Optimized Feature Extraction for Learning-Based Image Steganalysis

Ying Wang, *Student Member, IEEE,* and Pierre Moulin, *Fellow, IEEE*

Abstract

The purpose of image steganalysis is to detect the presence of hidden messages in cover photographic images. Supervised learning is an effective and universal approach to cope with the twin difficulties of unknown image statistics and unknown steganographic codes. A crucial part of the learning process is the selection of low-dimensional informative features. We investigate this problem from three angles and propose a three-level optimization of the classifier. First, we select a subband image representation that provides better discrimination ability than a conventional wavelet transform. Second, we analyze two types of features—empirical moments of probability density functions (PDFs) and empirical moments of characteristic functions of the PDFs—and compare their merits. Third, we address the problem of feature dimensionality reduction, which strongly impacts classification accuracy. Experiments show that our method outperforms previous steganalysis methods. For instance, when the probability of false alarm is fixed at 1%, the stegoimage detection probability of our algorithm exceeds that of its closest competitor by at least 15% and up to 50%.

Index Terms

Steganalysis, steganography, supervised learning, feature selection, detection theory, characteristic functions.

I. Introduction

Steganography, the art of covert communication, was already in use thousands of years ago in ancient Greece and China [1]. Today, steganography is an active research area due to the abundance of digital

Manuscript received August 11, 2006; revised October 26, 2006. This work was supported by NSF grant CCR 03-25924 and presented in part at the SPIE Conference on Security, Steganography, and Watermarking of Multimedia Contents, San Jose, CA, January 2006.

The authors are with the University of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA (email: ywang11@ifp.uiuc.edu; moulin@ifp.uiuc.edu).

media, which serve as cover signals, and to the wide availability of public communication networks such as the Internet. By secretly embedding information into an innocuous cover signal, the transmitter hopes that the message will reach the receiver without arousing suspicion. Cover signals with hidden information are called stegosignals. Steganalysis, the counter problem to steganography, aims at detecting the presence of hidden information from seemingly innocuous stegosignals.

This paper focuses on image steganography and steganalysis. Various techniques have been developed to hide data in digital photographic images. Among them, least significant bit (LSB) embedding, which replaces the LSB plane of image pixels with information bits, is easily detectable. This is because LSB embedding limits the pixel value transitions to $0 \leftrightarrow 1$, $2 \leftrightarrow 3$, $\cdots$, $254 \leftrightarrow 255$, and introduces unnatural statistical patterns [2], [3]. However, steganalysis of other embedding techniques, such as spread-spectrum (SS) embedding [4], [5], quantization index modulation (QIM) embedding [6], and stochastic quantization index modulation (SQIM) embedding [7], [8], is more difficult. One reason is that these embedding techniques, unlike LSB embedding, do not have an obvious Achilles' heel. Another reason is inherent to image steganography and steganalysis: unknown image statistics pose a serious challenge to both the steganographer and the steganalyzer. Although recent years have seen considerable progress in image modelling, universal image models still do not exist. However, given a training set consisting of two classes—cover photographic images and stegoimages with hidden information—we can extract features from images and learn their statistics through supervised learning [9], [10]. Hence, the steganographer faces the difficult challenge of approximately preserving statistics of *all* image features after data embedding, and the steganalyzer faces the opposite problem of finding *some* features whose statistics are distinguishably changed by data embedding.

Farid [10] was the first to propose a framework for steganalysis based on supervised learning and to demonstrate that supervised learning is an effective and universal approach to cope with the twin difficulties of unknown image statistics and unknown steganographic codes. The framework was further developed with various ingredients proposed and tested in subsequent papers. Farid [10], Harmsen and Pearlman [11], Xuan et al. [12], Holotyak et al. [13], and Goljan et al. [14] extracted features from image pixels (or wavelet coefficients) and their histograms, while Sullivan et al. [15] worked on the co-occurrence matrix of adjacent image pixels. In order to suppress the large cover interference, Farid [10] used cross-subband prediction errors of wavelet coefficients; Holotyak et al. [13] and Goljan et al. [14] used image denoising techniques to estimate the embedding noise. Given a group of image pixels or wavelet coefficients, two kinds of statistical moments have been used as features. The first is *empirical probability density function (PDF) moments* (often called *sample moments* in the probability and statistics

literature). They refer to the estimates of moments of a PDF from samples and were used by, e.g., Farid [10], Holotyak et al. [13], and Goljan et al. [14]. The second is *empirical characteristic function (CF) moments*, which refer to moments of the discrete CF of the histogram. They were used by, e.g., Harmsen and Pearlman [11] and Xuan et al. [12]. The latter approach appears to be more successful; the authors in [12] made the first attempt to explain this phenomenon, but gaps in the explanations remain (see Section III-C). Finally, different numbers of moments were used during the learning and testing phases: the first four orders of empirical PDF moments were used in [10]; only the first-order empirical CF moments were adopted in [11]; and the first three orders of empirical CF moments were selected in [12].

There are several fundamental questions one may ask: Which moment features are more informative in terms of discriminating between cover images and stegoimages? Is there a mathematical explanation for the superiority of these features? Until what point does steganalysis performance improve with the number of features used? These questions are all related to a crucial ingredient of any machine-learning system: feature extraction. This paper investigates the feature extraction problem for image steganalysis from three angles:

(1) **Image subband decomposition**. Given an image, we select an appropriate image subband representation. For instance, Farid's image representation includes wavelet subband coefficients and their cross-subband prediction errors [10]. However, we have discovered that decomposing the diagonal subband on the finest scale and combining the resulting detail subbands with Farid's representation is beneficial; see Section II.

(2) **Choice of features**. Given a sequence of data samples, we consider both empirical PDF and CF moments as features. These moments are good at capturing different statistical changes; see Sections III-A and III-B. To decide which moments should be used as features, we exploit our prior knowledge about images and commonly used steganographic algorithms. We observe that an effect of data embedding is to smooth out the peaky probability distributions that characterize wavelet coefficients of photographic images. 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, we show in Sections III-E to III-G, both qualitatively and quantitatively, that the empirical CF moments of subband histograms are more sensitive to embedding and hence are better features than empirical PDF moments of subband coefficients. Moreover, this conclusion also holds for those nonadditive embedding algorithms that smooth out the peaky distributions of subband coefficients. On the other hand, for our choice of cross-subband prediction errors (cf. Section II), statistical changes caused by embedding are different from those of

wavelet coefficients and instead the empirical PDF moments outperform the empirical CF moments in our simulations.

(3) **Feature evaluation and selection**. All features are not equally valuable to the learning system. Furthermore, using too many features is undesirable in terms of classification performance due to the *curse of dimensionality* [9]: one cannot reliably learn the statistics of too many features given a limited training set. Hence, we need to evaluate the features' usefulness and select the most relevant ones. In Section IV, we apply feature dimensionality reduction techniques from the pattern recognition and machine learning literature [16] to image steganalysis, thereby improving classification performance.

Finally, Section V applies our proposed image steganalysis method to images and reports experimental results.

1) Notation: We use uppercase letters for random variables, lowercase letters for individual values, and boldface fonts for sequences, e.g., $\mathbf{x} = (x_1, x_2, \cdots, x_N)$. We denote by $p(x)$, $x \in \mathcal{X}$, the probability mass function (PMF) of a random variable X if $\mathcal{X}$ is a set; we use the same notation if $\mathcal{X}$ is a continuum, in which case $p(x)$ is referred to as the PDF of X. We denote by E the mathematical expectation.

The characteristic function of a PDF $p(x)$ is defined as

$$\Phi(t) = \int_{-\infty}^{\infty} p(x) e^{jtx} dx, \tag{1}$$

where $j = \sqrt{-1}$, and the PDF can be recovered as

$$p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi(t) e^{-jtx} dx. \tag{2}$$

II. Multiresolution Image Representation

We decompose images into groups of data samples with similar statistics. A subband transform is often used to decorrelate image data. The resulting coefficients in each detail subband are assumed to be approximately independently and identically distributed (i.i.d.).

In this paper, images are first decomposed into three scales through a Haar wavelet transform[1] to obtain nine detail subbands (horizontal H_i, vertical V_i, and diagonal D_i, $i = 1, 2, 3$) and three approximation (lowpass) subbands (L_i, $i = 1, 2, 3$) as illustrated in Fig. 1. Let us denote by $\mathcal{I}_1$ the set of these 12 wavelet subbands plus the image itself. This image representation was used by Xuan et al. in [12].

[1]The type of wavelets has impact on steganalysis results. The optimal selection of wavelets is however not in the scope of this paper. We simply choose the Haar wavelet for its computational efficiency. The complexity of the fast wavelet transform, measured by the number of arithmetic operations (multiplications and additions) per sample, is directly proportional to N or $\log_2 N$, where N is the filter length [17]. The Haar wavelet is the simplest wavelet with filter length $N = 2$.

We propose to further decompose the first-scale diagonal subband D_1 to improve the performance of the learning system. We then obtain $\mathcal{I}_2$, the set of four extra subbands: lowpass L_2', horizontal H_2', vertical V_2', and diagonal D_2' as shown in Fig. 1. The reason for doing so is as follows. D_1 is the finest detail subband in the Haar wavelet transform, and each of its coefficients involves *diagonal differences* in a four-pixel block. The coefficients in H_2', V_2', and D_2' involve more neighboring pixels. For example, each coefficient in D_2' is essentially a function of adjacent 16 pixels. Hence, H_2', V_2', and D_2' reveal more information about the *difference of differences* between neighboring pixels. In contrast, H_2, V_2, and D_2 are the *averaged differences* because they are calculated from the first-scale lowpass subband L_1, where every coefficient is the average of a four-pixel block.

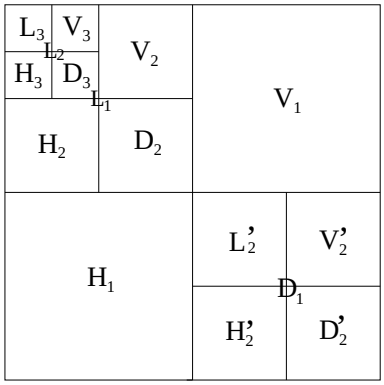

Fig. 1. Three-scale standard wavelet decomposition and an extra level of decomposition on the first-scale diagonal subband D_1.

Since wavelet coefficients possess strong intra- and intersubband dependencies, Farid [10] constructed a set $\mathcal{I}_3$ of nine prediction error subbands to exploit these dependencies as follows. Take a subband coefficient $H_i(j,k)$ as an example, where (j,k) denotes the spatial coordinates at scale i. The magnitude of $H_i(j,k)$ can be linearly predicted by those of its parent $H_{i+1}(j/2,k/2)$; neighbors $H_i(j+1,k)$, $H_i(j,k+1)$, $H_i(j-1,k)$, and $H_i(j,k-1)$; cousins $D_i(j,k)$ and $V_i(j,k)$; and aunts $D_{i+1}(j/2,k/2)$ and $V_{i+1}(j/2,k/2)$. Denote the predicted magnitude as $|\widehat{H_i(j,k)}|$. Then the logarithmic error $eH_i(j,k)$ is given by [10]

$$eH_i(j,k) = \log \frac{|H_i(j,k)|}{|\widehat{H_i(j,k)}|}. \tag{3}$$

This defines an error subband eH_i that corresponds to H_i. One can similarly define the error subbands eV_i and eD_i at scales $i = 1, 2, 3$. The prediction errors for a cover image and its stegoimage have different statistics, which is useful in steganalysis.

Features, such as the various moments to be defined in Sections III-A and III-B, are extracted from each subband in $\mathcal{I}_i$, $i = 1, 2, 3$. Experimental results in Section V-E will show that our best steganalysis performance comes from the more complete multiresolution representation: $\bigcup_{i=1}^{3} \mathcal{I}_i$.

III. Choice of Features: Moments

Given a group of data samples, e.g., coefficients in any subband of the image multiresolution representation $\bigcup_{i=1}^{3} \mathcal{I}_i$, the first important step of supervised-learning based image steganalysis is to choose representative features. Then a decision function is built based on the feature vectors extracted from the two classes of training images: photographic cover images and stegoimages with hidden information. The performance of the decision rule depends on the discrimination capabilities of the features. Also, if the feature vector has low dimension, the computational complexity of learning and implementing the decision function will decrease. In summary, we need to find *informative, low-dimensional* features.

In this section, we first introduce two kinds of such features—empirical PDF moments and empirical CF moments—and explain the interconnections between them. Then we will mainly focus on feature extraction from wavelet subbands in $\mathcal{I}_1$ and $\mathcal{I}_2$. We build statistical models of image steganography, under which we argue that the empirical CF moments are better features for wavelet subbands. For the error subbands in $\mathcal{I}_3$, we do not have a tractable model to analytically argue which kind of moments is better, but a heuristic answer is that the empirical PDF moments are better instead.

A. PDF Moments

For a sequence $\mathbf{x} = (x_1, \cdots, x_N)$ of i.i.d. samples drawn from an unknown PDF $p(x)$, a natural choice of descriptive statistics is a set of empirical PDF moments. The n^{th} empirical PDF moment is given by

$$\hat{\mathsf{m}}_n = \frac{1}{N} \sum_{i=1}^{N} x_i^n, \quad n \geq 1, \tag{4}$$

which is an unbiased estimate of the n^{th} PDF moment

$$\mathsf{m}_n = \mathsf{E} X^n = \int_{-\infty}^{\infty} p(x) x^n dx. \tag{5}$$

The first four moments define the mean, variance, skewness, and kurtosis of the PDF $p(x)$, respectively. Empirical PDF moments were used by Farid [10] and Holotyak et al. [13].

Often, image and stegoimage wavelet coefficients exhibit symmetry around 0, and hence empirical PDF moments of odd orders are approximately 0. Therefore, Goljan et al. in [14] chose to use the n^{th}

empirical absolute PDF moment

$$\hat{\mathsf{m}}_n^A = \frac{1}{N}\sum_{i=1}^{N}|x_i|^n, \quad n \geq 1, \tag{6}$$

which is an estimate of the n^{th} absolute PDF moment

$$\mathsf{m}_n^A = \mathsf{E}|X|^n = \int_{-\infty}^{\infty} p(x)|x|^n dx. \tag{7}$$

From (5) and (7), $p(x)$ is weighted by x^n and $|x|^n$, respectively, and any change in the *tails* of $p(x)$ is polynomially amplified in PDF moments. As is well known, $\hat{\mathsf{m}}_n$ and m_n in (4) and (5) relate to the n^{th} derivative of the CF $\Phi(t)$ of the PDF $p(x)$ at $t = 0$ by

$$\hat{\mathsf{m}}_n \approx \mathsf{m}_n = j^{-n}\frac{d^n}{dt^n}\Phi(t)\Big|_{t=0}. \tag{8}$$

Moreover,

$$\hat{\mathsf{m}}_n^A \approx \mathsf{m}_n^A \geq |\mathsf{m}_n| = \left|\frac{d^n}{dt^n}\Phi(t)\right|_{t=0}. \tag{9}$$

For a heavy-tailed PDF, m_n is large and it follows from (8) that $\Phi(t)$ has large derivatives at the origin, i.e., is peaky.

B. CF Moments

Analogously, for the CF $\Phi(t)$, its n^{th} moment is defined by

$$\mathsf{M}_n = \int_{-\infty}^{\infty}\Phi(t)t^n dt, \tag{10}$$

and its n^{th} absolute moment is

$$\mathsf{M}_n^A = \int_{-\infty}^{\infty}|\Phi(t)||t|^n dt. \tag{11}$$

In the above integral, $|\Phi(t)|$ is weighted by $|t|^n$. Any change in the *tails* of $|\Phi(t)|$, which correspond to the high-frequency components of $p(x)$, is thus polynomially amplified. Similarly to (8) and (9), the CF moments M_n and M_n^A relate to the n^{th} derivative of $p(x)$ at $x = 0$ by

$$\mathsf{M}_n = j^n 2\pi\frac{d^n}{dx^n}p(x)\Big|_{x=0} \tag{12}$$

and

$$\mathsf{M}_n^A \geq |\mathsf{M}_n| = 2\pi\left|\frac{d^n}{dt^n}p(x)\right|_{x=0}. \tag{13}$$

If a CF $\Phi(t)$ has heavy tails and M_n^A is large, then the corresponding PDF $p(x)$ is peaky. Equations (8), (9) and (12), (13) reveal a duality between PDF moments and CF moments that follows from the duality between the PDF $p(x)$ and its CF $\Phi(t)$.

To obtain the corresponding empirical CF moments from a sample sequence $\mathbf{x}$, we first estimate the PDF $p(x)$ using an M-bin histogram $\{h(m)\}_{m=0}^{M-1}$. Let $K = 2^{\lceil \log_2 M \rceil}$. The K-point discrete CF $\{\Phi(k)\}_{k=0}^{K-1}$ is then defined as

$$\Phi(k) = \sum_{m=0}^{M-1} h(m) \exp\left\{\frac{j2\pi mk}{K}\right\}, \quad 0 \le k \le K-1, \tag{14}$$

which is analogous to $\Phi(t)$ defined in (1) and can be easily computed using the fast Fourier transform (FFT) algorithms. Similarly to (2), the histogram

$$h(m) = \frac{1}{K} \sum_{k=0}^{K-1} \Phi(k) \exp\left\{-\frac{j2\pi mk}{K}\right\}, \; 0 \le m \le M-1 \tag{15}$$

can be recovered from the discrete CF $\Phi(k)$.

Harmsen and Pearlman defined the n^{th} absolute moment of the discrete CF $\{\Phi(k)\}_{k=0}^{K-1}$ as [11]

$$\hat{\mathsf{M}}'_n = \sum_{k=0}^{K/2-1} |\Phi(k)| k^n, \tag{16}$$

which is obtained by replacing the integral over t in (11) with a summation over k. We prefer to define the n^{th} moment of a discrete CF as

$$\hat{\mathsf{M}}_n = \sum_{k=0}^{K-1} \Phi(k) \sin^n\left(\frac{\pi k}{K}\right) \tag{17}$$

and the n^{th} absolute moment of a discrete CF as

$$\hat{\mathsf{M}}_n^A = \sum_{k=0}^{K-1} |\Phi(k)| \sin^n\left(\frac{\pi k}{K}\right). \tag{18}$$

The motivation is that $\hat{\mathsf{M}}_n^A$ in (18) provides an upper bound on the discrete derivatives of the histogram $\{h(m)\}_{m=0}^{M-1}$, just as in (13) M_n^A bounds the derivatives of the PDF $p(x)$ from above. Indeed, for the first discrete derivative of the histogram, we have

$$\begin{aligned} |h^{(1)}(m)| &= |h(m) - h(m-1)| \\ &\le \frac{2}{K} \sum_{k=0}^{K-1} |\Phi(k)| \cdot \sin\left(\frac{\pi k}{K}\right) \\ &= \frac{2}{K} \hat{\mathsf{M}}_1^A, \quad 1 \le m \le M-1, \end{aligned} \tag{19}$$

where the inequality follows directly from (15), and (19) is obtained by applying (18) with $n = 1$. Similarly, for the n^{th} discrete derivative,

$$\begin{aligned} |h^{(n)}(m)| &= \left| \sum_{i=0}^{n} (-1)^i C_n^i \; h\left(m + \lfloor \frac{n}{2} \rfloor - i\right) \right| \\ &\le \frac{2^n}{K} \hat{\mathsf{M}}_n^A, \quad \lceil \frac{n}{2} \rceil \le m \le M - \lceil \frac{n+1}{2} \rceil, \end{aligned} \tag{20}$$

where C_n^i is the binomial coefficient that denotes the number of size-k subsets from a size-n set.

We also define the normalized CF moments as

$$\tilde{\mathsf{M}}_n^A = \frac{\hat{\mathsf{M}}_n^A}{\hat{\mathsf{M}}_0^A}, \quad n \geq 1, \tag{21}$$

where $\hat{\mathsf{M}}_n^A$ is normalized by the zeroth-order moment $\hat{\mathsf{M}}_0^A$. A similar normalization was used by Harmsen and Pearlman [11]:

$$\tilde{\mathsf{M}}_n' = \frac{\hat{\mathsf{M}}_n'}{\hat{\mathsf{M}}_0'}, \quad n \geq 1. \tag{22}$$

The advantage of normalized CF moments over unnormalized ones will be evident in Section IV-A.

C. The Better Choice: PDF Moments or CF Moments?

If we casually examine $\hat{\mathsf{m}}_n^A$ and $\hat{\mathsf{M}}_n^A$ defined in (6) and (18), it is difficult to tell which one will serve better as a feature in image steganalysis. Compared to $\hat{\mathsf{m}}_n^A$, $\hat{\mathsf{M}}_n^A$ has some computational disadvantages: an appropriate bin width[2] is needed to obtain a histogram that is a good estimate of the underlying PDF; then a K-point FFT is used to calculate the discrete CF; and finally $\hat{\mathsf{M}}_n^A$ is obtained as a weighted sum of the magnitude of the discrete CF samples.

1) For wavelet subbands in $\mathcal{I}_1$ and $\mathcal{I}_2$: Image steganalysis experiments conducted by Xuan et al. [12] show that the moments $\hat{\mathsf{M}}_n'$ in (16) outperform both $\hat{\mathsf{m}}_n$ in (4) and $\hat{\mathsf{m}}_n^A$ in (6) on various data-embedding methods. In [12], Xuan et al. provided basically two arguments to explain the above phenomenon. Their first argument comes from a comparison of M_n^A and m_n^A for Gaussian embedding noise $\mathcal{N}(0, \sigma^2)$. They showed that M_n^A is proportional to $1/\sigma^{n+1}$ while m_n^A is proportional to σ^n, and from this they argued that M_n^A is more sensitive to embedding than m_n^A. However, this reasoning is not satisfactory in that during the process of supervised learning, we extract features from the cover signal samples and the stegosignal samples, but not directly from the embedding noise samples. The second argument in [12] is that since m_n^A "averages" the change of PDFs caused by embedding via "integration" and M_n^A catches the change of PDFs via "differentiation," M_n^A must be more sensitive to the change than m_n^A. However, it is not clear why "differentiation" *must* outperform "integration."

[2]For a fixed-resolution histogram, the bin width plays the primary role of a smoothing parameter, which controls the final appearance of the nonparametric PDF estimate. If the bin width is too large, the estimate may miss small details and key features due to over-smoothing; if the bin width is too small, the estimate exhibits volatile and extraneous wiggles. A good choice of the number of bins is in the range of $O(N^{1/3})$ to $O(N^{1/2})$, where N is the number of available samples [18]. The histogram of a typical image wavelet subband usually has 50 to 200 bins.

In Sections III-D to III-G, we will exploit our prior knowledge about *image steganography* in choosing the right features: image wavelet coefficients (those in $\mathcal{I}_1$ and $\mathcal{I}_2$) have peaky, heavy-tailed probability distributions; after data embedding, these peaky distributions are smoothed. We will build approximate statistical models for image steganography, examine the statistical differences between cover signals and stegosignals, and discuss which kind of moments, m_n^A or M_n^A, best captures these differences.

2) For prediction error subbands in $\mathcal{I}_3$: To our knowledge, only Farid [10] has reported steganalysis results using $\hat{\mathsf{m}}_n$ from the prediction error subbands in $\mathcal{I}_3$. Unfortunately, unlike for wavelet subbands, we do not have a tractable statistical model for these prediction error subbands: the errors are centered around zero for cover images, but we do not observe a clear law that governs how the statistics change after even the simple additive embedding. Based on our simulations, however, we conclude that for the prediction error subbands in $\mathcal{I}_3$, the empirical PDF moments $\hat{\mathsf{m}}_n$ outperform the CF moments $\hat{\mathsf{M}}_n^A$. We defer the experimental details to Sections IV-A and V-D.

Next, we focus on the wavelet subbands in $\mathcal{I}_1$ and $\mathcal{I}_2$ only. Also, we mainly deal with *additive* embedding for two reasons. First, many embedding algorithms, such as the widely used SS scheme [4], [5] and the $\pm k$ embedding scheme [19], have embedding noise that is independent of the cover signal. Second, with the constraint of additive embedding, the mathematical analysis is tractable. For simplicity and clarity, our following analysis is developed for continuous PDFs and uses the definitions of m_n^A in (7) and M_n^A in (11).

D. General Statistical Model for Additive Embedding

For additive embedding, the relationship between stegosignal $\mathbf{X}$, cover signal $\mathbf{S}$, and effective embedding noise $\mathbf{Z}$ is given by

$$\mathbf{X} = \mathbf{S} + \mathbf{Z}, \tag{23}$$

where $\mathbf{Z}$ is independent of $\mathbf{S}$ and is a function of transmitted messages and secret keys shared between the encoder and decoder. Under the i.i.d. model of Section III-A, the independence between $\mathbf{S}$ and $\mathbf{Z}$ leads to the following convolution equation between the marginal PDFs:

$$p_X(x) = \int_{s \in \mathcal{S}} p_S(s) p_Z(x - s) ds. \tag{24}$$

Therefore,

$$\Phi_X(t) = \Phi_S(t) \Phi_Z(t). \tag{25}$$

We also have

$$\mathsf{m}_{n,X} = \mathsf{E}(S+Z)^n. \tag{26}$$

By the uniqueness theorem of moment generating functions [20, p. B-11], $\mathsf{m}_{n,X}$ is different from $\mathsf{m}_{n,S} = \mathsf{E}S^n$ for at least one $n \geq 1$, unless $p_X = p_S$ almost surely. It is hard to compare $\mathsf{m}_{n,X}$ and $\mathsf{m}_{n,S}$ generally, but when the noise PDF p_Z is symmetric to the origin and n is even, it is easy to verify that

$$\begin{aligned} \mathsf{m}^A_{n,X} = \mathsf{m}_{n,X} &= \mathsf{E}(S+Z)^n \\ &\geq \mathsf{E}S^n + \mathsf{E}Z^n \\ &\geq \mathsf{E}S^n = \mathsf{m}_{n,S} = \mathsf{m}^A_{n,S}. \end{aligned} \tag{27}$$

From (24), an independent noise PDF p_Z may be thought of as a linear shift-invariant filter applied to p_S. In the frequency domain,[3] (25) shows that the stegosignal CF is a product of the CFs of the cover signal and additive noise. Since

$$|\Phi_Z(t)| \leq 1, \quad \forall\, t \in \mathbb{R}, \tag{28}$$

it is always true that

$$|\Phi_X(t)| \leq |\Phi_S(t)|, \quad \forall\, t \in \mathbb{R}. \tag{29}$$

From (11), it follows that, for *any additive* embedding noise,

$$\mathsf{M}^A_{n,X} \leq \mathsf{M}^A_{n,S}. \tag{30}$$

If p_Z is "smooth," as is the case for the Gaussian embedding noise in SS schemes [4, Section IV] or the uniform embedding noise in DC-QIM schemes [21], $|\Phi_Z(t)|$ decays quickly as $|t|$ becomes larger and its effect is equivalent to a lowpass filter to p_S: the resulting p_X has highly attenuated high-frequency components and is *smoother* than p_S. Interested readers are referred to [22, Chapter 2] and [23] for more details on the decay properties of characteristic functions.

E. An Image Embedding Model: Generalized Gaussian Cover Signal Plus Gaussian Noise

The additive embedding setup of Section III-D is quite general but does not tell us whether m^A_n or M^A_n is changed *more* in *image* steganography. Even though we do not know the exact image statistics

[3]The conjugate of the CF of a PDF $p(x)$, denoted as $\Phi^*(t)$, is proportional to the Fourier transform of the PDF; see (1) for the connection. Thus, we can regard the CF $\Phi_Z(t)$ as a frequency domain response of the noise PDF p_Z and study its filtering effects on $\Phi_S(t)$ and p_S.

and the underlying embedding algorithms, fortunately, we do have some prior knowledge about image statistics and characteristics of commonly used data-embedding techniques. Next, we incorporate those specifics into the additive embedding model of (23).

Image wavelet coefficients in high-pass subbands serve as cover signal $\mathbf{S}$ and are well modelled by generalized Gaussian distributions (GGDs) [24]. This model is widely used in image coding [25], denoising [26], and other applications. A GGD is given by

$$p_{\alpha,\beta}(s) \triangleq \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} \exp\left\{-\left(\frac{|s|}{\alpha}\right)^{\beta}\right\},\ \alpha>0,\ \beta>0,\ s\in\mathbb{R}, \tag{31}$$

where $\Gamma(\cdot)$ is the Gamma function, α is the scale parameter, and β is the shape parameter. The Gaussian and Laplacian PDFs are special cases of GGD with $\beta = 2$ and 1, respectively.

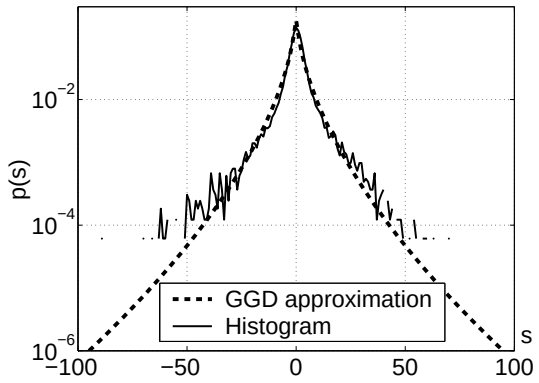

Fig. 2. Histogram of Haar wavelet coefficients from the finest diagonal subband of the *Lena* image and the maximum likelihood GGD estimate of the underlying PDF.

We model the effective embedding noise as a mixture of zeros (with probability $1-\gamma$) and Gaussian noise $\mathcal{N}(0,\sigma^2)$ (with probability γ):

$$Z_\gamma \sim (1-\gamma)\delta(0) + \gamma\mathcal{N}(0,\sigma^2), \quad \gamma\in[0,1]. \tag{32}$$

The justification for this mixture model is as follows. First, many embedding algorithms only embed data in a fraction ($\gamma\in[0,1]$) of either image pixels or transform domain coefficients (see, e.g., [4], [8], and [19]). The embedding locations are randomized and are part of the secret key. When γ is small, the noise also has a peaky PDF. Second, conditioned on the embedding locations, Gaussian embedding noise is a reasonable model for many data embedding methods (e.g., SS methods and $\pm k$ methods). Thus, besides the embedding fraction γ, we also use the reference noise-to-cover ratio (RNCR)

$$\text{RNCR} = \frac{\mathsf{E}Z_1^2}{\mathsf{E}S^2} = \frac{\sigma^2}{\mathsf{E}S^2} \tag{33}$$

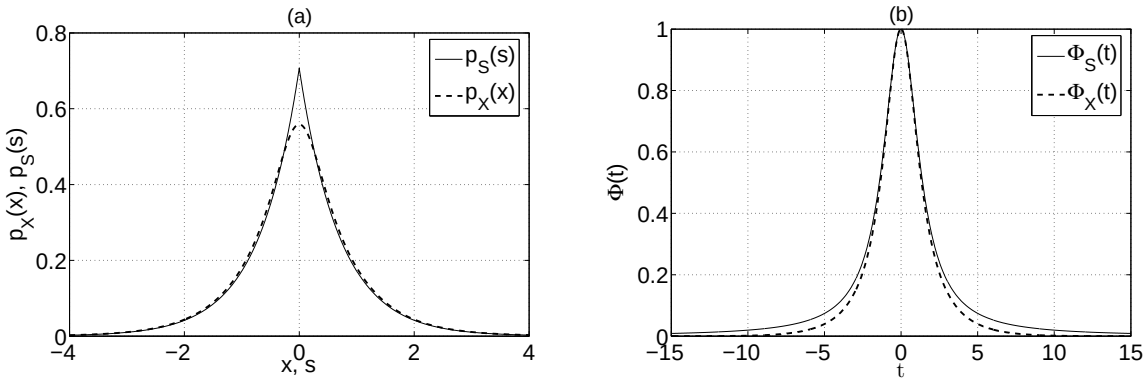

Fig. 3. PDFs and corresponding CFs of a Laplacian distributed cover signal S and its stegosignal $X = S + Z$, where $Z \sim \mathcal{N}(0, \sigma^2)$ with RNCR $= \frac{\sigma^2}{\mathsf{E}S^2} = 0.05$ and $\gamma = 1$. (a) PDFs: p_S and p_X. (b) CFs: Φ_S and Φ_X.

as an indicator of the embedding strength.

In summary, we consider the following image embedding model in the wavelet domain:

$$\begin{cases} X = S + Z_\gamma, \\ S \sim p_{\alpha,\beta}(s), \quad \alpha > 0,\ \beta > 0, \\ Z_\gamma \sim (1-\gamma)\delta(0) + \gamma\mathcal{N}(0, \sigma^2), \quad \gamma \in [0, 1]. \end{cases} \tag{34}$$

F. Remarks

For image wavelet coefficients, histograms and estimated $p_{\alpha,\beta}(s)$ are often peaky at $s = 0$ while having heavy tails at large s; see Fig. 2 for an example. Usually, $\beta \in (0.3, 2)$ [27]. When $p_{\alpha,\beta}(s)$ is linear-filtered by a smooth $p_Z(z)$ such as a Gaussian PDF, the peak is levelled much more than tails. Thus, the most significant difference between $p_S(s)$ and $p_X(x)$ appears in the vicinity of the origin; see Fig. 3(a) for an illustration. According to (7), PDF absolute moments m_n^A are obtained by weighting the PDF $p(x)$ with $|x|^n$, which gives zero or little weight to the vicinity of the origin. Thus m_n^A *discounts* the part of the PDF that is most changed by embedding instead of emphasizing it. Remember from (9) that m_n^A relates to the n^{th} derivative of the corresponding CF at the origin: the two before- and after-embedding CFs shown in Fig. 3(b) correspond to the two PDFs in Fig. 3(a) and have little difference in the vicinity of $t = 0$.

In contrast, CF absolute moments M_n^A are obtained by weighting the CF $\Phi(t)$ with polynomially increasing weight $|t|^n$. As illustrated by Fig. 3(b), distinguishable differences between $\Phi_S(t)$ and $\Phi_X(t)$ appear at large t's and these differences are *emphasized* by M_n^A. This may also be seen by examining (13): M_n^A relates to the n^{th} derivative of the corresponding PDF at the origin; we see from Fig. 3(a) that

$p_S(s)$ and $p_X(x)$ have considerably different derivatives at the origin. Therefore, M_n^A is more sensitive to embedding than m_n^A for the image embedding model of (34).

G. Quantitative Analysis

Next we compare the ratio between $\mathsf{m}_{n,S}^A$ and $\mathsf{m}_{n,X}^A$ and the ratio between $\mathsf{M}_{n,S}^A$ and $\mathsf{M}_{n,X}^A$ for the model of (34). The ratios are defined as

$$r_{\mathsf{m},n} = \max\left(\frac{\mathsf{m}_{n,X}^A}{\mathsf{m}_{n,S}^A}, \frac{\mathsf{m}_{n,S}^A}{\mathsf{m}_{n,X}^A}\right) \tag{35}$$

and

$$r_{\mathsf{M},n} = \max\left(\frac{\mathsf{M}_{n,X}^A}{\mathsf{M}_{n,S}^A}, \frac{\mathsf{M}_{n,S}^A}{\mathsf{M}_{n,X}^A}\right). \tag{36}$$

The more a ratio deviates from one, the more sensitive the corresponding moment is to embedding. Furthermore, if M_n^A is a better feature choice than m_n^A, the ratio

$$A_n = \frac{r_{\mathsf{M},n}}{r_{\mathsf{m},n}} \tag{37}$$

exceeds 1, and we call A_n the *advantage* of M_n^A over m_n^A.

1) $\beta = 2$: For the Gaussian cover distribution ($\beta = 2$), the calculation of the above ratios is given in Appendix I. We have

$$r_{\mathsf{m},n} = 1 - \gamma + \gamma(1 + \text{RNCR})^{\frac{n}{2}} \tag{38}$$

and

$$r_{\mathsf{M},n} = \frac{1}{1 - \gamma + \gamma(1 + \text{RNCR})^{-\frac{n+1}{2}}}. \tag{39}$$

See Fig. 4(a) for the case of RNCR $= 0.05$ and $\gamma = 1$. Both $r_{\mathsf{m},n}$ and $r_{\mathsf{M},n}$ are monotonically increasing functions of the moment order n and embedding strength indicators γ and RNCR. The advantage A_n is a function of γ and RNCR:

$$A_n = \frac{\left[1 - \gamma + \gamma(1 + \text{RNCR})^{-\frac{n+1}{2}}\right]^{-1}}{\left[1 - \gamma + \gamma(1 + \text{RNCR})^{\frac{n}{2}}\right]}. \tag{40}$$

Clearly, $A_n = 1$ when $\gamma = 0$ or RNCR $= 0$, and $A_n = (1 + \text{RNCR})^{\frac{1}{2}} \geq 1$ when $\gamma = 1$. Moreover, it is derived in Appendix I that $A_n \geq 1$ if $\gamma \in [\max(0, \gamma_1), 1]$, where

$$\gamma_1 \triangleq 1 - \frac{(1 + \text{RNCR})^{\frac{n+1}{2}} - (1 + \text{RNCR})^{\frac{n}{2}}}{\left[(1 + \text{RNCR})^{\frac{n}{2}} - 1\right]\left[(1 + \text{RNCR})^{\frac{n+1}{2}} - 1\right]}. \tag{41}$$

Also, $A_n \geq 1$ for all $\gamma \in [0, 1]$ only when $\gamma_1 < 0$, i.e.,

$$(1 + \text{RNCR})^{-\frac{n+1}{2}} + (1 + \text{RNCR})^{\frac{n}{2}} < 2. \tag{42}$$

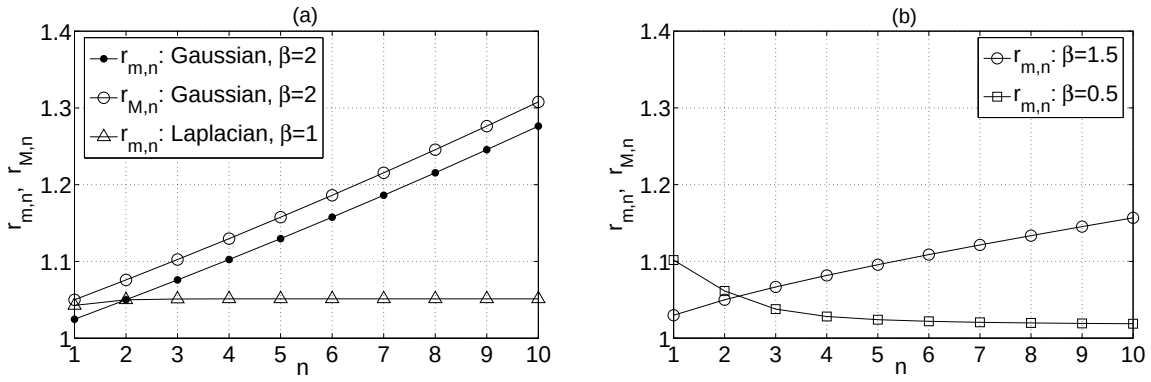

Fig. 4. Ratios $r_{\mathsf{m},n}$ and $r_{\mathsf{M},n}$ for GGD cover signals, assuming an additive Gaussian embedding noise. RNCR $= 0.05$ and $\gamma = 1$. (a) Gaussian cover signal and Laplacian cover signal. (b) GGD cover signal with $\beta = 1.5$ and GGD cover signal with $\beta = 0.5$. For the GGD cover signal with $\beta = 1.5$, $r_{\mathsf{M},n}$ is not shown since $r_{\mathsf{M},n} \approx 2$ when $n = 1$ and $r_{\mathsf{M},n}$ is infinite when $n > 1$. Also, for the Laplacian cover signal and the GGD cover signal with $\beta = 0.5$, $r_{\mathsf{M},n}$ is infinite when $n > 1$ and not shown.

Thus, for the case of a Gaussian cover signal, when the embedding fraction γ is close to 1, M_n^A is a better feature choice than m_n^A; otherwise, m_n^A may have advantage over M_n^A for either large RNCR or large n.

2) $1 < \beta < 2$*:* The cover GGDs with $1 < \beta < 2$ are first-order differentiable but higher-order nondifferentiable at the origin. So $\mathsf{M}_{n,S}^A$ ($n > 1$) is unbounded according to (13). When RNCR > 0 and $\gamma > 0$, numerical calculation shows that $\mathsf{M}_{n,X}^A$ is finite and hence $r_{\mathsf{M},n} = \infty$ for $n > 1$. It also shows that $r_{\mathsf{m},n}$ is always finite and $r_{\mathsf{m},n=1} < r_{\mathsf{M},n=1}$. Fig. 4(b) displays $r_{\mathsf{m},n}$ when $\beta = 1.5$. Thus, for any RNCR > 0 and $\gamma \in (0, 1]$, $A_{n=1} > 1$ and $A_n = \infty$ when $n > 1$. Hence, for the case of $1 < \beta < 2$, M_n^A is always a better feature choice than m_n^A.

3) $0 < \beta \leq 1$*:* When $\beta \leq 1$, $\mathsf{M}_{n=1,S}^A$ is also unbounded. Numerical calculation shows that when RNCR > 0 and $\gamma > 0$, $\mathsf{M}_{n,X}^A$ is finite and so is $r_{\mathsf{m},n}$; see Figures 4(a) and 4(b) for $r_{\mathsf{m},n}$ at $\beta = 1$ and 0.5, respectively. So, we have $r_{\mathsf{M},n} = \infty$ and $A_n = \infty$ for any RNCR > 0 and $\gamma \in (0, 1]$. Again, for the case of $0 < \beta \leq 1$, M_n^A is always a better feature choice than m_n^A.

In summary, the advantage A_n increases to ∞ when β decreases from 2 to 1. That is, when the cover distribution becomes more peaky, M_n^A is *much more sensitive* to embedding and hence is a *better feature* than m_n^A.

H. Discussion

The above analysis of the CF moment M_n^A versus the PDF moment m_n^A is performed on Gaussian/GGD PDFs and CFs with infinite precision. In practice, we handle M-bin histograms $\{h(m)\}$ and K-point discrete CFs $\{\Phi(k)\}$; moreover, actual marginal PDFs of image wavelet and DCT coefficients may not belong exactly to the GGD family. See Fig. 2 for the example of the *Lena* image. The empirical CF moment $\hat{\mathsf{M}}_n^A$ defined in (18) is always finite, and the theoretical advantage of $\hat{\mathsf{M}}_n^A$ over the empirical PDF moment $\hat{\mathsf{m}}_n^A$ may be reduced by factors such as finite precision, suboptimal histogram bin width, and uncertainty about the underlying cover PDF.

The remarks in Section III-F are not limited to the model in (34), but are applicable to *any* model where the marginal cover PDF is peaky and the marginal stego PDF is smooth at the origin (see Fig. 5). As long as this property holds, the CF moment M_n^A is generally a better feature than the PDF moment m_n^A, even when the embedding noise PDF is non-Gaussian and/or nonadditive.

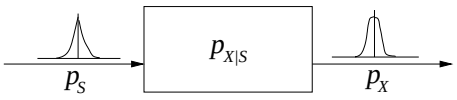

Fig. 5. An embedding black box that smoothes the peaky cover signal PDF.

IV. Feature Selection

Given a multiresolution image representation, e.g., $\bigcup_{i=1}^{3} \mathcal{I}_i$, we can calculate an arbitrary number of moments from each subband. In the current literature on moment-based image steganalysis, the number of moments used in training and testing is somewhat arbitrary: the first four PDF moments $\hat{\mathsf{m}}_n$ were used in [10]; the first CF moment $\tilde{\mathsf{M}}_n'$ was adopted in [11]; and the first three CF moments $\tilde{\mathsf{M}}_n'$ were selected in [12]. However, we learn form Fig. 4 that in some cases $r_{\mathsf{M},n}$ and $r_{\mathsf{m},n}$ increase with the order n—the higher order a moment is, the more sensitive it is to embedding. So why do we not use higher order moments as features instead? And why do we not use as many moments as possible? Next we will address these issues.

A. Feature Evaluation

Each feature is a statistic of data samples, and its impact on classification accuracy is determined by the feature-label distribution. Several criteria from the pattern recognition and machine learning literature

may be used to evaluate the usefulness of a feature in discriminating between classes [28]. In this paper, we choose to use the Bhattacharyya distance

$$B(p_0, p_1) = -\log \int_{\mathcal{X}} \sqrt{p_0(x) p_1(x)}\, dx, \tag{43}$$

where x is a feature (or a feature vector), $\mathcal{X}$ is the feature space, and $p_0(x)$ and $p_1(x)$ are the feature PDFs under Class 0 and Class 1, respectively.

From its definition, the Bhattacharyya distance has the nice property that it is additive over independent features:

$$B\big(p_0(x)q_0(y), p_1(x)q_1(y)\big) = B(p_0, p_1) + B(q_0, q_1), \tag{44}$$

where p_i and q_i ($i = 0, 1$) are the respective PDFs of two independent features X and Y. The Bhattacharyya distance also provides bounds on P_e, the average probability of error in discrimination between two equally likely classes, through [29], [30]

$$\frac{1}{2}\left[1 - \left(1 - e^{-2B(p_0,p_1)}\right)^{\frac{1}{2}}\right] \leq P_e \leq \frac{1}{2} e^{-B(p_0,p_1)}. \tag{45}$$

The larger the $B(p_0, p_1)$ for a feature, the better the suitability of that feature for classification. Always, $B(p_0, p_1) \geq 0$; only when $p_0 = p_1$, $B(p_0, p_1) = 0$ and 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.

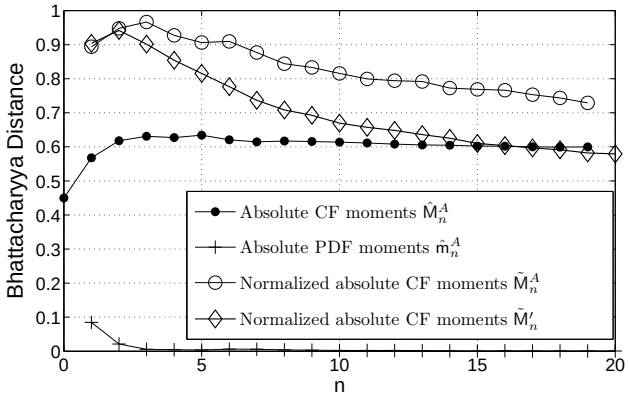

Fig. 6. Empirical Bhattacharyya distance for features $\hat{\mathsf{M}}_n^A$ from (18), $\hat{\mathsf{m}}_n^A$ from (6), $\tilde{\mathsf{M}}_n^A$ from (21), and $\tilde{\mathsf{M}}_n'$ from (22), $1 \leq n \leq 20$. Data are gathered from the first diagonal subband of the Haar-wavelet transform of 1370 photographic images, and their corresponding stegoimages with additive Gaussian noise $\mathcal{N}(0, 4)$ (quantized to integers) in the pixel domain ($\gamma = 1$).

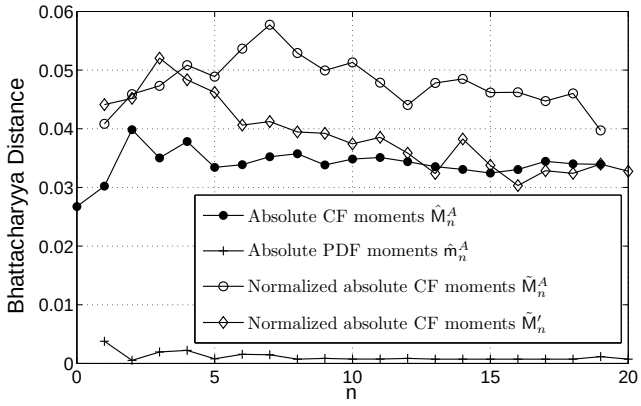

Fig. 7. Empirical Bhattacharyya distance for features $\hat{\mathsf{M}}_n^A$ from (18), $\hat{\mathsf{m}}_n^A$ from (6), $\tilde{\mathsf{M}}_n^A$ from (21), and $\tilde{\mathsf{M}}'_n$ from (22), $1 \leq n \leq 20$. Data are gathered from the first horizontal subband of the Haar-wavelet transform of 1370 photographic images, and their corresponding stegoimages with full LSB embedding ($\gamma = 1$).

1) For wavelet subbands in $\mathcal{I}_1$ and $\mathcal{I}_2$: We compare the empirical Bhattacharyya distance of several features in Fig. 6. The moments are calculated from the first diagonal subband (D_1 in $\mathcal{I}_1$) coefficients of the Haar-wavelet transform of 1370 photographic images[4] and their corresponding stegoimages generated by adding Gaussian noise $\mathcal{N}(0, 4)$ (quantized to integers) everywhere in the pixel domain ($\gamma = 1$). The RNCR ranges from -35 dB to -20 dB because the cover signal variance varies from image to image.

We first observe that $\hat{\mathsf{M}}_n^A$ from (18) is a better feature than $\hat{\mathsf{m}}_n^A$ from (6) since the empirical Bhattacharyya distance of $\hat{\mathsf{M}}_n^A$ is larger than that of $\hat{\mathsf{m}}_n^A$. This is consistent with our analysis in Section III. Also, observe that the empirical Bhattacharyya distance of the normalized CF moment $\tilde{\mathsf{M}}_n^A$ from (21) is larger than that of the unnormalized feature $\hat{\mathsf{M}}_n^A$. The reason is that the class of cover images is so broad that there is a large overlap between the range of $\hat{\mathsf{M}}_n^A$ of cover images and that of stegoimages; however, the self-calibration using the zeroth-order moment reduces the dynamic range of moments and the overlap. We also see from Fig. 6 that our $\tilde{\mathsf{M}}_n^A$ has a larger empirical Bhattacharyya distance and is a better feature than the normalized CF moment $\tilde{\mathsf{M}}'_n$ from (22) used by Harmsen and Pearlman [11]. However, it is interesting to observe that for $\hat{\mathsf{M}}_n^A$, $\tilde{\mathsf{M}}_n^A$, and $\tilde{\mathsf{M}}'_n$, the empirical Bhattacharyya distance increases till $n = 2$ or 3, then decreases as n increases; for $\hat{\mathsf{m}}_n^A$, the empirical Bhattacharyya distance decreases all the way down to zero as n increases. Therefore, for real images, higher-order moments

[4]More details on image datasets are available in Section V-A.

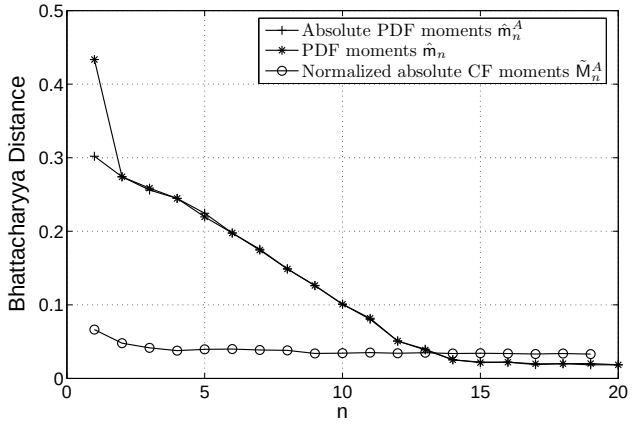

Fig. 8. Empirical Bhattacharyya distance for features $\hat{\mathsf{m}}_n$ from (4), $\hat{\mathsf{m}}_n^A$ from (6), and $\tilde{\mathsf{M}}_n^A$ from (21), $1 \leq n \leq 20$. Data are gathered from the prediction error subband eD_1 (in $\mathcal{I}_3$) of the Haar-wavelet transform of 1370 photographic images, and their corresponding stegoimages with additive Gaussian noise $\mathcal{N}(0, 4)$ (quantized to integers) in the pixel domain ($\gamma = 1$).

are not necessarily more sensitive to data embedding than lower-order moments; this partially justifies previous work [11]–[14], which used moments of the first few orders as features.

The above phenomena have been fairly consistently observed across all the wavelet subbands in $\mathcal{I}_1$ and $\mathcal{I}_2$ and for nonadditive embedding noise as well. For example, Fig. 7 shows the empirical Bhattacharyya distance of moment features from the first horizontal subband (H_1 in $\mathcal{I}_1$) when the stegoimages are generated by full LSB embedding ($\gamma = 1$). Note that the effective embedding noise for LSB embedding is dependent on images.

2) For prediction error subbands in $\mathcal{I}_3$: In Fig. 8, we compare the empirical Bhattacharyya distance of features from the prediction error subband eD_1. The stegoimages are again generated by adding Gaussian noise $\mathcal{N}(0, 4)$ (quantized to integers) everywhere in the pixel domain ($\gamma = 1$). Contrary to the case of wavelet subbands, the empirical Bhattacharyya distance of the PDF moments is consistently greater than or comparable to that of the CF moments across the nine error subbands in $\mathcal{I}_3$. Hence, the PDF moment $\hat{\mathsf{m}}_n$ from (4) is the best feature choice for $\mathcal{I}_3$.

B. Peaking Effect and Feature Selection

All moments whose associated Bhattacharyya distance is positive are potentially useful in image steganalysis. If we do so in practical image steganalysis, however, we will observe the *peaking effect*—there is an optimal number of features beyond which steganalysis performance will deteriorate. The

peaking effect is due to the finite size of the training set.[5] As the dimensionality of the feature space grows, estimating feature PDFs from the finite training set becomes harder and more inaccurate. It is an instance of the *curse of dimensionality* problem [9].

Given a finite set of training samples and a total of J available features, the problem of finding the optimal number of features has been studied extensively in the pattern recognition and machine learning literature; see [16] and references therein. It is a complicated problem involving many factors [32]: the discrimination abilities of features vary, features may have high correlations, and the optimal number of features may depend on the choice of the classifier (e.g., linear discriminant analysis, support vector machine, and so on [9]). The optimal solution can be found by an exhaustive search over 2^J possibilities, which is computationally infeasible when J is large.

We propose two methods with reduced computational complexity to find *sub*optimal feature sets and to improve image steganalysis performance. Suppose that for each image, we extract the first N moments from l wavelet (or prediction error) subbands. When N increases from 1, the steganalysis performance of the lN moments will improve till we reach some number $N = N_p$, after which the performance will degrade. We take the lN_p moments to form the feature set $\mathcal{F}_1$ and call this the *threshold selection* algorithm. Our second proposed method identifies a smaller feature set $\mathcal{F}_2 \subset \mathcal{F}_1$ that potentially has better performance, using a more sophisticated feature selection algorithm called *sequential forward floating selection* (SFFS) proposed by Pudil et al. in [33]. This method achieves better performance at the cost of higher computational complexity.

V. Experimental Results

So far, we have addressed three aspects of feature extraction: image representation, choice of features, and feature evaluation and selection. We thus propose a three-pronged approach to improve image steganalysis performance: use the multiresolution image representation $\bigcup_{i=1}^{3} \mathcal{I}_i$, the normalized absolute CF moments $\tilde{\mathsf{M}}_n^A$ in (21) for the wavelet subbands in $\mathcal{I}_1 \cup \mathcal{I}_2$ and the PDF moments $\hat{\mathsf{m}}_n$ in (4) for the prediction error subbands in $\mathcal{I}_3$, and the two feature selection algorithms of Section IV-B.

In this section, first we describe the experimental setups in Sections V-A to V-C. Then we present our experimental results in Sections V-D to V-F, by successively examining the three aforementioned aspects of feature extraction. Finally, we show in Section V-G that our optimized steganalysis method outperforms previous methods.

[5]For example, the size of the training image set is 300 in [14], 896 in [12], 1800 in [10], smaller than 2000 in [13], and 32,000 in [31].

A. Image Datasets

1) Cover image dataset: Our cover image dataset consists of 1370 256×256 8-bit graylevel photographic images, including standard test images such as *Lena*, *Baboon*, and images from the Uncompressed Colour Image Database (UCID) constructed by Schaefer and Stich [34]. Our cover images contain a wide range of outdoor/indoor and daylight/night scenes, including nature (e.g., landscapes, trees, flowers, and animals), portraits, man-made objects (e.g., ornaments, kitchen tools, architectures, cars, signs, and neon lights), etc.

2) SSIS stegoimage dataset: Our first stegoimage dataset is generated by the spread-spectrum image steganography (SSIS) method [5] proposed by Marvel et al. The embedding noise is additive and approximately Gaussian with variance $\sigma^2 = 4$. The RNCRs of 1370 SSIS stegoimages range from -35 dB to -20 dB, and the embedding fraction is $\gamma = 1$.

3) LSB stegoimage dataset: Our second stegoimage dataset is generated by full LSB embedding ($\gamma = 1$), which means that about half of the image pixels' LSBs are flipped. The RNCRs of our 1370 LSB stegoimages range from -44 dB to -29 dB.

4) F5 stegoimage dataset: Our final stegoimage dataset is generated by the steganography software F5 [35], which embeds information bits in the LSB plane of quantized DCT coefficients and employs matrix embedding to minimize the number of modified coefficients. We choose F5 because recent results [31], [36] have shown that F5 is harder to crack than other publicly available steganography software such as Jsteg [37], Outguess [38], Steghide [39], and Jphide [40]. We choose a JPEG quality factor of 80 for both cover images and stegoimages. The stegoimages are generated by embedding up to the maximum payload defined by the F5 software. The RNCRs of our 1370 F5 stegoimages range from -42 dB to -11 dB.

B. Classifier

We adopt the Fisher linear discriminator (FLD) for training and testing; see [9, Chapter 3.8.2] for full implementation details. An important step before applying the classifier is to scale the features so that they have comparable dynamic ranges. The scaling is done as follows. For a feature f, we find its maximum value $f_{\max}$ and minimum value $f_{\min}$ from all the training images. For any training or test image, the feature f is extracted and scaled as

$$\tilde{f} = \frac{f - f_{\min}}{f_{\max} - f_{\min}}. \tag{46}$$

For all the training images, $\tilde{f} \in [0,1]$; for most test images, it is expected that $\tilde{f}$ will also be between 0 and 1. This scaling step prevents features with large numerical ranges from dominating those with small numerical ranges, avoids numerical ill-conditioning, and dramatically improves classification accuracy [41].

C. Steganalysis Performance Evaluation

A receiver operating characteristic (ROC) curve displays the detection probability P_D (the fraction of the stegoimages that are correctly classified) in terms of the false alarm probability P_{FA} (the fraction of the cover images that are misclassified as stegoimages). We use the area under the ROC curve (AUC) [30]

$$\text{AUC} = \int_0^1 P_D(P_{FA})\, dP_{FA} \tag{47}$$

to measure the overall goodness of the ROC curve. The ideal ROC curve is $P_D(P_{FA}) = 1$ for any $P_{FA} \in [0,1]$ and has AUC $= 1$; the worst ROC curve is $P_D(P_{FA}) = P_{FA}$ and has AUC $= 0.5$. The AUC is connected to P_e, the average probability of error in discrimination between two equally likely hypotheses, through [30]

$$1 - \text{AUC} \leq P_e \leq \sqrt{\frac{1 - \text{AUC}}{2}}. \tag{48}$$

The steganalysis performance at low P_{FA}, say less than 0.1, is of particular interest because a steganalyzer presumably wants to keep the risk of wrongly accusing an innocent low. Thus, we plot ROC curves with P_{FA} in a logarithmic scale to illustrate the performance at small P_{FA} better.

We randomly choose 700 cover images and their corresponding stegoimages for training, then the remaining 670 cover images and their corresponding stegoimages for testing.[6] If not specified, all the following reported results are averaged over 30 such random training/testing splits in order to avoid flukes for any particular split.

D. Best Feature Choice

In this subsection, we compare only the merit of different moments (or normalized moments)—our proposed $\tilde{\mathsf{M}}_n^A$ in (21), Harmsen and Pearlman's $\tilde{\mathsf{M}}_n'$ in (22) [11], Goljan et al.'s $\hat{\mathsf{m}}_n^A$ in (6) [14], and Farid's

[6]Since the Uncompressed Color Image Database by Schaefer and Stich [34] only consists of about 1370 images and we allocate 700 of them for the training purpose, the data for $P_{FA} \leq 0.005$ are not very trustworthy since the number of test cover images are limited to 670 and P_{FA} can easily fluctuate by $1/670 = 0.0015$, give or take one false positive due to systematic errors (e.g., the limited number of images, the limited range of scenes, etc.). So in Figs. 9-15, we only show the ROC curves for $P_{FA} \geq 0.005$.

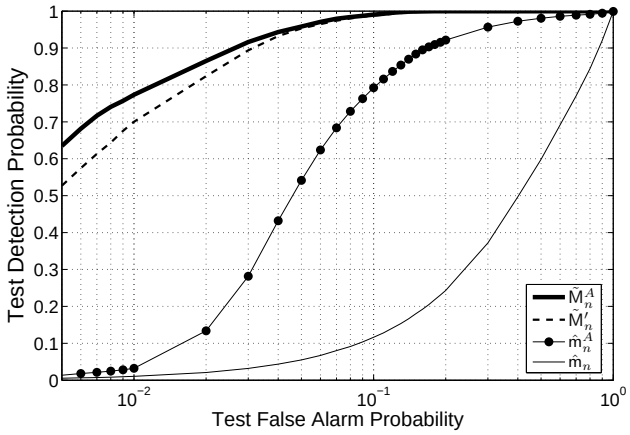

Fig. 9. Test ROC curves of feature choices—our proposed $\tilde{\mathsf{M}}_n^A$, Harmsen and Pearlman's $\tilde{\mathsf{M}}'_n$, Goljan et al.'s $\hat{\mathsf{m}}_n^A$, and Farid's $\hat{\mathsf{m}}_n$—on the cover and SSIS stegoimage dataset using image representation $\mathcal{I}_1$. In all cases, the first three moments, $1 \leq n \leq 3$, are used. From the best ROC curve to the worst one, AUC $= 0.9904$, 0.9875, 0.9166, and 0.5551, respectively.

$\hat{\mathsf{m}}_n$ in (4) [10]—with all other classifier parameters being equal. Further improvements by our proposed approach over the image steganalysis methods in [10]–[12] will be reported later in Section V-G.

1) For wavelet subbands in $\mathcal{I}_1$ and $\mathcal{I}_2$: Fig. 9 shows the test ROC curves when the first three moments are extracted from each of the 13 wavelet subbands in $\mathcal{I}_1$, that is, a total of 39 features.

Our proposed $\tilde{\mathsf{M}}_n^A$ outperforms Harmsen and Pearlman's $\tilde{\mathsf{M}}'_n$, which is consistent with Fig. 6 that shows the empirical Bhattacharyya distance of $\tilde{\mathsf{M}}_n^A$ being larger than that of $\tilde{\mathsf{M}}'_n$. The difference between $\tilde{\mathsf{M}}_n^A$ and $\tilde{\mathsf{M}}'_n$ lies with the weighting functions—$\sin^n\left(\frac{\pi k}{K}\right)$ and k^n (cf. (18) and (16)), respectively—that are applied to the magnitude of the discrete CF $|\Phi(k)|$. The former emphasizes the midfrequency components of CFs more than the latter, especially when n is small. Note that weighting functions that lead to more efficient representations of CFs and better steganalysis performance may exist.

The empirical CF moments $\tilde{\mathsf{M}}_n^A$ and $\tilde{\mathsf{M}}'_n$ are indeed far better than the empirical PDF moments $\hat{\mathsf{m}}_n^A$ and $\hat{\mathsf{m}}_n$. For $\tilde{\mathsf{M}}_n^A$, $\tilde{\mathsf{M}}'_n$, $\hat{\mathsf{m}}_n^A$, and $\hat{\mathsf{m}}_n$, the AUCs are 0.9904, 0.9875, 0.9166, and 0.5551, respectively. This confirms our conclusion in Section III that in image steganalysis, empirical CF moments are better feature choices for wavelet subbands than empirical PDF moments.

2) For prediction error subbands in $\mathcal{I}_3$: Fig. 10 shows the test ROC curves when the first three moments are extracted from each of the nine prediction error subbands in $\mathcal{I}_3$, that is, a total of 27 features. As predicted by the empirical Bhattacharyya distance in Fig. 8, the PDF moments $\hat{\mathsf{m}}_n$ outperform the CF moments $\tilde{\mathsf{M}}_n^A$, contrary to the phenomenon observed for wavelet subbands in $\mathcal{I}_1$ and $\mathcal{I}_2$.

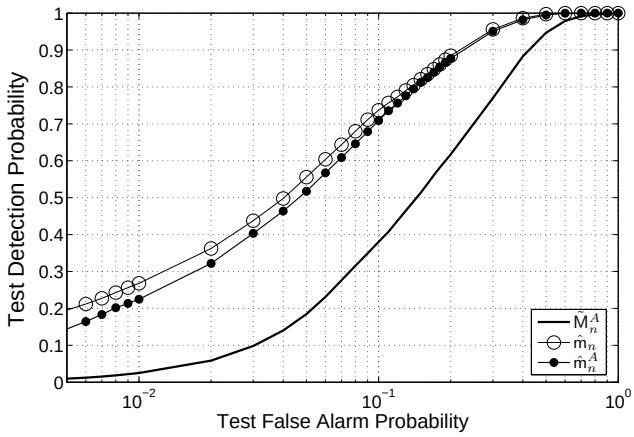

Fig. 10. Test ROC curves of feature choices—our proposed $\tilde{\mathsf{M}}_n^A$, Farid's $\hat{\mathsf{m}}_n$, and Goljan et al.'s $\hat{\mathsf{m}}_n^A$—on the cover and SSIS stegoimage dataset using image representation $\mathcal{I}_3$. In all cases, the first three moments, $1 \leq n \leq 3$, are used. From the best ROC curve to the worst one, AUC $= 0.9216$, 0.9152, and 0.8069, respectively.

In all subsequent experiments, therefore, we will associate the CF moment $\tilde{\mathsf{M}}_n^A$ as the best feature choice with the wavelet subbands in $\mathcal{I}_1$ and $\mathcal{I}_2$, and the PDF moment $\hat{\mathsf{m}}_n$ with the prediction error subbands in $\mathcal{I}_3$.

E. Multiresolution Image Representation

This subsection compares the performance obtained using the multiresolution image representations $\mathcal{I}_1$, $\mathcal{I}_1 \cup \mathcal{I}_2$, $\mathcal{I}_1 \cup \mathcal{I}_3$, and $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3$ introduced in Section II. We extract $\tilde{\mathsf{M}}_n^A$ (resp. $\hat{\mathsf{m}}_n$), $1 \leq n \leq N$, from every subband in $\mathcal{I}_1$ and $\mathcal{I}_2$ (resp. $\mathcal{I}_3$) as features. Fig. 11 shows test ROC curves in all four cases when $N = 5$. The multiresolution representation $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3$ gives the best detection performance with $\text{AUC}_{\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3} = 0.9917$, in comparison to $\text{AUC}_{\mathcal{I}_1 \cup \mathcal{I}_2} = 0.9912$, $\text{AUC}_{\mathcal{I}_1 \cup \mathcal{I}_2} = 0.9902$, and $\text{AUC}_{\mathcal{I}_1} = 0.9893$. Especially in the low P_{FA} range $[0.005, 0.1]$, using $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3$ improves P_D over using $\mathcal{I}_1$, $\mathcal{I}_1 \cup \mathcal{I}_2$, or $\mathcal{I}_1 \cup \mathcal{I}_3$.

F. Peaking Effect and Feature Selection

Fig. 12 illustrates the peaking effect (Section IV-B) for a finite set of training images: 700 cover images and 700 SSIS stegoimages. The features are $\tilde{\mathsf{M}}_n^A$ from $\mathcal{I}_1$ and $\mathcal{I}_2$, and $\hat{\mathsf{m}}_n$ from $\mathcal{I}_3$, $1 \leq n \leq N$. The feature set size is lN with $l = 26$ being the number of subbands in $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3$. Steganalysis performance measured by AUC improves when N increases, peaks at $N_p = 6$ with AUC $= 0.9948$, and eventually deteriorates quickly for $N \geq 10$.

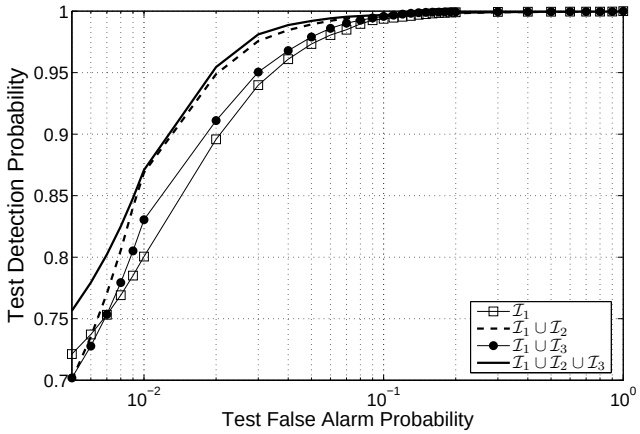

Fig. 11. Test ROC curves for multiresolution image representations $\mathcal{I}_1$, $\mathcal{I}_1 \cup \mathcal{I}_2$, $\mathcal{I}_1 \cup \mathcal{I}_3$, and $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3$ on the cover and SSIS stegoimage dataset. Features are $\tilde{\mathsf{M}}_n^A$ for $\mathcal{I}_1$ and $\mathcal{I}_2$, and $\hat{\mathsf{m}}_n$ for $\mathcal{I}_3$, $1 \leq n \leq 5$. The areas under the ROC curves are $\text{AUC}_{\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3} = 0.9917$, $\text{AUC}_{\mathcal{I}_1 \cup \mathcal{I}_2} = 0.9912$, $\text{AUC}_{\mathcal{I}_1 \cup \mathcal{I}_3} = 0.9902$, and $\text{AUC}_{\mathcal{I}_1} = 0.9893$, respectively.

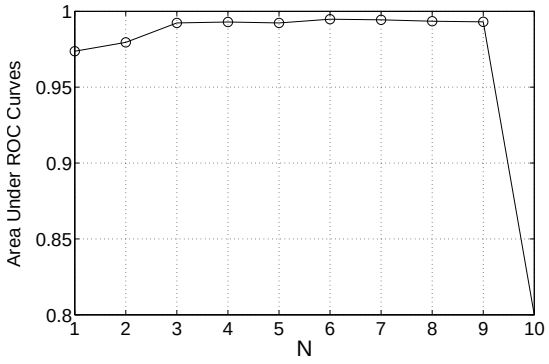

Fig. 12. AUC for the cover and SSIS stegoimage dataset, using the threshold selection procedure. Features are $\tilde{\mathsf{M}}_n^A$ from $\mathcal{I}_1$ and $\mathcal{I}_2$, and $\hat{\mathsf{m}}_n$ from $\mathcal{I}_3$, $1 \leq n \leq N$. The performance peaks at $N_p = 6$ with $\text{AUC} = 0.9948$.

The threshold feature selection algorithm that we proposed in Section IV-B identifies N_p and forms a feature subset $\mathcal{F}_1$ that consists of those $26N_p = 156$ features. Then we use the SFFS algorithm [33] to search for a smaller feature subset $\mathcal{F}_2$ with a possibly larger AUC. Note that the cost function for optimization of $\mathcal{F}_2$ is not limited to the AUC and can be an arbitrary objective, e.g., the detection probability P_D for a fixed false alarm probability P_{FA}. In our example, $|\mathcal{F}_2| = 73$ with $\text{AUC} = 0.994$. The test ROC curves for the feature sets $\mathcal{F}_1$ and $\mathcal{F}_2$ are shown in Fig. 13. The performance of SFFS is vastly better in the low P_{FA} range. However, the SFFS algorithm consumes hours, even days, in

our simulations, in contrast to the minutes taken by the threshold selection approach. Hence, there is a tradeoff between performance and training time if computational complexity is a concern.

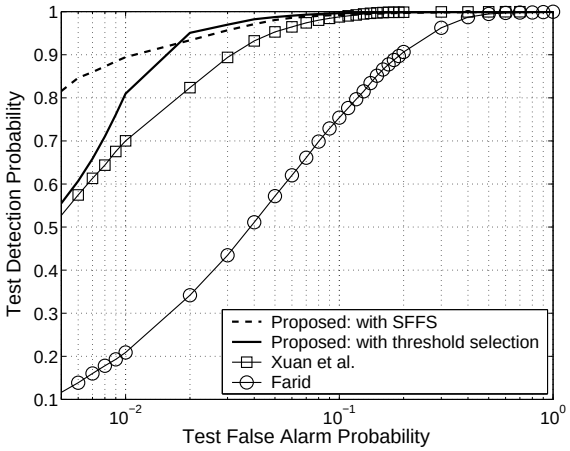

Fig. 13. Test ROC curves on the cover and SSIS stegoimage dataset. Our proposed method extracts $\tilde{\mathsf{M}}_n^A$ (resp. $\hat{\mathsf{m}}_n$) from $\mathcal{I}_1$ and $\mathcal{I}_2$ (resp. $\mathcal{I}_3$), $1 \leq n \leq N$. The threshold selection algorithm takes $N_p = 8$, and its feature set $\mathcal{F}_1$ has 208 features that yield AUC $= 0.9922$; the SFFS algorithm has a feature set $\mathcal{F}_2 \subset \mathcal{F}_1$ with 73 features that yield AUC $= 0.994$. Xuan et al.'s method [12] extracts $13N$ features $\tilde{\mathsf{M}}'_n$, $1 \leq n \leq N$, from $\mathcal{I}_1$; $N = 3$ leads to 39 features and yields AUC $= 0.9875$. Farid's method [10] extracts $18N$ features $\hat{\mathsf{m}}_n$, $1 \leq n \leq N$, from $\mathcal{I}_3$ and all the high-pass subbands in $\mathcal{I}_1$; $N = 4$ leads to 72 features and yields AUC $= 0.9249$.

G. Comparison with State-of-the-Art Methods

Finally, we propose a method that combines a multiresolution image representation $\bigcup_{i=1}^{3} \mathcal{I}_i$, a feature choice $\tilde{\mathsf{M}}_n^A$ (resp. $\hat{\mathsf{m}}_n$) for $\mathcal{I}_1$ and $\mathcal{I}_2$ (resp. $\mathcal{I}_3$), and a feature selection algorithm such as the threshold selection and SFFS algorithms. We compare the steganalysis performance of our method to Xuan et al.'s method[7] [12] and Farid's method [10] on three kinds of steganographic embedding algorithms. Clearly, from Figs. 13-15, our proposed method consistently outperforms these two state-of-the-art methods.

Fig. 13 shows the test ROC curves for our cover image set and SSIS stegoimage set. From the best ROC curve to the worst, the AUCs are 0.994 for the SFFS algorithm, 0.9922 for our threshold selection algorithm, 0.9875 for Xuan et al.'s method, and 0.9249 for Farid's method. Fixing $P_{FA} = 0.01$, our methods with the threshold selection and SFFS algorithms yield $P_D = 0.81$ and $P_D = 0.895$ respectively, which are significantly better than Xuan et al.'s $P_D = 0.7$ and Farid's $P_D = 0.21$.

[7]Xuan et al.'s method [12] is an improved version of Harmsen and Pearlman's method [11]. The former method uses $\tilde{\mathsf{M}}'_n$'s with $1 \leq n \leq 3$ while the latter only uses $\tilde{\mathsf{M}}'_1$.

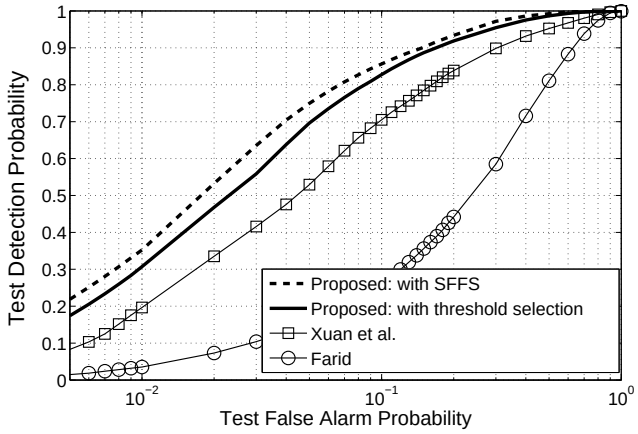

Fig. 14. Test ROC curves on the cover and LSB stegoimage dataset. Our proposed threshold selection algorithm extracts $26N_p$ features: $\tilde{\mathsf{M}}_n^A$ from $\mathcal{I}_1 \cup \mathcal{I}_2$ and $\hat{\mathsf{m}}_n$ from $\mathcal{I}_3$, $1 \leq n \leq N_p$; $N_p = 6$ yields a 156-feature set $\mathcal{F}_1$ and AUC $= 0.9365$. The SFFS algorithm is applied to obtain a feature set $\mathcal{F}_2 \subset \mathcal{F}_1$ with 43 features that yield AUC $= 0.9483$. Xuan et al.'s method [12] extracts $13N$ features $\tilde{\mathsf{M}}'_n$, $1 \leq n \leq N$, from $\mathcal{I}_1$; $N = 3$, AUC $= 0.8901$. Farid's method [10] extracts $18N$ features $\hat{\mathsf{m}}_n$, $1 \leq n \leq N$, from $\mathcal{I}_3$ and all the high-pass subbands in $\mathcal{I}_1$; $N = 4$, AUC $= 0.7122$.

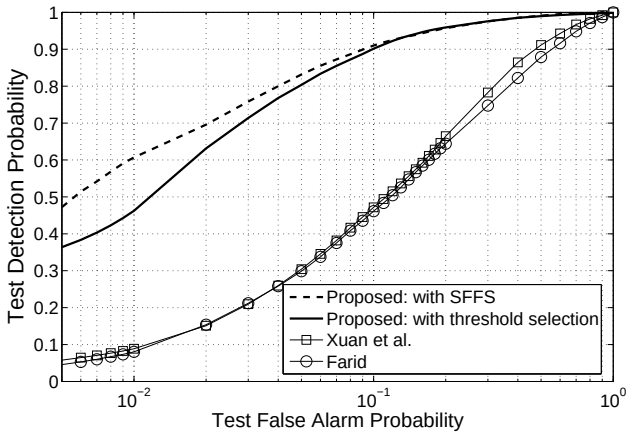

Fig. 15. Test ROC curves on the cover and F5 stegoimage dataset. Our proposed threshold selection algorithm extracts $26N_p$ features $\tilde{\mathsf{M}}_n^A$, $1 \leq n \leq N_p$, from the multiresolution image representation $\mathcal{I}_1 \cup \mathcal{I}_2 \cup \mathcal{I}_3$; $N_p = 10$ leads to a 260-feature set $\mathcal{F}_1$; AUC $= 0.9591$. The SFFS algorithm is applied to obtain a feature set $\mathcal{F}_2 \subset \mathcal{F}_1$ with 96 features and yields AUC $= 0.9657$. Xuan et al.'s method [12] extracts $13N$ features $\tilde{\mathsf{M}}'_n$, $1 \leq n \leq N$, from $\mathcal{I}_1$; $N = 3$, AUC $= 0.8132$. Farid's method [10] extracts $18N$ features m_n, $1 \leq n \leq N$, from $\mathcal{I}_3$ and all the high-pass subbands in $\mathcal{I}_1$; $N = 4$, AUC $= 0.7934$.

Fig. 14 shows the steganalysis results of LSB embedding, where the embedding noise is dependent on the cover image. Again, our steganalysis method with feature selection algorithms outperforms other methods: fixing $P_{FA} = 0.01$, we obtain $P_D = 0.31$ using the threshold selection algorithm ($N_p = 6$, $|\mathcal{F}_1| = 156$) and $P_D = 0.35$ using the SFFS algorithm ($|\mathcal{F}_2| = 43$), compared to Xuan et al.'s $P_D = 0.2$ and Farid's $P_D = 0.03$. The AUC is 0.9484 for our method with the SFFS algorithm, 0.9365 for our method with the threshold selection algorithm, 0.8901 for Xuan et al.'s method, and 0.7122 for Farid's method.

Fig. 15 shows the steganalysis results of F5 embedding. Again, our steganalysis method with feature selection algorithms outperforms other methods: fixing $P_{FA} = 0.01$, we obtain $P_D = 0.47$ using the threshold selection algorithm ($N_p = 10$, $|\mathcal{F}_1| = 260$) and $P_D = 0.6$ using the SFFS algorithm ($|\mathcal{F}_2| = 96$), compared to Xuan et al.'s $P_D = 0.09$ and Farid's $P_D = 0.08$. The AUC is 0.9657 for our method with the SFFS algorithm, 0.9591 for our method with the threshold selection algorithm, 0.8132 for Xuan et al.'s method, and 0.7934 for Farid's method.

VI. Discussion

In practice, both the steganographer and steganalyzer have only partial knowledge of the cover signal statistics. However, the steganalyzer may extract appropriate features and learn their statistics from training data. The steganalyzer's success largely depends on the ability to identify the most changed statistics by embedding and to extract reliable features that are sensitive to these changes. For example, multiresolution representations of photographic images are sparse, which implies that the PDF of wavelet coefficients exhibits a sharp peak near zero. In contrast, the embedding noise PDF is smooth for many watermarking and steganographic algorithms such as spread-spectrum, dithered quantization index modulation, $\pm k$ embedding, etc. Thus, a prominent characteristic of stegoimages is that the marginal PDF of their wavelet coefficients is smoothed.

We analyzed the statistical effects of additive embedding and explained why empirical characteristic function moments of wavelet coefficients are better choices than empirical PDF moments in image steganalysis. We also studied which moment orders, higher or lower, are more suitable features. And, in light of the inevitable peaking effect caused by the finite training sample size, we explored some feature selection algorithms to find informative, low-dimensional feature sets. In addition, we proposed a new multiresolution image representation that is more informative than existing ones. Our image steganalysis results—on both additive embedding represented by the spread-spectrum method and nonadditive embedding represented by the LSB embedding method and the F5 embedding algorithm—demonstrated the

effectiveness of our method: it has significantly better performance than methods recently proposed by Farid [10] and Xuan et al. [12].

Of course, the proposed features and steganalysis methods in this paper are by no means optimal. To achieve better performance, various improvements can be made. For example, one could look for new features that are more sensitive to weak-noise embedding, use better classifiers, etc. However, the strategy of a steganalyzer will remain the same: describe the cover signal statistics as completely as possible and seek a small number of informative, reliable features as inputs to the classifier.

APPENDIX I
CALCULATION OF m_n^A, M_n^A, $r_{\mathsf{m},n}$, $r_{\mathsf{M},n}$, AND A_n FOR GAUSSIAN COVER SIGNALS

For a Gaussian distributed random variable, $S \sim \mathcal{N}(0, \sigma^2)$, its n^{th} absolute PDF moment $\mathsf{m}_{n,S}^A$ is given by

$$\mathsf{m}_{n,S}^A = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} |x|^n dx.$$

Simple calculus yields

$$\mathsf{m}_{n,S}^A = \begin{cases} \sqrt{\frac{2}{\pi}}\,\sigma & \text{for } n = 1, \\ \sqrt{\frac{2}{\pi}}\,\sigma^n \prod_{i=1}^{\frac{n-1}{2}} 2i & \text{for odd } n > 1, \\ \sigma^n \prod_{i=1}^{\frac{n}{2}} (2i-1) & \text{for even } n > 1. \end{cases} \tag{49}$$

The CF of $S \sim \mathcal{N}(0, \sigma^2)$ is given by

$$\begin{aligned} \Phi_S(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{x^2}{2\sigma^2}} \cdot e^{jtx} dx \\ &= e^{-\frac{\sigma^2 t^2}{2}}, \quad t \in \mathbb{R}. \end{aligned}$$

Thus the n^{th} absolute moment of the CF is given by

$$\mathsf{M}_{n,S}^A = \int_{-\infty}^{\infty} e^{-\frac{\sigma^2 t^2}{2}} |t|^n dt.$$

Similarly, simple calculus yields

$$\mathsf{M}_{n,S}^A = \begin{cases} 2\,\sigma^{-2} & \text{for } n = 1, \\ 2\,\sigma^{-(n+1)} \prod_{i=1}^{\frac{n-1}{2}} 2i & \text{for odd } n > 1, \\ \sqrt{2\pi}\,\sigma^{-(n+1)} \prod_{i=1}^{\frac{n}{2}} (2i-1) & \text{for even } n > 1. \end{cases} \tag{50}$$

For the stegosignal $X = S + Z_\gamma$, where

$$Z_\gamma \sim (1-\gamma)\,\delta(0) + \gamma\,\mathcal{N}(0, \text{RNCR}\ \sigma^2),$$

$\gamma \in [0,1]$, and RNCR ≥ 0, we have

$$X \sim \begin{cases} \mathcal{N}(0,\sigma^2) & \text{with probability } 1-\gamma, \\ \mathcal{N}\left(0,(1+\text{RNCR})\sigma^2\right) & \text{with probability } \gamma. \end{cases} \tag{51}$$

The moments $\mathsf{m}_{n,X}^A$ and $\mathsf{M}_{n,X}^A$ are obtained by applying (49) and (50):

$$\mathsf{m}_{n,X}^A = c_n \left[1-\gamma+\gamma(1+\text{RNCR})^{\frac{n}{2}}\right]\sigma^n$$

and

$$\mathsf{M}_{n,X}^A = C_n \left[1-\gamma+\gamma(1+\text{RNCR})^{-\frac{n+1}{2}}\right]\sigma^{-(n+1)},$$

where c_n and C_n are the respective constant terms in (49) and (50).

Therefore, the ratio $r_{\mathsf{m},n}$ defined in (35) is given by

$$r_{\mathsf{m},n}(\gamma) = \frac{\mathsf{m}_{n,X}^A}{\mathsf{m}_{n,S}^A} = 1-\gamma+\gamma(1+\text{RNCR})^{\frac{n}{2}}. \tag{52}$$

Clearly, $r_{\mathsf{m},n}(0)=1$, $r_{\mathsf{m},n}(1)=(1+\text{RNCR})^{\frac{n}{2}}$, and $r_{\mathsf{m},n}(\gamma)$ is a monotonically increasing function of γ. Similarly, the ratio $r_{\mathsf{M},n}$ defined in (36) is given by

$$r_{\mathsf{M},n}(\gamma) = \frac{\mathsf{M}_{n,S}^A}{\mathsf{M}_{n,X}^A} = \frac{1}{1-\gamma+\gamma(1+\text{RNCR})^{-\frac{n+1}{2}}}. \tag{53}$$

Clearly, $r_{\mathsf{M},n}(0)=1$, $r_{\mathsf{M},n}(1)=(1+\text{RNCR})^{\frac{n+1}{2}}$, and $r_{\mathsf{M},n}(\gamma)$ is also a monotonically increasing function of γ.

From (52) and (53), the ratio $A_n(\gamma) \stackrel{\triangle}{=} \frac{r_{\mathsf{M},n}}{r_{\mathsf{m},n}}$ is then given by

$$A_n(\gamma) = \frac{\left[1-\gamma+\gamma(1+\text{RNCR})^{-\frac{n+1}{2}}\right]^{-1}}{\left[1-\gamma+\gamma(1+\text{RNCR})^{\frac{n}{2}}\right]}, \tag{54}$$

for which $A_n(0)=1$ and $A_n(1)=(1+\text{RNCR})^{\frac{1}{2}} \geq 1$. The denominator of $A_n(\gamma)$, or $A_n^{-1}(\gamma)$ in this case, is a *quadratic* function of γ. Therefore $A_n(\gamma)=1$ only if $\gamma=0$ or

$$\gamma = \gamma_1 \stackrel{\triangle}{=} 1 - \frac{(1+\text{RNCR})^{\frac{n+1}{2}} - (1+\text{RNCR})^{\frac{n}{2}}}{\left[(1+\text{RNCR})^{\frac{n}{2}}-1\right]\left[(1+\text{RNCR})^{\frac{n+1}{2}}-1\right]}.$$

Clearly, $\gamma_1 \leq 1$ when RNCR > 0. Simple algebra shows that $\gamma_1 < 0$ if and only if

$$(1+\text{RNCR})^{-\frac{n+1}{2}} + (1+\text{RNCR})^{\frac{n}{2}} < 2.$$

The second derivative of $A_n^{-1}(\gamma)$ is given by

$$\frac{d^2 A_n^{-1}(\gamma)}{d\gamma^2} = 2[(1+\text{RNCR})^{\frac{n}{2}}-1][(1+\text{RNCR})^{-\frac{n+1}{2}}-1], \tag{55}$$

which is negative when RNCR > 0. Hence, $A_n^{-1}(\gamma)$ is a concave quadratic function of γ, with $A_n^{-1}(\gamma) = 1$ at $\gamma = 0$ and γ_1. It follows that the function $A_n(\gamma)$ is convex, and if $\gamma > \max(0, \gamma_1)$, $A_n(\gamma)$ is greater than 1 and monotonically increasing with γ.

REFERENCES

[1] D. Kahn, *The Codebreakers*. New York: Macmillan, 1967.

[2] J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color and gray-scale images," *IEEE Multimedia*, vol. 8, no. 4, pp. 22–28, Oct. 2001.

[3] S. Dumitrescu, X. Wu, and Z. Wang, "Detection of LSB steganography via sample pair analysis," *IEEE Trans. Signal Processing*, vol. 51, no. 7, pp. 1995–2007, July 2003.

[4] I. J. Cox, J. Killian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[5] L. M. Marvel, C. G. Boncelet, and C. T. Retter, "Spread spectrum image steganography," *IEEE Trans. Image Processing*, vol. 8, no. 8, pp. 1075–1083, Aug. 1999.

[6] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[7] Y. Wang and P. Moulin, "Steganalysis of block-structured stegotext," in *Proc. of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI*, San Jose, CA, Jan. 2004, pp. 477–488.

[8] P. Moulin and A. Briassouli, "A stochastic QIM algorithm for robust, undetectable image watermarking," in *Proc. Int. Conf. on Image Processing*, vol. 2, Singapore, Oct. 2004, pp. 1173–1176.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2001.

[10] H. Farid, "Detecting hidden messages using higher-order statistical models," in *Proc. IEEE Int. Conf. on Image Processing*, New York, Sept. 2002, pp. 905–908.

[11] J. J. Harmsen and W. A. Pearlman, "Steganalysis of additive noise modelable information hiding," in *Proc. of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI*, San Jose, CA, Jan. 2003, pp. 131–142.

[12] G. Xuan, Y. Q. Shi, J. Gao, D. Zou, C. Yang, Z. Zhang, P. Chai, C. Chen, and W. Chen, "Steganalysis based on multiple features formed by statistical moments of wavelet characteristic functions," in *Proc. Information Hiding Workshop*, Barcelona, Spain, June 2005, pp. 262–277.

[13] T. Holotyak, J. Fridrich, and S. Voloshynovskiy, "Blind statistical steganalysis of additive steganography using wavelet higher order statistics," in *Proc. of the 9th IFIP TC-6/TC-11 Conference on Communications and Multimedia Security*, Salzburg, Austria, Sept. 2005, pp. 273–274.

[14] M. Goljan, J. Fridrich, and T. Holotyak, "New blind steganalysis and its implications," in *Proc. of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents VI*, San Jose, CA, Jan. 2006, pp. 1–13.

[15] K. Sullivan, U. Madhow, S. Chandrasekaran, and B. Manjunath, "Steganalysis for Markov cover data with applications to images," *IEEE Trans. Inform. Forensics and Security*, vol. 1, no. 2, pp. 275–287, June 2006.

[16] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[17] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Upper Saddle River, New Jersey: Prentice Hall, 1995.

[18] J. E. Gentle, W. Härdle, and Y. Mori, *Handbook of Computational Statistics*. New York: Springer, 2004.

[19] T. Sharp, “An implementation of key-based digital signal steganography,” in *Proc. 4th Int. Workshop on Information Hiding*, Pittsburgh, PA, Apr. 2001, pp. 13–26.

[20] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. New York: John Wiley & Sons, 1994.

[21] J. J. Eggers, R. Bauml, R. Tzschoppe, and B. Girod, “Scalar Costa scheme for information embedding,” *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 1003–1019, Apr. 2003.

[22] N. G. Ushakov, *Selected Topics in Characteristic Functions*. Utrecht, The Netherlands: VSP, 1999.

[23] E. Lukacs, *Characteristic Functions*, 2nd ed. New York: Hafner Pub. Co., 1970.

[24] S. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 7, pp. 674–693, July 1989.

[25] S. M. Lopresto, K. Ramchandran, and M. T. Orchard, “Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework,” in *Proc. IEEE Data Compression Conf.*, Snowbird, UT, Mar. 1997, p. 271.

[26] P. Moulin and J. Liu, “Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors,” *IEEE Trans. Inform. Theory*, vol. 45, no. 3, pp. 909–919, Mar. 1999.

[27] E. P. Simoncelli, “Higher-order statistical models of visual images,” in *Proc. IEEE Signal Processing Workshop on Higher-Order Statistics*, Ceasarea, Israel, June 1999, pp. 54–57.

[28] M. Ben-Bassat, “Use of distance measures, information measures and error bounds on feature evaluation,” in *Handbook of Statistics: Classification, Pattern Recognition and Reduction of Dimensionality*, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam: North-Holland Publishing Company, 1987, pp. 773–791.

[29] H. V. Poor, *An Introduction to Detection and Estimation Theory*. London, UK: Springer-Verlag, 1994.

[30] J. H. Shapiro, “Bounds on the area under the ROC curve,” *J. Opt. Soc. Am. A*, vol. 16, pp. 53–57, Jan. 1999.

[31] S. Lyu and H. Farid, “Steganalysis using higher-order image statistics,” *IEEE Trans. Inform. Forensics and Security*, vol. 1, no. 1, pp. 111–119, Mar. 2006.

[32] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, “Optimal number of features as a function of sample size for various classification rules,” *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, Apr. 2005.

[33] P. Pudil, J. Novovicova, and J. Kittler, “Floating search methods in feature selection,” *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, Nov. 1994.

[34] G. Schaefer and M. Stich, “UCID—An uncompressed colour image database,” in *Proc. of the SPIE, Storage and Retrieval Methods and Applications for Multimedia*, San Jose, CA, Jan. 2004, pp. 472–480.

[35] A. Westfeld. F5. [Online]. Available: http://wwwrn.inf.tu-dresden.de/~westfeld/f5.html

[36] D. Fu, Y. Q. Shi, D. Zou, and G. Xuan, “JPEG steganalysis using empirical transition matrix in block DCT domain,” in *Proc. Int. Workshop on Multimedia Signal Processing*, Victoria, BC, Canada, Oct. 2006.

[37] D. Upham. Jsteg. [Online]. Available: ftp://ftp.funet.fi/pub/crypt/steganography/

[38] N. Provos. Outguess. [Online]. Available: http://www.outguess.org

[39] S. Hetzl. Steghide. [Online]. Available: http://steghide.sourceforge.net

[40] A. Latham. Jpeg hide-and-seek. [Online]. Available: http://linux01.gwdg.de/~alatham/stego.html

[41] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf