Feature mining and pattern classification for steganalysis of LSB matching steganography in grayscale images

Qingzhong Liu\textsuperscript{a}, Andrew H. Sung\textsuperscript{a,b,*}, Zhongxue Chen\textsuperscript{c}, Jianyun Xu\textsuperscript{d}

\textsuperscript{a}Department of Computer Science, New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA
\textsuperscript{b}Institute for Complex Additive Systems Analysis, New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA
\textsuperscript{c}Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, USA
\textsuperscript{d}Microsoft Corporation, One Microsoft Way, Redmond, WA 98052-6399, USA

Received 4 July 2006; received in revised form 16 May 2007; accepted 6 June 2007

Abstract

In this paper, we present a scheme based on feature mining and pattern classification to detect LSB matching steganography in grayscale images, which is a very challenging problem in steganalysis. Five types of features are proposed. In comparison with other well-known feature sets, the set of proposed features performs the best. We compare different learning classifiers and deal with the issue of feature selection that is rarely mentioned in steganalysis. In our experiments, the combination of a dynamic evolving neural fuzzy inference system (DENFIS) with a feature selection of support vector machine recursive feature elimination (SVMRFE) achieves the best detection performance. Results also show that image complexity is an important reference to evaluation of steganalysis performance.

\textcopyright 2007 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Steganalysis; LSB matching; DENFIS; SVMRFE; Image complexity

1. Introduction

Steganography is the art and science of communicating hidden messages in such a way that no one apart from the intended recipient knows of the existence of the covert message. Conversely, steganalysis aims to expose the presence of hidden data. To this date, some steganographic embedding methods, such as LSB embedding, spread spectrum steganography, F5 algorithm, have been successfully attacked\cite{1–7}. However, several other embedding paradigms, such as stochastic modulation\cite{8,9} and LSB matching steganography (or plus/minus one embedding), first described in Ref.\cite{10}, are much more difficult to detect statistically.

The literature does contain a few detectors for LSB matching steganography. One of them is the measure of histogram characteristic function center of mass (HCFCOM) proposed by Harmsen and Pearlman\cite{6}. To improve the probability of detection for LSB matching in grayscale images, Ker proposed two novel ways of applying the histogram characteristic function (HCF): calibrating the output using a down-sampled image and computing the adjacency histogram instead of the usual histogram. The best discriminators are Adjacency HCFCOM (A.HCFCOM) and Calibrated Adjacency HCFCOM (C.A.HCFCOM)\cite{11}. Farid and Lyu described an approach for detecting hidden messages in images by using a wavelet-like decomposition to build high-order statistical models of natural images\cite{12,13}. Fridrich et al.\cite{14} presented a maximum likelihood estimator for estimating the number of embedding changes for non-adaptive $\pm K$ embedding in images; based on the stego-signal estimation, a blind steganalysis classifying on high-order statistics of the estimation signal is presented in Ref.\cite{15}. Unfortunately, “the ML estimator starts to fail to reliably estimate the message length $p$ once the variance of $X^F$ exceeds 9”\cite{14}.

The publications mentioned above, however, did not fully address the issue of image complexity that is very important in evaluating the detection performance. Recently, Liu et al.
2.1. Statistical model and image complexity

Several papers [21–24] describe the statistical models of images such as Markov random field models (MRFs), Gaussian mixture model (GMM) and GGD model in transform domains, such as DCT, DFT, or DWT.

Experiments show that a good probability distribution function (PDF) approximation for the marginal density of coefficients at a particular sub-band produced by various types of wavelet transforms may be achieved by adaptively varying two parameters of the GGD [24–26], which is given as

$$p(x; \alpha, \beta) = \frac{\beta}{2\pi \Gamma(1/\beta)} e^{-|x|^{\beta}/\alpha},$$  (1)

where $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt$, $z > 0$ is the Gamma function and $\alpha$ models the width of the PDF peak (standard deviation), while $\beta$ is inversely proportional to the decreasing rate of the peak. Sometimes $\alpha$ is referred to as the scale parameter and $\beta$ is called the shape parameter. The GGD model contains the Gaussian and Laplacian PDFs as special cases, using $\beta=2$ and 1, respectively.

Generally, an image with complicated texture has a high shape parameter of the GGD in the wavelet domain. Three 256x256 grayscale images with different complexity are shown on the left of Fig. 1, the histograms of the Haar wavelet HH wavelet transforms may be achieved by adaptively taking two coefficients at a particular sub-band produced by various types of wavelet transforms.

On the other side, many detection methods in steganalysis are based on feature mining and pattern classification techniques. In feature mining, besides feature extraction, another general problem is feature selection. Avcibas et al. [17] utilized analysis of variance (ANOVA) to choose good image quality metrics. In detail, the higher the $F$ statistic, the lower the $p$-value, and the better the feature is. This feature selection is simple and runs fast. It is good in evaluating the statistical significance of the individual feature, but it does not consider the interaction of the features, and probably, the final feature set is not optimal. Otherwise there has been little research that deals with the feature selection problem with specific respect to steganalysis.

To improve the performance in detecting LSB matching steganography in grayscale images, based on our previous work [16], in this paper, we propose five types of features and introduce a dynamic evolving neural fuzzy inference system (DENFIS) [18,19]. We also adopt the feature selection of support vector machine recursive feature elimination (SVMRFE) [20] to choose the features in our steganalysis.

Comparing against other well-known methods in terms of steganalysis performance, our feature set performs the best. DENFIS is superior to other compared learning classifiers including SVM and adaboost. SVMRFE outperforms DENFIS-based sequential forward selection (SFS) and statistical significance-based feature selection-like T-test.

Our experimental results also indicate that image complexity is an important parameter to evaluation of the detection performance. At a certain information-hiding ratio, it is much more difficult to detect the information-hiding behavior in high-image complexity than that in low complexity.

2. Feature extraction

2.1. Statistical model and image complexity

The high-peak distribution (y-axis) of the wavelet coefficients corresponds to the value zero (x-axis) as shown in Fig. 1. It implies that adjacent pixels are highly correlated and the probability of what the values of adjacent pixels equal to each other is pretty high. To clearly demonstrate the relation of adjacent pixels, an 8-bit grayscale image is shown in Fig. 2(a). The grayscale value at the point $(i, j)$ and the grayscale value at the point $(i+1, j)$ is $v(i, j)$. The joint probability distribution of the pair $(v(i, j), v(i+1, j))$, shown in Fig. 2(b), indicates that the adjacent pixels are highly correlated.

Generally, an image with complicated texture has a high shape parameter of the GGD in the wavelet domain. Three 256x256 grayscale images with different complexity are shown on the left of Fig. 1, the histograms of the Haar wavelet HH sub-band coefficients and the GGD simulations are shown on the right.

2.2. Entropy and high-order statistics of the histogram of the nearest neighbors (NNH)

The entropy of NNH (NNH_E) is calculated as follows:

$$\text{NNH}_E = -\sum \rho_H \log_2 \rho_H,$$  (2)

where $\rho_H$ denotes the distribution density of the NNH.

The high-peak distribution (y-axis) of the wavelet coefficients corresponds to the value zero (x-axis) as shown in Fig. 1. It implies that adjacent pixels are highly correlated and the probability of what the values of adjacent pixels equal to each other is pretty high. To clearly demonstrate the relation of adjacent pixels, an 8-bit grayscale image is shown in Fig. 2(a). The grayscale value at the point $(i, j)$ and the grayscale value at the point $(i+1, j)$ is $v(i, j)$. The joint probability distribution of the pair $(v(i, j), v(i+1, j))$, shown in Fig. 2(b), indicates that the adjacent pixels are highly correlated.

2.2. Entropy and high-order statistics of the histogram of the nearest neighbors (NNH)

The high-peak distribution (y-axis) of the wavelet coefficients corresponds to the value zero (x-axis) as shown in Fig. 1. It implies that adjacent pixels are highly correlated and the probability of what the values of adjacent pixels equal to each other is pretty high. To clearly demonstrate the relation of adjacent pixels, an 8-bit grayscale image is shown in Fig. 2(a). The grayscale value at the point $(i, j)$ and the grayscale value at the point $(i+1, j)$ is $v(i, j)$. The joint probability distribution of the pair $(v(i, j), v(i+1, j))$, shown in Fig. 2(b), indicates that the adjacent pixels are highly correlated.

$$\text{NNH}_HOS(r) = \frac{1/N^3 \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \sum_{z=0}^{N-1} (H(x, y, z) - (1/N^3 \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \sum_{z=0}^{N-1} H(x, y, z))^r}{\sigma_H^r},$$  (3)
where $N$ is the number of possible greyscales of the image, e.g., for an 8-bit grayscale image, $N = 256$.

### 2.3. Probabilities of the equal neighbors (PEN)

Besides the features on the NNH, the PFN are extracted. The structures of the equal neighbors are shown in Fig. 3, where $a$ represents the pixel value. Equal neighbors mean that the pixel values in the structure equal to each other.

### 2.4. Correlation features

The following correlation features are defined.

1. The correlation between the Least significant bit plane (LSBP) and the second least significant bit Plane (LSBP2) and the correlation in the LSBP.

   $M_1(1 : m, 1 : n)$ denotes the binary bits of the LSBP and $M_2(1 : m, 1 : n)$ denotes the binary bits of the LSBP2.

   $C_1 = \text{cor}(M_1, M_2) = \frac{\text{Cov}(M_1, M_2)}{\sigma_{M_1} \sigma_{M_2}}$, \hspace{1cm} (4)
Fig. 2. An 8-bit grayscale image (a) and the joint distribution (b) of the adjacent pixel pair \((v(i, j), v(i + 1, j))\). The horizontal axis in (b) shows the value of the pixel \((i, j)\) and the vertical axis marks the value of the pixel \((i + 1, j)\). The joint distribution indicates the probability of the pair \((v(i, j), v(i + 1, j))\).

\[
\text{Fig. 3. The structures of the equal neighbors.}
\]

where \(\text{Cov}(M_1, M_2) = E[(M_1 - E(M_1))(M_2 - E(M_2))]\), \(E(\bullet)\) is the mathematical expectation, and

\[
\sigma^2_{M_1} = \text{Var}(M_1) \quad \text{and} \quad \sigma^2_{M_2} = \text{Var}(M_2).
\]

The autocorrelation of LSBP \(C(k, l)\) is defined as

\[
C(k, l) = \text{cor}(X_k, X_l), \quad (5)
\]

where \(X_k = M_1(1 : m - k, 1 : n - l); X_l = M_1(k + 1 : m, l + 1 : n)\), the variables \(k\) and \(l\) are the lag distances in the vertical direction and the horizontal direction.

2. The autocorrelation in the image histogram.

The variable \(\rho_k\) denotes the histogram probability density of the original image at the intensity \(k\) \((k = 0, 1, \ldots, N - 1)\) for 8-bit grayscale image, \(N = 256\). The variable \(\rho'_k\) represents the histogram probability density of the adulterated image at the intensity \(k\). The LSBP hiding rate \(r\) is the relative length of hidden binary data, assume the hidden data is i.i.d., for 8-bit grayscale image, \(\rho'_k\) is given as follows:

\[
\rho'_k = (1 - r/2) * \rho_k + (r/4) * \rho_{k-1} + (r/4) * \rho_{k+1},
\]

\[2 \leq k \leq 253,\]

\[
\rho'_0 = (1 - r/2) * \rho_0 + (r/4) * \rho_1,
\]

\[
\rho'_1 = (1 - r/2) * \rho_1 + (r/4) * \rho_2 + (r/2) * \rho_0,
\]

\[
\rho'_{255} = (1 - r/2) * \rho_{255} + (r/4) * \rho_{254},
\]

\[
\rho'_{254} = (1 - r/2) * \rho_{254} + (r/4) * \rho_{253} + (r/2) * \rho_{252}. \quad (6)
\]

Obviously, LSB matching steganography does modify the distribution density of the histogram. Based on this point, we present the correlation features on the histogram. The histogram probability density, \(H\), is denoted as \((\rho_0, \rho_1, \rho_2 \ldots \rho_{N-1})\). The histogram probability densities, \(H_e, H_0, H_{11}, \text{and} H_{12}\) are represented as

\[
H_e = (\rho_0, \rho_2, \rho_4 \ldots \rho_{N-2}), \quad H_0 = (\rho_1, \rho_3, \rho_5 \ldots \rho_{N-1});
\]

\[
H_{11} = (\rho_0, \rho_1, \rho_2 \ldots \rho_{N-1-1}),
\]

\[
H_{12} = (\rho_1, \rho_{1+1}, \rho_{1+2} \ldots \rho_{N-1}).
\]

The autocorrelation coefficients \(C_2\) and \(C_H(l)\) \((l\) is the lag distance\), are defined as follows:

\[
C_2 = \text{cor}(H_e, H_0), \quad (7)
\]

\[
C_H(l) = \text{cor}(H_{11}, H_{12}). \quad (8)
\]

3. The correlation in the difference between the image and the denoised version.

The original cover image is denoted as \(CI\), the stego-image is denoted as \(SI\), \(D(\cdot)\) denotes some denoising function, the difference between the image and the denoised are given as follows:

\[
E_{CI} = CI - D(CI), \quad (9)
\]

\[
E_{SI} = SI - D(SI). \quad (10)
\]
Generally, the statistics of $E_{CI}$ and $E_{SI}$ are different. The correlation features in the difference domain are extracted as follows. First, decompose the testing image by using wavelet transform, find the coefficients in HL, LH and HH sub-bands and the absolute values are smaller than some threshold $t$, set zero to these coefficients, and reconstruct the image by using the inverse wavelet transform. The reconstructed is treated as the denoised image. The difference between original testing image and the reconstructed is $E_t$ ($t$ is the threshold value). The autocorrelation of $E_t$ is given as

$$C_E(t; k, l) = \text{cor}(E_{t,k}, E_{t,l}),$$

where $E_{t,k} = E_t(1 : m - k, 1 : n - l)$; $E_{t,l} = E_t(k + 1 : m, l + 1 : n)$. The variables $k$ and $l$ denote the lag distances.

3. Pattern classification and feature selection

3.1. Introduction to DENFIS

Neuron-fuzzy inference systems consist of a set of rules and an inference method that are embodied or combined with a connectionist structure for better adaptation. Evolving neuron-fuzzy inference systems are such systems, where both the knowledge and the mechanism evolve and change in time, with more examples presented to the system [19]. The DENFIS [18], uses the Takagi–Sugeno type of fuzzy inference method [27]. The inference used in DENFIS is performed on $m$ fuzzy rules indicated as follows:

If $x_1$ is $R_{11}$ and $x_2$ is $R_{12}$ and ... and $x_q$ is $R_{1q}$,

then $y$ is $f_1(x_1, x_2, \ldots, x_q)$

If $x_1$ is $R_{21}$ and $x_2$ is $R_{22}$ and ... and $x_q$ is $R_{2q}$,

then $y$ is $f_2(x_1, x_2, \ldots, x_q)$

... ... ...

If $x_1$ is $R_{mq1}$ and $x_2$ is $R_{mq2}$ and ... and $x_q$ is $R_{mqq}$,

then $y$ is $f_m(x_1, x_2, \ldots, x_q)$.

where “$x_j$ is $R_{ij}$”, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, q$, are $m \times q$ fuzzy propositions that form $m$ antecedents for $m$ fuzzy rules respectively; $x_j$, $j = 1, 2, \ldots, q$, are antecedent variables defined over universes of discourse $X_j$, $j = 1, 2, \ldots, q$, and $R_{ij}$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, q$, are fuzzy sets defined by their fuzzy membership functions: $X_j \rightarrow [0, 1], i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, q$. In the consequent parts of the fuzzy rules, $y$ is the consequent variable, and crisp functions $f_i$, $i = 1, 2, \ldots, m$, are employed.

In the DENFIS model, all fuzzy membership functions are triangular type functions defined by the three parameters, $a, b$, and $c$, as given below:

$$\mu(x) = mf(x, a, b, c) = \max(\min((x - a)/(b - a), (c - x)/(c - b)), 0),$$

where $b$ is the value of the cluster center on the $x$ dimension, $a = b - d \times \text{Dthr}$, $d = 1.2 - 2$. The threshold value, $\text{Dthr}$, is a clustering parameter.

For an input vector $x^0 = [x_1^0, x_2^0, \ldots, x_q^0]$, the result of the inference, $y^0$, or the output of the system, is the weighted average of each rule’s output indicated as follows:

$$y^0 = \sum_{i=1}^{m} w_i f_i(x_1^0, x_2^0, \ldots, x_q^0)\Bigg/\sum_{i=1}^{m} w_i,$$

where, $w_i = \prod_{j=1}^{q} R_{ij}(x_j^0)$; $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, q$.

In the DENFIS on-line model, the first-order Takagi–Sugeno type fuzzy rules are employed. In the DENFIS off-line models, the first-order and an extended high-order Takagi–Sugeno inference engines are used, corresponding to a linear model and an MLP-based model, respectively. The experiments indicate that the DENFIS with MLP-based model has the best prediction performance. The details of the DENFIS off-line learning process is presented in Ref. [19].

3.2. Feature selection in steganalysis

To detect the information-hiding behaviors in steganography, many articles proposed different features or measures. In steganalysis, feature selection should be a general problem; to our knowledge, however, few publications deal with this issue except Avciabas et al. [17] presented a universal steganalysis based on image quality metrics and utilized ANOVA to choose the good measures. Essentially, this feature selection belongs to filtering approach and the final feature set may not be optimal.

Generally, feature selection can be grouped into two categories: filtering and wrapper methods. Filter methods select feature subsets independently from the learning classifiers and do not incorporate learning [28,29]. A weakness of filtering methods is that they just consider the individual feature in isolation and ignore the possible interaction of features among them. Yet, the combination of these features may have a combined effect that does not necessarily follow from the individual performance of features in the group. If there is a limit on the number of features to be chosen, we may not be able to include all informative features.

Wrapper methods wrap around a particular learning algorithm that can assess the selected feature subsets in terms of the estimated classification errors and then build the final classifiers [30]. One of the well-known methods is SVMRFE, which refines the optimum feature set by using SVM in a wrapper approach to address the problem of gene selection in the analysis of microarray data [20]. Additionally, SFS is a greedy search algorithm in wrapper methods. To deal with the issue of feature selection in our steganalysis, we compare these three feature selections: DENFIS-SFS, SVMRFE, and T-test, a filtering feature selection which is similar to the approach in Ref. [17].
4. Experiments

4.1. Experimental setup

The original images in our experiments are 5000 TIFF raw format digital pictures, taken in USA during 2003–2005. These images are 24-bit, 640 × 480 pixels, lossless true color and never compressed. According to the method in Refs. [12,13], we cropped the original images into 256 × 256 pixels in order to get rid of the low complexity parts of the images. The cropped color images are converted into grayscales and LSB matching stego-images are produced by hiding data in these grayscales. The hiding ratio (the ratio of the file size of the hidden data to the file size of the cover image) is 12.5%. The hidden data in any two images are different.

We categorize the grayscale images (covers and stego-images) according to the image complexity which is measured by the shape parameter $\beta$ of the GGD of the HH wavelet sub-band coefficients. Fig. 4 lists some cover samples with different shape parameters in our experiments.

4.2. Feature extraction and comparison

The following features are extracted:

1. Shape parameter $\beta$ of the GGD of the HH wavelet sub-band that measures the image complexity.
2. Entropy of the histogram of the nearest neighbors, $\text{NNH}_E$, defined in Eq. (2).
3. The high-order statistics of the histogram of the nearest neighbors, $\text{NNH}_HOS(r)$ in Eq. (3), and $r$ is set from 3 to 22, total 20 high-order statistics.
4. PEN, described in Section 2.3.
5. Correlations features consist of $C1$ in Eq. (4), $C(k, l)$ in Eq. (5), $C2$ in Eq. (7), $CH(l)$ in Eq. (8), and $CE(t; k, l)$ in Eq. (11).

We set the following lag distance to $k$ and $\text{lin} C(k, l)$ and get 14 features:

1) $k = 0, l = 1, 2, 3, \text{and} 4$; $l = 0, k = 1, 2, 3, \text{and} 4$.
2) $k = 1, l = 1; k = 2, l = 2; k = 3, l = 3; k = 4$ and $l = 4$.
3) $k = 1, l = 2; k = 2, l = 1$.

In Eq. (8), $l$ is set to 1, 2, 3, and 4. In Eq. (11), we set the following lag distances to $k$ and $l$ in $CE(t; k, l)$ and get following pairs:

$CE(t; 0, 1), CE(t; 0, 2), CE(t; 1, 0), CE(t; 2, 0), CE(t; 1, 1), CE(t; 1, 2),$ and $CE(t; 2, 1)$. $t$ is set 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, and 5.

We record the fifth type of correlation features as CF: types 1–5 as EHPCC (Entropy, high-order statistics, probabilities of the equal neighbors, correlation features, and complexity).

To compare EHPCC with other well-known features, the HCFCOM [6] is extracted because the hiding process of LSB matching steganography can be modeled in the context of additive noise. We extend the HCFCOM to the high-order moments. HCFCOM stands for HCF center of mass high-order moments; HCFCOM($r$) denotes the $r$th order statistics. In our experiments, the HCFCOM feature set consists of HCFCOM and HCFCOM($r$) ($r = 2, 3, \text{and} 4$). We also compare A. HCFCOM and C.A. HCFCOM proposed by Ker [11].

Additionally, Farid and Lyu [12,13] presented an approach to detecting hidden messages in images by building high-order moment statistics in multi-scale decomposition domain (we call these features HOMMS), which consists of 72-dimension features in grayscale images.

All the features mentioned above are listed in Table 1. The probability distributions of these features are Gaussian and non-Gaussian. Parametric tests work well with large samples even if the population is non-Gaussian [31]. Fig. 5 lists the $F$ statistics and $p$-values of $\text{NNH}_E$ and $\text{NNH}_HOS$, shape parameter $\beta$ and correlation features, PEN, HOMMS features, HCFCOM features, A. HCFCOM, and C.A. HCFCOM features, extracted from the 5000 grayscale covers and the 5000 LSB matching stego-images. Fig. 5 indicates that, regarding the statistical significance, on the average, $\text{NNH}_E$, $\text{NNH}_HOS$, correlation features, and PEN with high $F$ statistics and very small $p$-values are better than HCFCOM, A. HCFCOM, and C.A. HCFCOM features; and HOMMS features are not good because the $p$-values of most HOMMS features are high and the $F$ statistics are small, it implies that the discrimination ability of HOMMS features is very weak. Fig. 5 also shows that the $F$ statistic of the shape parameter $\beta$ is small and the $p$-value is close to 0, which means that the information-hiding changes the image complexity of the original cover, but the affection is very weak.

4.3. Detection performance on feature sets

To compare the detection performances on these feature sets with different classifiers, besides DENFIS, we apply the following classifiers to each feature sets. These classifiers are naive Bayes classifier (NBC), support vector machines (SVM), quadratic Bayes normal classifier (QDC), and adaboost that produces a classifier composed from a set of weak rules [32–35].

Thirty experiments are done on each feature set using each classifier. In each experiment, training sets are randomly chosen and the remaining sets are tested. The testing results consist of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In each category of the image complexity, the number of cover samples is approximately equal to the number of stego-samples, so the testing accuracy is calculated by $(\text{TP + TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$. The average testing accuracy and the standard error of the 30 experiments are compared.

Table 2 lists the testing results (mean values and standard deviations) on each feature set with the use of SVM, ADABOOST, NBC, and QDC. In each category of image complexity, the number of cover samples is approximately equal to the number of stego-samples, so the testing accuracy is calculated by $(\text{TP + TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$. The average testing accuracy and the standard error of the 30 experiments are compared.
complexity, the best testing accuracy is in bold. In the five categories of image complexity, all the highest testing results happen to the feature set of EHPCC. The results indicate that EHPCC is superior to its subset CF; CF is better than HCFHOM, A.HCFCOM, and C.A.HCFCOM; the detection performance of HOMMS is not good. The results in Table 2 are in agreement with the demonstration of the statistical significance in Fig. 5. Regarding the detection performance of these four learning classifiers, SVM and adaboost are better than NBC and QDC.

Since EPHCC is the best feature set, we compare the detection performance by applying DENFIS to EPHCC against
Table 1
Proposed and compared features in our experiments

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Description of the features</th>
<th>The source</th>
<th>The number of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHPCC</td>
<td>Entropy of NNH (NNH_E)</td>
<td>Defined in Eq. (2)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>High-order statistics of NNH (NNH-HOS (r), (r = 3, 4, \ldots, 22))</td>
<td>Defined in Eq. (3)</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Probabilities of equal neighbors (PEN)</td>
<td>Described in Section 2.3</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Correlation features (CF)</td>
<td>Fig. 3 presents the structures of the equal neighbors</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>Complexity measure</td>
<td>The shape parameter (\beta) in Eq. (1), Refs. [16,24]</td>
<td>1</td>
</tr>
<tr>
<td>HCFHOM</td>
<td>Histogram characteristic function center of mass and the high-order statistics HCFHOM ((r)) ((r = 2, 3, 4))</td>
<td>Ref. [6]</td>
<td>4</td>
</tr>
<tr>
<td>HOMMS</td>
<td>High-order moment statistics in multi-scale decomposition domain (grayscale)</td>
<td>Refs. [12,13]</td>
<td>72</td>
</tr>
</tbody>
</table>

Fig. 5. \(F\) statistics and \(p\)-values of NNH-E (feature dimension 1 on the upper left), NNH-HOS (feature dimension 2–21 on the upper left), image complexity measure \(\beta\) (feature dimension 1 on the middle left), correlation features (feature dimension 2–84 on the middle left), probabilities of equal neighbors, HOMMS, HCFHOM, A. HCFCOM, and C.A. HCFCOM features.
Table 3  

The testing results of applying DENFIS to EHPCC vs. the best results in Table 2

<table>
<thead>
<tr>
<th>β</th>
<th>Feature set</th>
<th>SVM</th>
<th>ADABOOST</th>
<th>NBC</th>
<th>QDC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>91.8 ± 0.9</td>
<td>89.0 ± 1.0</td>
<td>76.0 ± 1.9</td>
<td>70.3 ± 1.8</td>
</tr>
<tr>
<td>&lt; 0.4</td>
<td>EHPCC</td>
<td>85.9 ± 1.0</td>
<td>82.0 ± 1.2</td>
<td>77.0 ± 2.1</td>
<td>80.9 ± 1.7</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>60.9 ± 1.3</td>
<td>57.6 ± 1.5</td>
<td>57.5 ± 1.5</td>
<td>53.4 ± 1.0</td>
</tr>
<tr>
<td></td>
<td>HCFHOM</td>
<td>53.6 ± 1.0</td>
<td>50.6 ± 2.0</td>
<td>46.9 ± 1.7</td>
<td>42.1 ± 1.4</td>
</tr>
<tr>
<td></td>
<td>HOMMS</td>
<td>55.3 ± 0.6</td>
<td>54.3 ± 1.1</td>
<td>53.8 ± 1.1</td>
<td>55.4 ± 1.1</td>
</tr>
<tr>
<td></td>
<td>A.HCFCOM</td>
<td>55.6 ± 0.9</td>
<td>55.4 ± 1.8</td>
<td>54.7 ± 1.4</td>
<td>55.5 ± 1.1</td>
</tr>
<tr>
<td>0.4–0.6</td>
<td>EHPCC</td>
<td>86.2 ± 0.6</td>
<td>79.9 ± 1.0</td>
<td>66.8 ± 0.9</td>
<td>65.1 ± 0.8</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>77.6 ± 0.4</td>
<td>72.2 ± 1.0</td>
<td>67.6 ± 1.3</td>
<td>70.6 ± 1.3</td>
</tr>
<tr>
<td></td>
<td>HCFHOM</td>
<td>58.4 ± 0.6</td>
<td>56.6 ± 1.1</td>
<td>56.1 ± 0.9</td>
<td>54.5 ± 0.6</td>
</tr>
<tr>
<td></td>
<td>HOMMS</td>
<td>48.8 ± 1.6</td>
<td>47.6 ± 1.0</td>
<td>47.1 ± 0.8</td>
<td>44.0 ± 1.5</td>
</tr>
<tr>
<td></td>
<td>C.A.HCFCOM</td>
<td>58.1 ± 0.7</td>
<td>57.0 ± 1.5</td>
<td>57.8 ± 1.1</td>
<td>57.9 ± 0.8</td>
</tr>
<tr>
<td></td>
<td>A.HCFCOM</td>
<td>57.3 ± 0.6</td>
<td>56.6 ± 0.9</td>
<td>56.8 ± 0.7</td>
<td>56.6 ± 0.6</td>
</tr>
<tr>
<td>0.6–0.8</td>
<td>EHPCC</td>
<td>73.7 ± 1.3</td>
<td>69.4 ± 1.2</td>
<td>61.4 ± 1.4</td>
<td>62.8 ± 0.9</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>66.7 ± 0.7</td>
<td>63.9 ± 1.2</td>
<td>62.1 ± 1.1</td>
<td>62.3 ± 1.2</td>
</tr>
<tr>
<td></td>
<td>HCFHOM</td>
<td>57.6 ± 0.9</td>
<td>55.3 ± 1.1</td>
<td>54.2 ± 1.3</td>
<td>53.1 ± 0.7</td>
</tr>
<tr>
<td></td>
<td>HOMMS</td>
<td>47.3 ± 0.7</td>
<td>43.7 ± 1.3</td>
<td>45.4 ± 1.2</td>
<td>40.6 ± 2.4</td>
</tr>
<tr>
<td></td>
<td>C.A.HCFCOM</td>
<td>56.0 ± 1.1</td>
<td>56.4 ± 1.0</td>
<td>55.8 ± 1.0</td>
<td>56.2 ± 0.8</td>
</tr>
<tr>
<td></td>
<td>A.HCFCOM</td>
<td>56.6 ± 0.6</td>
<td>54.9 ± 1.2</td>
<td>55.2 ± 1.1</td>
<td>55.5 ± 1.2</td>
</tr>
<tr>
<td>0.8–1</td>
<td>EHPCC</td>
<td>63.7 ± 1.0</td>
<td>63.0 ± 1.4</td>
<td>56.5 ± 1.2</td>
<td>61.4 ± 1.0</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>60.0 ± 1.0</td>
<td>57.4 ± 1.8</td>
<td>57.8 ± 1.5</td>
<td>57.5 ± 1.6</td>
</tr>
<tr>
<td></td>
<td>HCFHOM</td>
<td>53.9 ± 1.2</td>
<td>52.0 ± 1.6</td>
<td>53.2 ± 1.4</td>
<td>51.7 ± 0.6</td>
</tr>
<tr>
<td></td>
<td>HOMMS</td>
<td>–</td>
<td>42.0 ± 1.5</td>
<td>44.5 ± 0.8</td>
<td>41.6 ± 2.8</td>
</tr>
<tr>
<td></td>
<td>C.A.HCFCOM</td>
<td>52.4 ± 0.7</td>
<td>52.6 ± 1.5</td>
<td>52.1 ± 1.3</td>
<td>53.1 ± 1.2</td>
</tr>
<tr>
<td></td>
<td>A.HCFCOM</td>
<td>53.3 ± 0.6</td>
<td>50.3 ± 1.3</td>
<td>51.8 ± 1.2</td>
<td>50.8 ± 1.6</td>
</tr>
<tr>
<td>&gt;1</td>
<td>EHPCC</td>
<td>54.6 ± 0.2</td>
<td>61.3 ± 1.2</td>
<td>58.0 ± 1.2</td>
<td>60.0 ± 0.5</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>59.7 ± 1.7</td>
<td>58.9 ± 2.3</td>
<td>57.1 ± 1.5</td>
<td>58.4 ± 1.3</td>
</tr>
<tr>
<td></td>
<td>HCFHOM</td>
<td>54.4 ± 0.8</td>
<td>52.7 ± 1.6</td>
<td>51.9 ± 1.7</td>
<td>53.2 ± 1.8</td>
</tr>
<tr>
<td></td>
<td>HOMMS</td>
<td>–</td>
<td>46.7 ± 1.8</td>
<td>50.4 ± 1.4</td>
<td>43.1 ± 1.5</td>
</tr>
<tr>
<td></td>
<td>C.A.HCFCOM</td>
<td>54.7 ± 0.5</td>
<td>52.7 ± 1.7</td>
<td>53.1 ± 1.4</td>
<td>54.4 ± 0.9</td>
</tr>
<tr>
<td></td>
<td>A.HCFCOM</td>
<td>54.3 ± 0.3</td>
<td>51.2 ± 1.6</td>
<td>51.6 ± 2.0</td>
<td>53.5 ± 1.4</td>
</tr>
</tbody>
</table>

In each category of image complexity, the highest test accuracy is in bold with the use of SVM, adaboost, NBC, and QDC. As β > 0.8, SVM fails to classify the HOMMS feature set.

Table 3  

The testing results of applying DENFIS to EHPCC vs. the best results in Table 2

<table>
<thead>
<tr>
<th>β</th>
<th>DENFIS</th>
<th>Best testing in Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.4</td>
<td>93.2 ± 1.1</td>
<td>91.8 ± 0.9</td>
</tr>
<tr>
<td>0.4–0.6</td>
<td>87.7 ± 1.2</td>
<td>86.2 ± 0.6</td>
</tr>
<tr>
<td>0.6–0.8</td>
<td>72.6 ± 1.6</td>
<td>73.7 ± 1.3</td>
</tr>
<tr>
<td>0.8–1</td>
<td>62.5 ± 2.2</td>
<td>63.7 ± 1.0</td>
</tr>
<tr>
<td>&gt;1</td>
<td>62.8 ± 1.8</td>
<td>61.3 ± 1.2</td>
</tr>
</tbody>
</table>

the best testing values in Table 2; the results are shown in Table 3. On the average, DENFIS is better than SVM and adaboost.

4.4. Comparison of feature selections

Although EHPCC has the best detection performance, Fig. 5 shows that not all the features in EHPCC are good, not all the elements of HOMMS are useless. If we combine all the features listed in Table 1, how to choose the features?

Since Tables 2 and 3 show that DENFIS is better than SVM and adaboost and SFS is classical approach in wrapper feature selections, we compare DENFIS-based SFS (DENFIS-SFS) with SVMRFE and T-test. Fig. 6 plots the cross-validation detection performances under the feature dimension 1–40 with the application of DENFIS and SVM to the feature selections: SVMRFE, DENFIS-SFS, and T-test. It shows that, while β > 0.8, by applying SVM to all the feature sets from feature dimension 1–40, it fails to detect the steganography; on the contrary, DENFIS works well. Fig. 6 indicates that, regarding the testing accuracy and the stability spanning over different image complexity, the classifier DENFIS outperforms SVM; the feature selection SVMRFE is superior to DENFIS-SFS and DENFIS-SFS is better than T-test; the combination of DENFIS with SVMRFE achieves the best detection performance.
Fig. 6. The detection performance with the use of SVM and DENFIS to the feature selections: SVMRFE, DENFIS-SFS, and T-test. In the lower subfigures $(0.8 < \beta < 1$ and $1.0 < \beta$), SVM fails to classify the testing sets of covers and stego-images.

5. Conclusions

In this paper, we present a scheme of detecting LSB matching steganography in grayscale images based on feature mining and pattern recognition techniques. Five types of features are extracted and several learning classifiers are applied. Experimental results indicate that the proposed feature set is better than other well-known feature sets including HCFHOM, HOMMS, A.HCFCOM, and C.A.HCFCOM. DENFIS is superior to adaboost, SVM, NBC, and QDC. To deal with the issue of feature selection in steganalysis, we compared three feature selections: SVMRFSE, DENFIS-SFS, and T-test. SVMRFSE performs the best. The learning classifier DENFIS combining with the feature selection of SVMRFSE achieves the best detection performance.

The experimental results also show that image complexity is an important reference to evaluation of the steganalysis performance. At a certain information-hiding ratio, the detection performance is highly different in different image complexity. It is still very challenging in detecting the information-hiding behavior in the grayscale images with high complexity.

Acknowledgments

Partial support for this research received from ICASA (Institute for Complex Additive Systems Analysis, a division of New Mexico Tech), a DoD IASP Capacity Building grant is gratefully acknowledged. We thank Dr. Lyu for offering us the HOMMS code and the reviewers for their insightful comments and helpful suggestions for us to improve our previous work.

References


About the Author—QINGZHONG LIU received his B.E. in 1993 from Northwestern Polytechnical University and M.S. in 1997 from Sichuan University, both in China. He is currently a Ph.D. candidate in the Computer Science Department of New Mexico Tech of USA. His research interests include computational intelligence, data mining, and the applications in bioinformatics, image processing and analysis, information security, and other engineering issues. He is a reviewer for the journal of Pattern Recognition.

About the Author—ANDREW H. SUNG is a professor and the chairman of the Computer Science Department of New Mexico Tech. He received his Ph.D. in 1984 from the State University of New York at Stony Brook. His current research interests include information security, application algorithms, and soft computing and its engineering applications.

About the Author—ZHONGXUE CHEN received his Ph.D. in Statistical Science in 2007 from Southern Methodist University of USA. His research interests include machine learning and bioinformatics.

About the Author—JIANYUN XU received his B.S. in Computer Science and Engineering in 1999 from Zhejiang University of China and M.S. and Ph.D. in Computer Science in 2005 from New Mexico Tech of USA. He is currently working on network security in Microsoft. His research interests include fraud detection, malware detection, network security, watermarking & steganography, steganalysis, and computer graphics.