensional elevated space as the Reproducing Kernel Hilbert Space (RKHS). Examples of kernels are as follows:

$$\begin{cases} Gaussian : k(x, y) = exp\left(-\frac{\|x-y\|^2}{2\delta^2}\right) \\ Polynomial\ with\ degree\ d : k(x, y) = (c + \langle x, y \rangle)^d \\ Sigmoid : k(x, y) = \tanh(\langle x, y \rangle + \alpha) \end{cases} \tag{12}$$

In this paper, we choose Gaussian kernel k(x,x')= exp $\left(-\frac{\|x-x'\|^2}{2\delta^2}\right)$, $\delta > 0$ as the kernel function and denote as $k_{i,j} = k(x_i, x_j) =< \phi(x_i), \phi(x_j) >$.

To obtain an even better projection matrix, we consider the following optimization problem:

$$\begin{aligned} \beta = & \underset{w^T k B k^T w = d}{\operatorname{argmin}} \sum_{i \neq j} \|\beta^T k_i - \beta^T k_j\|^2 W_{i,j} \\ = & \underset{w^T k B k^T w = d}{\operatorname{argmin}} w^T k L k^T w \end{aligned} \tag{13}$$

where k_i indicates the i-th column vector of the kernel gram matrix K and w is defined in (11). The solutions of (13) can be converted to the following generalized eigenvalue problem:

$$KBK\beta = \lambda KLK\beta \tag{14}$$

where L is defined in (9). To obtain the weight coefficient S_w^{weight} of intra class scatter matrix, we need to optimize S_w in (5) and the expression of S_w^{weight} is as follows:

$$S_w^{\text{weight}} = \frac{1}{l_k} \tag{15}$$

Similarly, optimizing (6) we can obtain the solution of S_b as follows:

$$S_b^{\text{weight}} = \frac{1}{l} - \frac{1}{l_k} \tag{16}$$

We cannot directly solve the generalized eigenproblem of (14), since it is ill posed as mentioned in KLK [12]. Instead, we solve its following regularization problem:

$$KBK\beta = \lambda(KLK + \varepsilon I)\beta \tag{17}$$

where ε is a constant with small value. As mentioned above, the core idea of our IKLDA is to combine the advantages of LDA and Laplacian method. That is, the advantage that the nearby points remain nearby and far apart points remain far apart in dimension reduction and the advantage that concentrating the points of intra-classes and repulsing the points of inter-classes. In fact, the former advantage means that the intrinsic geometric property of the data is well preserved during dimension reducing. In addition, the weights $W_{i,j}$ defined in (8) means that the influence of the data on S_w^{weight} and S_b^{weight} may be reduced as the increase of the distances among the points of the data.

Many improvements of LDA algorithms are based on the redefinitions of weighting factors of S_w^{weight} and S_b^{weight}. For example, Loog et al. [4] proposed a criterion called approximate pairwise accuracy (aPAC); Sugiyama [12] proposed the LFDA algorithm by setting $S_w^{'} = \frac{1}{l_k} S_w, S_b^{'} = \frac{1}{l} S_b + (1 - \frac{1}{l_k}) e_l e_l^T + \frac{1}{l} e_{lk} e_{lk}^T$; Yan et al. [16] proposed a cluster algorithm, named ICBKM, by setting $S_w^{'} = I - \sum_{k=1}^{c} \frac{1}{l_k} e_{lk} e_{lk}^T, S_b^{'} = \sum_{k=1}^{c} \frac{1}{l_k} e_{lk} e_{lk}^T - \frac{1}{l} e_l e_l^T$

According to (15)–(17), we redefine the weight factors S_w and S_b as follows:

$$\begin{cases} S_w^{'} = \alpha \sqrt{\frac{1}{l_k}} KLK & where\ 0 < \alpha < 1;\ D_{ii} = \sum_{i \neq j} W_{i,j};\ L = D - W \\ S_b^{'} = (1 - \alpha) \left(\sqrt{\frac{1}{l} - \frac{1}{l_k}} \right) KLK & same\ as\ above \end{cases} \tag{18}$$

Equation (17) can be expressed as follows:

$$S_b^{'} \beta = \lambda(S_w^{'} + \varepsilon I)\beta \tag{19}$$

Form the above (8), (9), (12), (18), (19), we obtain the following optimization equation of our IKLDA method:

$$(1-\alpha)\left(\sqrt{\frac{1}{l}-\frac{1}{l_k}}\right)KLK\beta=\lambda\left(\alpha\sqrt{\frac{1}{l_k}}KLK+\varepsilon I\right)\beta \tag{20}$$

where K= $\exp(-\frac{\|x-x^{'}\|^2}{2\delta^2})$, L is defined in (9).

Let $\{\beta_i\}_{i=1}^{r}$ be the eigenvectors of (20) corresponding to the leading eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$. The reduced dimension Y of the data points in X is thus can be given by Y= $\beta^T K$.

For the sake of completeness we will use correlation dimension to describe the assumed inner structure. Intrinsic dimensionality is the minimum number of parameters that is necessary account for all information in the data. It is hard to say how many dimensions are appropriate to describe the data set in real world applications. The fewer dimensions, the fewer properties that may fall short of describing the abundance of images, and therefore may lead to misclassification. However, the more the dimensions the feature vector has, the higher probability the redundancy and dependency exists. Usually, such redundancy or dependency of the feature vector tends to induce misclassification. Techniques for intrinsic dimensionality estimation can be divided into two groups: those based on local properties, and those based on global properties of the data. Local intrinsic dimensionality estimators are based on the following principle: the number of data points covered by a hypersphere around a data point with radius *r* grows proportional to r^d, where *d* is the intrinsic dimensions. So *d* can be estimated by measuring the number of data points covered by a hypersphere with a growing radius *r*.

Correlation dimension estimator [2] is roughly one kind of local estimator, the relative amount of data points in a hypersphere with radius *r* are as follows:

$$C(r)=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{c} c \ where \ c=\begin{cases}1 & if\|x_i-x_j\|\leqslant r\\ 0 & if\|x_i-x_j\|>r\end{cases} \tag{21}$$

We use C(*r*) to estimate the dimensions *d*:

$$d=\lim_{r\to 0}\frac{\log C(r)}{\log r} \tag{22}$$

It is pretty hard to solve (22), so we use the following expression instead:

$$\bar{d}=\frac{\log(C(r_2)-C(r_1))}{\log(r_2-r_1)} \tag{23}$$

By (23) we calculate the intrinsic dimensions and get 2.91741 in the steganalysis data set.

3 Experimental results

The purpose of this section is to verify the efficiency of our IKLDA method in steganalysis by observing the difference of the distributions between stego-images and cover images of the dimension reduced data. The experimental results show that the detective rate of steganalysis is also increased.

To this end, we implement two groups of experiments. The first group involves five data hiding methods and the second includes one data hiding method. We take 1096 sample images from different picture sets, such as Nature, Ocean, Food, Animals, Architecture, Places, Leisure, Misc, of CorelDRAW Version 10.0 software CD♯3 to complete the first group of experiments. The following five typical data hiding methods [11] are used in our first group of experiments: Cox et al.'s non-blind SS, Piva et al.'s blind SS, Huang and Shi's 8 by 8 block SS, a generic QIM (0.1 bpp(bit per pixel)), and a generic LSB(0.3 bpp, both the pixel position used for embedding data and the to-be-embedded are randomly selected). For each sample cover image, five stego-images are generated with these five data hiding methods, respectively. For all the data hiding methods, different random signals are embedded into different images. The evaluation of the proposed steganalysis system is hence more general.

Shi et al. [11] use the statistical moments of the characteristic functions of wavelet subbands as features for steganalysis. The characteristic function (CF) and the PDF (here, histogram) are similar to a Fourier transform pair. We denote the histogram by h(x_j), and the characteristic function by h(f_k),The n-th statistical moment of a characteristic function M_n is defined as follows:

$$M_n = \frac{\sum_{k=1}^{N/2} f_k^n |H(f_k)|}{\sum_{k=1}^{N/2} |H(f_k)|}$$

where $H(f_k)$ is the magnitude of the CF.

In order to handle the noise introduced by data hiding, Shi et al. [11] proposed to predict each pixel grayscale value in the original cover image by using its neighboring pixels' grayscale values, and obtain the prediction-error image by subtracting the predicted image from the test image. If the hidden data are unrelated to the cover media, the prediction-error image can remove all other informations other than that caused by data hiding and this makes the steganalysis more efficient. Fortunately, the hidden data are usually unrelated to the cover media.

In this first group of experiments, each test image is applied with Haar wavelet transformation three times to obtain a 3-level decomposition. For each level, there are four subbands, resulting in 12 subbands in total. If the original image is referred to level-0 LL subband, we have a total of 13 subbands. For each subband, the first three moments of characteristic functions are derived, resulting in a set of 39 features. Similarly, for the prediction-error image, another set of 39 features can be generated. Thus, a 78-Dimention feature vector is produced for a test image. In fact, the experiments show that using more than three-level wavelet decomposition and employing more than the first three order moments cannot further improve the steganalysis performance other than leading to higher computational complexity. Hence the 78-Dimention feature vectors are used in this proposed steganalysis system.

We use the Improved Kernel Linear Discriminant Analysis algorithm to project the sample image data onto R^3 which is good enough for image steganalysis. Figures 1, 2, 3, 4 and 5 show the spatial distributions obtained from IKLDA to capture the statistical differences between cover images and stego-images. All the results show that the distributions of the projected stego-images are exactly clustered

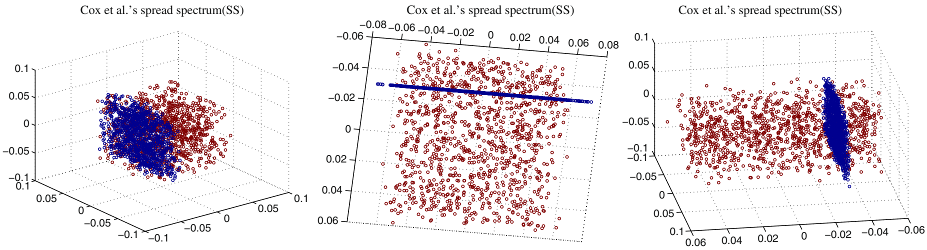


Fig. 1 Cox et al.'s spread spectrum

in a plane, respectively, while the distributions of the projections of all cover images are scattered more evenly and have no other clusters. Based on this distribution analysis, we get a 100 % detective rate after further projecting the projected data in R^3 into one dimensional space.

Any image is decomposed into four subimages after applying a wavelet transformation and the four subimages are called horizontal, vertical, diagonal and lowpass subbands, respectively, and denoted by H-subband, V-subband, D-subband and L-subband, respectively.

In the second group of experiments, we also test IKLDA for other data sets. To complete the second group of experiments, we choose a number of 1813 gray-scale JPEG images which are 256 × 256 pixel and are embedded data by using steghide [10](a steganography method that is able to hide data in various kinds of images). We first fuse the feature space for each data set by two different methods to produce a total of 150 features. The first 72 features are from the method of Lyu and Farid [6] and the remaining 78 features are from the method of Shi et al. [11]. The first 72 features include the mean, variance, skewness and kurtosis of the subbands H1, V1, D1, H2, V2, D2, H3, V3, D3, EH1, EV1, ED1, EH2, EV2, ED2, EH3, EV3, ED3, where H1, V1, D1, H2, V2, D2, H3, V3, D3 are the H-subbands, V-subbands and D-subbands obtained by applying wavelet transformation three times and EH1, EV1, ED1, EH2, EV2, ED2, EH3, EV3, ED3 are corresponding to that of the prediction-error image. We arrange these 72 features as follows: 1:meanV, the mean of V1, 2:meanH, 3:meanD, 4:varV, 5:varH, 6:varD, 7:kurV, 8:kurH, 9:kurD, 10:skeV, 11:skeH, 12:skeD, 13:meanEV, 14:meanH, 15:meanED, 16:varEV, 17:varEH, 18:varED, 19:kurEV, 20:kurEH, 21:kurED, 22:skeEV, 23:skeEH, 24:skeED. Similarly, the corresponding

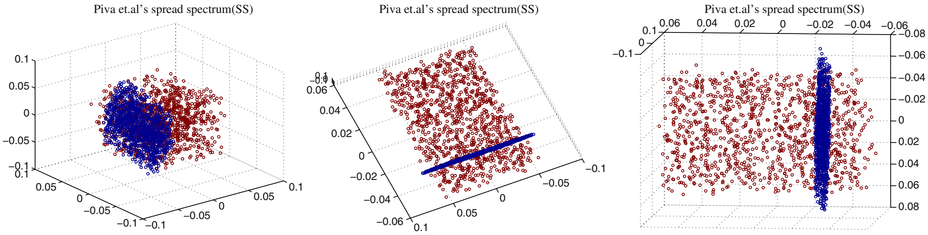


Fig. 2 Piva et al.'s spread spectrum

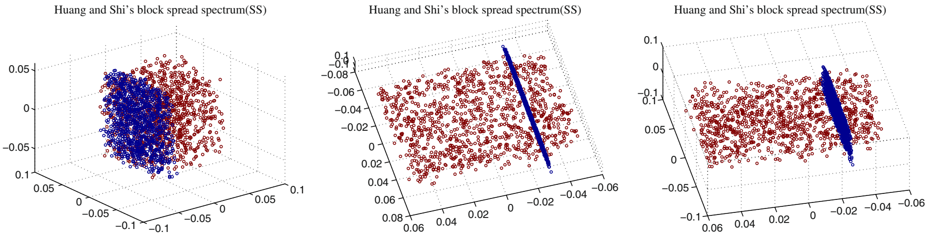


Fig. 3 Huang and Shi's spread spectrum

values of the second level and third level are put into components from 25th to 48th and from 49th to 72nd, respectively. The rest 78 features are as follows: 1st–3rd features are the mean, variance and kurtosis of the original images; 4th–6th features are the mean, variance and kurtosis of the prediction-error image; from 7th to 42nd is followed by the means, variances and kurtosises of the H-subbands, V-subbands, D-subbands and L-subbands LLi, HLi, LHi, HHi, i=1,2,3, obtained by applying Haar wavelet transformation three times, say meanLL1, varLL1, kurLL1, meanHL1, varHL1, kurHL1, meanLH1, varLH1, kurLH1, meanHH1, varHH1, kurHH1, meanLL2, varLL2, kurLL2, meanHL2, varHL2, kurHL2, meanLH2, varLH2, kurLH2, meanHH2, varHH2, kurHH2, meanLL3, varLL3, kurLL3, meanHL3, varHL3, kurHL3, meanLH3, varLH3, kurLH3, meanHH3, varHH3, kurHH3; finally, from 43rd to 78th positions the values of the prediction-error image, corresponding to that from 7th to 42nd.

We index +1 for any original image and −1 for any stego-image. The following experiments show the efficiency of IKLDA for steganalysis after projecting the above 150 features onto R^3, by comparing the difference between the cover image and the stego-image, obtained by steghide, in the reduced dimensional space.

In order to show how IKLDA works, we consider two of its important parameters ϵ and α in (20) and their functions of impacting on the distributions of features of the cover images and stego-images in the reduced dimension. ϵ is a small constant. By fixing α and changing ϵ, we get feature clouds for steganalysis: from Fig. 6a–f we can observe that the less value of ϵ,the farther between the inter-classes and the closer in intra-classes. But with ϵ decreasing, a downturn appears; after that, a better result shows up again.

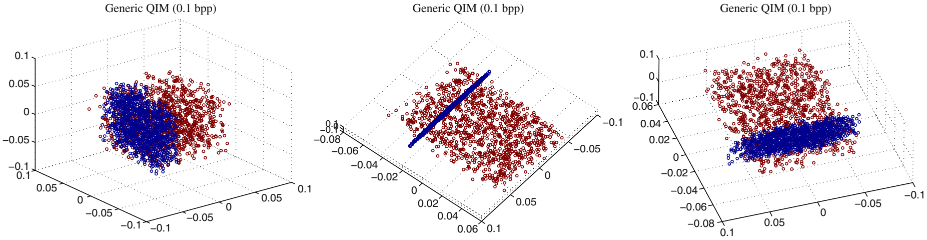
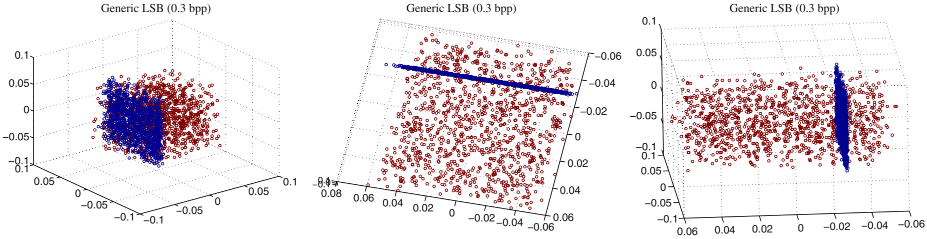
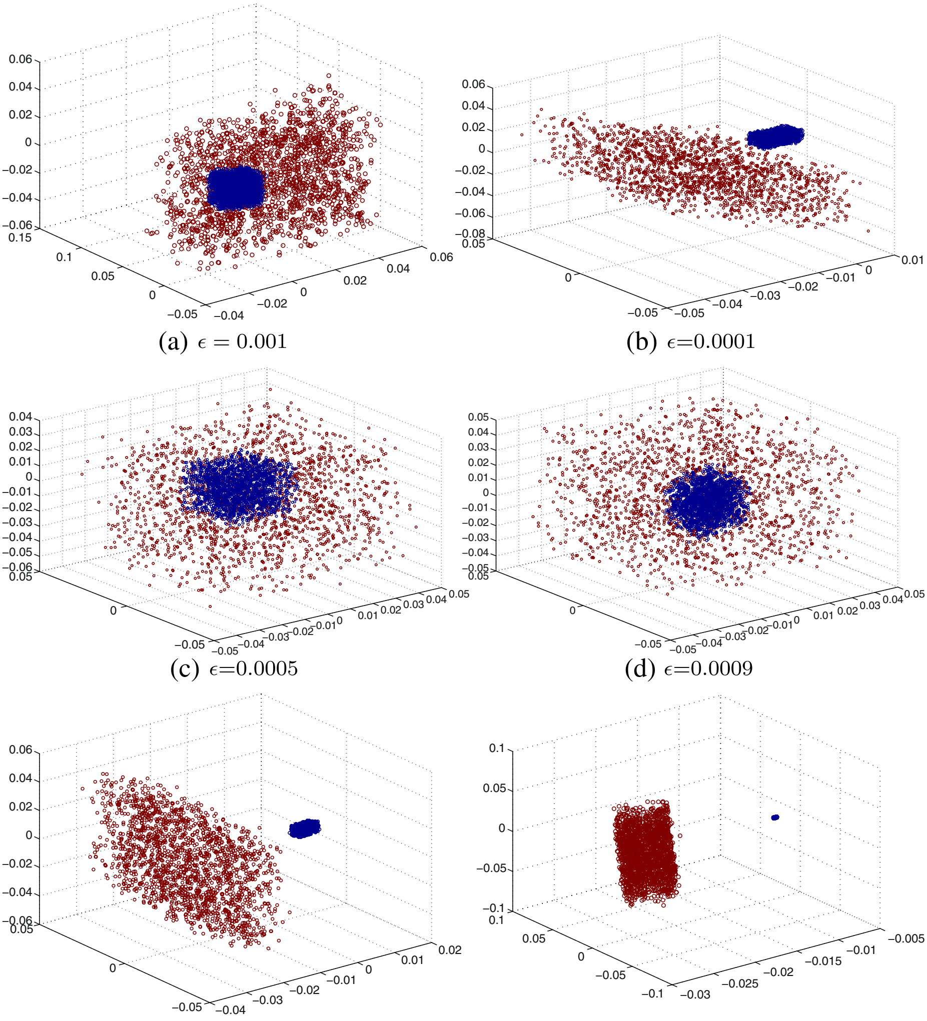


Fig. 4 Generic QIM(0.1 bpp)



Generic LSB (0.3 bpp) Generic LSB (0.3 bpp) Generic LSB (0.3 bpp)

Fig. 5 Generic LSB(0.3 bpp)

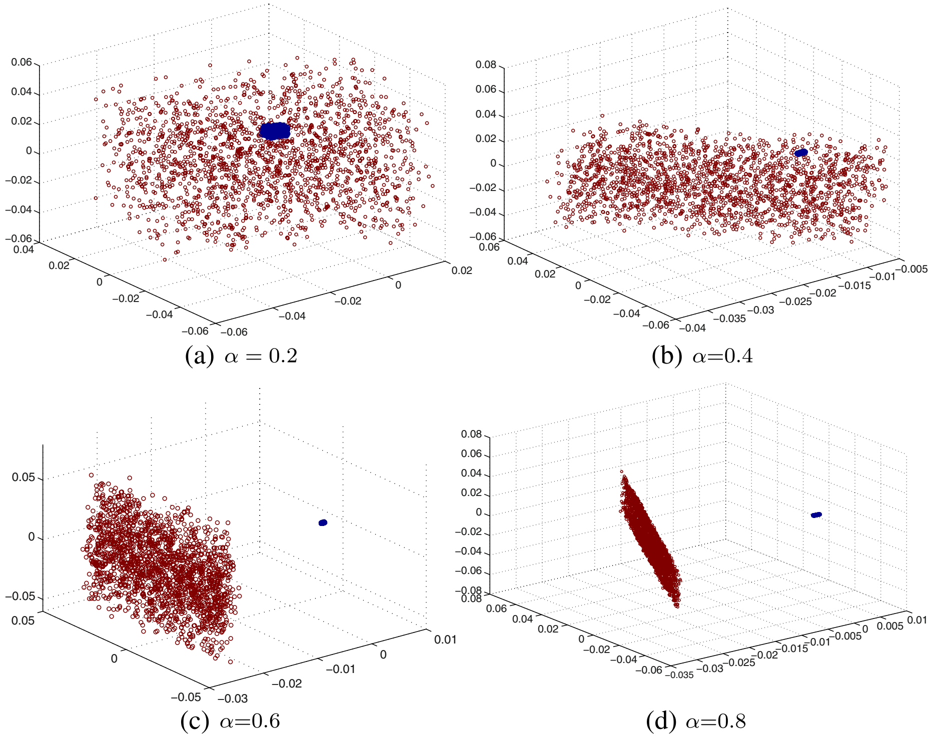


(a) $\epsilon = 0.001$ (b) ϵ=0.0001

(c) ϵ=0.0005 (d) ϵ=0.0009

(e) ϵ=0.00001 (f) ϵ=0.000001

Fig. 6 The distributions of IKLDA with the changing of parameter ϵ



(a) $\alpha = 0.2$ (b) α=0.4

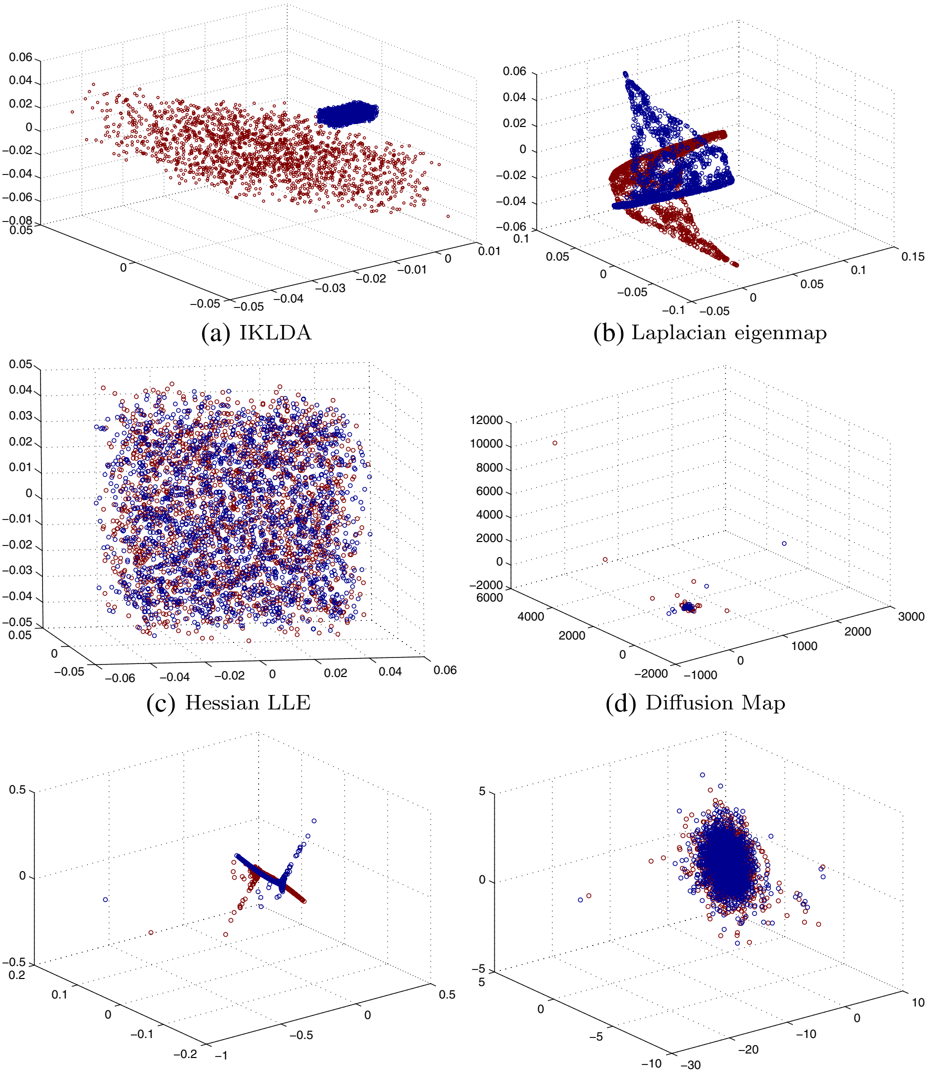
(c) α=0.6 (d) α=0.8

Fig. 7 The distributions of IKLDA with the changing of parameter α

Similarly, Fig. 7a–d are obtained by fixing ϵ and changing α in (20). The feature clouds of Fig. 7a–d show that, with the increasing of α, the samples in one intra-class are getting closer while samples in the other intra-class are getting farther. The samples between inter-classes are also getting farther.

In the second group of experiments, we still compare IKLDA with other dimension reduction methods [13] to show the efficiency of IKLDA. The goal is to find out which method can map the input data into a feature space in which samples from different classes can be clearly separated.

The possibilities of distributions can be classified into three cases: (1) the distances of the samples among inter-classes are maximized while that of intra-classes are maximized as well; (2) the distances of the samples among inter-classes are maximized while that of intra-classes are minimized; (3) the distances of the samples among inter-classes are minimized while that of intra-classes are minimized. From the following experiments we can get their corresponding distributions where Fig. 8a is the reduced dimension space obtained from IKLDA($\epsilon = 0.0001$) which can maximize the distances of the samples among inter-classes while minimize that of intra-classes. Figure 8b is the reduced dimension space obtained from Laplacian eigenmap which can get a nice symmetry geometry structure but fail in maximizing the distances of the samples among inter-classes. Figure 8h is the reduced dimension



(a) IKLDA

(b) Laplacian eigenmap

(c) Hessian LLE

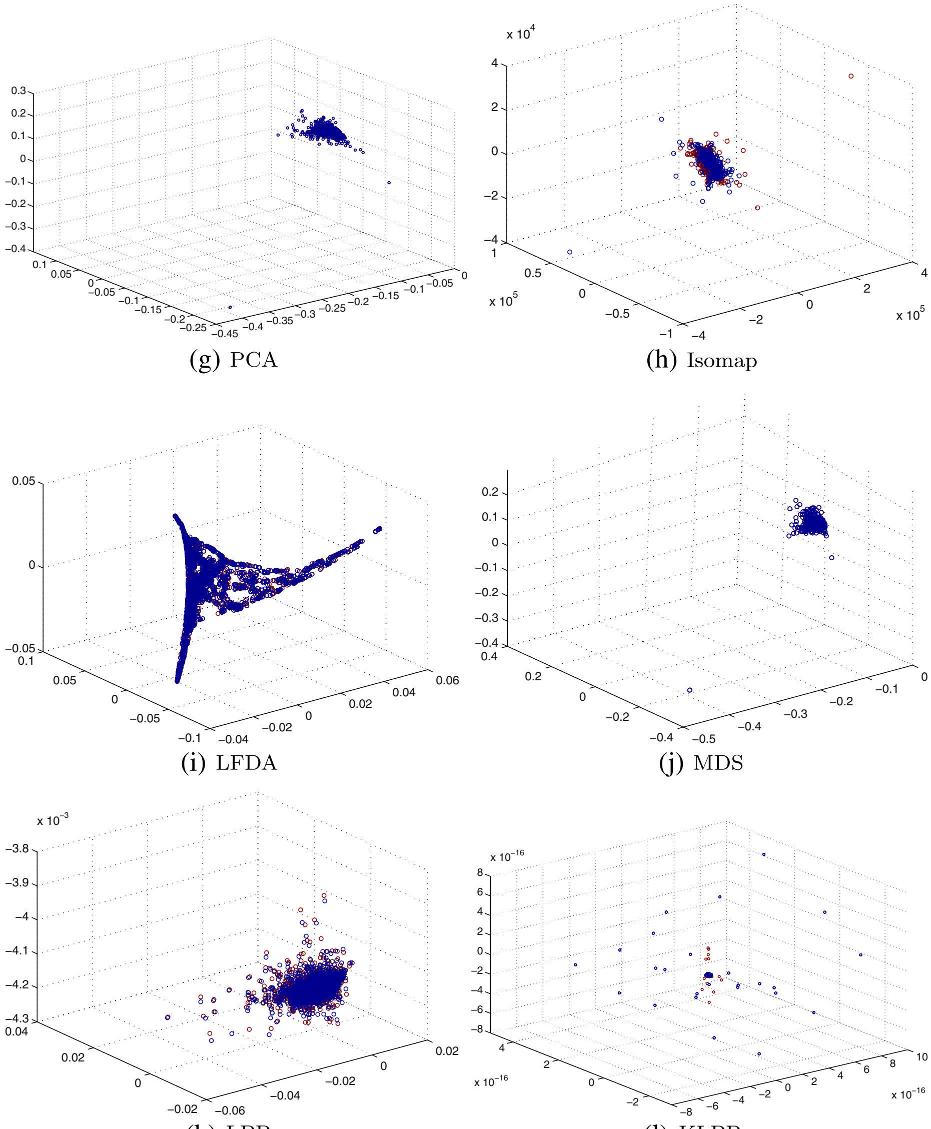
(d) Diffusion Map

(e) LTSA

(f) LDA

Fig. 8 The distributions of different dimension reduction methods

space obtained from Isomap, Fig. 8d is the reduced dimension space obtained from Diffusion Map and Fig. 8e is the reduced dimension space obtained from LTSA, they show that their methods can minimize the distances of the samples among intra-classes but fail in maximizing that of inter-classes. Figure 8c is the reduced dimension space obtained from Hessian LLE, Fig. 8f is the reduced dimension space obtained from LDA, Fig. 8g is the reduced dimension space obtained from PCA, Fig. 8i is the reduced dimension space obtained from LFDA, Fig. 8j is the reduced dimension space obtained from MDS, Fig. 8k is the reduced dimension



(g) PCA

(h) Isomap

(i) LFDA

(j) MDS

(k) LPP

(l) KLPP

Fig. 8 (continued)

space obtained from LPP and Fig. 8l is the reduced dimension space obtained from LPP's kernel extension called KLPP, they cannot get favorable results either. The distributions of visualization experiments show that our new IKLDA algorithm gets better performance in our steganalysis scheme and is better than the other traditional dimension reduction methods.

Most important of all, we should investigate the way the reduced dimensional space by our improved kernel linear discriminant analysis algorithm will work for the image steganalysis. We compare this phenomenon with that of Kocal's to the best of our knowledge. Since Koçal et al. [3] found chaotic-type features for speech steganalysis where Lyapunov exponents (LE) and fraction of false neighbors (FNFS) have been used as chaotic features to detect the existence of the speech stego signal. The rest may be deduced by analogy, considering that data hiding within a speech signal can distort the chaotic properties of the original speech signal. Although there is no direct evidence for the existence of chaotic phenomena in image signals, which linear modeling cannot cover, there should be self-similar distribution in typical natural images. We would like to call this inner structure a possible fractal structure. Data hiding within image signals has distorted this inner structure and characteristic distribution of these signals, resulting in a change in distribution characters of the reduced dimension. Exploring this change made by data hiding can lead to the design of a learning-based steganalyzer.

Furthermore, the data-hiding affects the neighborhood distances between cover and stego-images in the proposed reduced dimension. The details of the data hiding effects can be shown in the above mentioned experiments. In speech steganalysis, some similar patterns come into effect that there is more significant distinction between cover and stego speech signals in phase space representation than in time-series representation. Accompanied with the image steganalysis, the reduced dimensional feature can uncover more useful information if proper reduction algorithm has been proposed.

In order to go a step further and explore our steganalysis schema in another way, we take speech steganalysis in phase space for example. The Lyapunov exponent, a quantitative measure for the divergence of nearby trajectories, should be calculated for all of the nearest neighbor pairs on different trajectories. A positive exponent means that the trajectories, which are initially close to each other, move apart over time (divergence); while for negative exponents, the trajectories move closer to each other (convergence); the magnitude of the exponent determines how rapidly the trajectories move. In image steganalysis our IKLDA is to maximize the distances of the samples among inter-classes and minimize that of intra-classes in the proposed reduced dimension space, and the key parameters in (20) act as the ratio to control how well this process will go, just as the Lyapunov exponent does in speech steganalysis.

As explained above, a new steganalysis scheme is proposed which focuses on the alteration and differences of statistical self-similarity features of stego-images which are formed by embedding algorithms, and has something to do with the distribution characters of stego-images. We believe that the proposed reduced dimensional space can act as characterization and modeling of image steganalysis and have the same effect as the phase-space used to distinguish stego-signals from cover signals in speech steganalysis.

4 Conclusions

In this paper, we propose an Improved Kernel Linear Discriminant Analysis algorithm to analyze the distribution difference between the cover image and the stego-image in the dimensional reduced space. We observe that after dimension reduction the steganographically embedded hidden information in stego-images are clustered in a plane while all other information of cover images are scattered more evenly in this space and have no other clusters. Based on this fact, we develop a passive steganalysis scheme to discriminate stego-images from innocent images. The experiment results verify the effectiveness of the propose approach.

Acknowledgements This work is partially supported by the State Scholarship Fund (2007 104752), National Natural Science Foundation of China (61100170), Guangdong Province ordinary university science and technology innovation project(2012KJCX0079), Guangdong Modern Information Service Industry Develop Particularly item (No.13090), the Fundamental Research Funds for the Central Universities (12lgpy37).

References

1. Cai D, He X, Han J (2007) Semi-supervised discriminant analysis. In: Proceedings of 11th IEEE International Conference on Computer Vision (ICCV), pp 1–7
2. Kawaguchi A, Yonemoto K, Yanagawa T (2005) Estimating the correlation dimension from a chaotic system with dynamic noise. Japan Statist 35(2):287–302
3. Koçal O, Yuruklu E, Avcibas I (2008) Chaotic-type features for speech steganalysis. IEEE Trans Inf Forensics Secur 3(4):651–661
4. Loog M, Duin R, Haeb-Umbach R (2001) Multiclass linear dimension reduction by weighted pairwise fisher criteria. IEEE Trans Pattern Anal Mach Intell 23(7):762–766
5. Lou DC, Chou CL, Tso HK, Chiu CC (2012) Active steganalysis for histogram-shifting based reversible data hiding. Opt Commun 285(10):2510–2518
6. Lyu S, Farid H (2004) Steganalysis using color wavelet statistics and one-class support vector machines. In: Electronic imaging 2004, international society for optics and photonics, pp 35–45
7. Lyu S, Simoncelli E (2009) Modeling multiscale subbands of photographic images with fields of gaussian scale mixtures. IEEE Trans Pattern Anal Mach Intell 31(4):693–706
8. Martinez A, Kak A (2001) Pca versus lda. IEEE Trans Pattern Anal Mach Intell 23(2):228–233
9. Qi D, Xijian P (2010) Steganalysis of compressed speech based on histogram features. In: Proceedings of 6th IEEE international conference on Wireless Communications networking and mobile computing (WiCOM), pp 1–4
10. Savoldi A, Gubian P (2007) Blind multi-class steganalysis system using wavelet statistics. In: Proceedings of 3rd IEEE international conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP), vol 2, pp 93–96
11. Shi Y, Xuan G, Yang C, Gao J, Zhang Z, Chai P, Zou D, Chen C, Chen W (2005) Effective steganalysis based on statistical moments of wavelet characteristic function. In: Proceedings of IEEE international conference on Information Technology: Coding and Computing (ITCC), vol 1, pp 768–773
12. Sugiyama M (2006) Local fisher discriminant analysis for supervised dimensionality reduction. In: Proceedings of the 23rd international conference on machine learning, pp 905–912. ACM
13. Van der Maaten L (2007) An introduction to dimensionality reduction using matlab. Report 1201:07–07
14. Wu Y, Shih F (2006) Genetic algorithm based methodology for breaking the steganalytic systems. IEEE Trans Syst Man Cybern, Part B 36(1):24–31
15. Xuan G, Shi Y, Huang C, Fu D, Zhu X, Chai P, Gao J (2006) Steganalysis using high-dimensional features derived from co-occurrence matrix and Class-wise Non-Principal Components Analysis (CNPCA), pp 49–60. Springer
16. Yan S, Hu Y, Xu D, Zhang H, Zhang B, Cheng Q (2007) Nonlinear discriminant analysis on embedded manifold. IEEE Trans Circuits Syst Video Technol 17(4):468–477
17. Yan S, Xu D, Zhang B, Zhang H, Yang Q, Lin S (2007) Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell 29(1):40–51
18. Zhao H, Yuen P (2008) Incremental linear discriminant analysis for face recognition. IEEE Trans Syst Man Cybern, Part B 38(1):210–221

Guoming Chen received the M.S. and Ph.D. degree from School of Information Science and Technology, Sun Yat-sen University, China, in 2003 and 2009. He has been awarded a scholarship under the State Scholarship Fund to pursue his study in Eastern Washington University, USA as a joint PHD student for 12 months in 2008. His areas of interest include data mining, machine learning, pattern recognition and image processing.

Qiang Chen received his B.S. degree from Mathematics Department, South China Normal University, Guangzhou, in 1984. He is a Professor of Guangdong University of Education, China. His current research interests are data base system, and knowledge engineering and knowledge management.

Dong Zhang received his B.S.E.E. and M. S. degrees from Nanjing University, China, in 1999 and 2003, respectively, and Ph.D. degree from Sun Yat-sen University, China, in 2009. He is currently a lecturer of the school of information science and technology, Sun Yat-sen University. His research interests include information hiding, image processing, and pattern recognition.

Duanning Zhou is a Professor of Management Information Systems, Chair of the Accounting and Information Systems Department, at the College of Business and Public Administration, Eastern Washington University, Washington, USA. He received his Ph.D. in Information Systems from the City University of Hong Kong in 2000. His current research interests include electronic commerce, data mining, and decision support systems. He has published in ACM Transaction on Internet Technology, Communication of Association for Information Systems, Group Decision and Negotiation, IEEE Transactions on Engineering Management, IEEE Transactions on Education, IEEE Transactions on Systems, Man, and Cybernetics, Information & Management, and other journals.