Generic Adoption of Spatial Steganalysis to Transformed Domain

Andreas Westfeld

Technische Universität Dresden
Faculty of Computer Science
Institute of Systems Architecture
01062 Dresden, Germany
mailto:westfeld@inf.tu-dresden.de

Abstract. There are several powerful steganalytic methods for images in the spatial domain, which are based on higher order statistics. We propose a generic methodology to prepare higher order steganalytic methods from spatial domain for application in the transformed domain. This paper presents 72 new systematically designed methods that are derived from the spatial domain. Their reliability and the precision of their length estimation is evaluated based on 1700 million attacks. We present the contribution of the proposed methods in terms of detection power and precision compared to prior art and determine how properties like image size and JPEG quality influence the ranking of the proposed attacks.

1 Steganography in JPEG Files

Steganography is the art and science of invisible communication. Its aim is the transmission of information embedded invisibly into carrier data. The goal of *steganalysis* is to discover steganographic alterations to carrier data. We can distinguish *targeted* attacks, which are fine-tuned to one particular embedding method, and *blind* (or universal) attacks, which detect steganographic changes using a classifier that is trained with features from known steganograms and carrier media. Targeted methods are often used to extract features for the training. The distinction of blind and targeted attacks is blurred in this respect.

The most simple embedding methods are at the same time the most widely used. Common carrier data include digitised images or audio files. In many cases the image data are stored in the spatial domain (e.g., in BMP or PNM format). The message is embedded by replacing the least significant bits of the colour or brightness values by the (encrypted) message bits. The human eye will usually not notice the changes that are caused by the embedding in the image.

Sometimes, however, the embedding can still be detected by statistical means. One of the first steganalytic attacks was based on the global histogram of the colour or brightness values and is also applicable to DCT values in JPEG files [1]. JPEG files are more common as e-mail attachments than spatial domain images. As such, they are a more important carrier medium for steganography.

K. Solanki, K. Sullivan, and U. Madhow (Eds.): IH 2008, LNCS 5284, pp. 161–177, 2008.

All previous targeted attacks to Jsteg embedding [2] only evaluated global frequencies of DCT values. They ignored the location of these values and the spatial frequency in the context in which they occurred. This is also called marginal distributions or first order statistics. Part of this concept are, e.g., the attack by Zhang and Ping [3] (ZP, cf. Sect. 3.1), the attack by Yu et al. [4] (Yu, cf. Sect. 3.2), and the category attack by Lee et al. [5] (CA, cf. Sect. 3.3) together with its generalisation (GCA) [6]. Blind attacks also employ *higher* order statistics, however, unlike the aforementioned attacks, they are rarely used to estimate the length of the embedded message [7,8,9]. A recent evaluation concluded that blind attacks to JPEG images are less reliable than the CA, at least for heterogeneous sources of images [6].

There are several steganalytic techniques for the spatial domain that use higher order statistics, e.g., RS [10], WS [11], SPA [12,13], and Pairs [14]. To our knowledge, they have never been applied directly to coefficients of the frequency domain before, while this was done from the beginning with the histogram based chi-square attack. It seems not very obvious how to extract the sample sequences from the DCT frequency domain to use them with spatial domain higher order statistical attacks. It is even doubtful because the goal of the JPEG compression is to remove irrelevance, i.e., to reduce the local dependency that is inherent to images and exploited by the higher order attacks. We tried it, though, contributing partly considerable improvements in terms of reliability and precision.

Although we think the proposed methodology is applicable in general to exploit local dependencies in JPEG files, there is essentially only one steganographic algorithm used for both, spatial *and* DCT domain: LSB replacement. At the same time this is the most completely analysed algorithm in the spatial domain, which offers a wide range of different steganalytical approaches for testing our methodology. While we apply these approaches to Jsteg [2] in our evaluation, an adaption to JPhide [15] is easy, however, the ranking of attacks could be different.

This paper is organised as follows: Section 2 looks at several scanpaths (intra block and inter block) to exploit local dependencies in the DCT domain. Section 3 briefly describes histogram-based attacks to JPEG files. In Sect. 4 we derive 72 new attacks for JPEG media (6 fundamental attacks × 6 scanpaths × with and without DC coefficients) from the spatial domain attacks mentioned at the beginning (RS, WS, SPA, and Pairs). The new attacks are evaluated in Sect.5. All in all we produced 1700 million length estimations for a variety of different messages of several lengths in a set of images with different quality and size to pick the best scanpath and the most suitable attack from the spatial domain. The paper is concluded in Sect. 6.

2 Methods

2.1 Local Correlation

A well-known feature of spatial domain images is the dependency between neighbouring pixels. This has been exploited for steganalysis using fractal scanpaths [16].

Correlations of DCT coefficients have been studied in the context of entropy coding in JPEG images. Tu and Tran distinguish three kinds of correlations [17]:

Intrablock or intersubband correlation. The magnitudes of (unquantised) AC coefficients decrease as the frequency increases. This results in a dependency between neighbouring coefficients. However, the intrablock dependency is limited to a small block of 8×8 pixels only. It is also weak, since the coefficients of one block belong to different, (nearly) orthogonal subbands. Most coefficients are quantised to zero and do not carry much information.

Interblock or intrasubband correlation. Low frequency coefficients and their neighbours are mutually dependent on each other within the same subband. Generally, the interblock dependency is much stronger than the intrablock dependency. However, it is limited especially for higher frequencies because, compared to the raster of pixels, their distance is increased to the block raster (e.g., factor 8).

Sign correlation. Deever and Hemami analysed the sign behaviour of wavelet coefficients [18]. A similar intrasubband dependency can be found for low frequency coefficients in JPEG images.

Fridrich uses co-occurrence matrices of neighbouring coefficients in the same subband in her blind attack using 23 DCT features [7]. Even individual co-occurrence features very reliably detect F5, Outguess, and MB1 [19,20,21] for full embedding rate.

Fu et al. rearrange the DCT coefficients to exploit both, intrablock and interblock dependency for their blind attack based on Markov empirical transition matrices [22]. They apply a zigzag scan for intrablock dependency as well as a slalom scan for interblock dependency.

2.2 Scanning Orders for DCT Coefficients

The JPEG compression reorders the DCT coefficients of one 8×8 block from their natural order to zigzag order (cf. Fig. 1) to take advantage of the intrablock dependency. We scan DCT coefficients in natural and zigzag order to exploit intrablock dependency by our attacks.

The dependency in different subbands may have different orientation. From spatial domain we know that recursive scanning paths increase the dependency between pixel values [16]. This is also valid for DC coefficients. AC coefficients might favour the interblock dependency in vertical or horizontal lines. We apply recursive and linear scanning paths (cf. Fig. 2) to evaluate the interblock dependency by the attacks that we propose in the sequel.

3 Prior Attacks to Jsteg-Like Embedding

Jsteg is probably the first algorithm that was developed for JPEG images [2]. It overwrites the least significant bits of the quantised DCT coefficients with secret message bits. Jsteg is detectable by the chi-square attack [1]. However, the attack

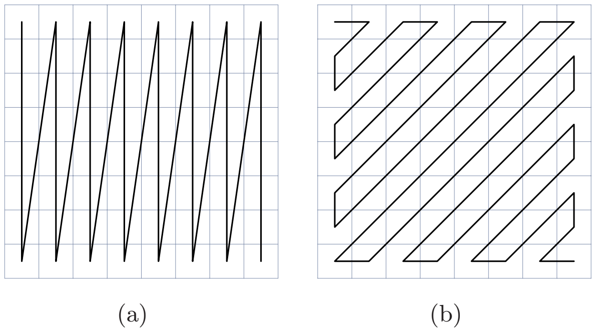

Fig. 1. Intrablock scanpath in natural order (a) and zigzag order (b)

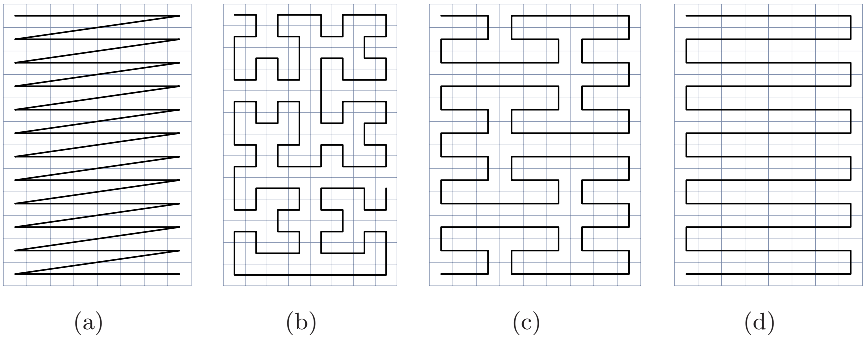

Fig. 2. Row by row scan (a), fractal scan along Hilbert 1 curve (b) and Hilbert 2 curve (c), slalom scan (d)

fails if the message is pseudo-randomly straddled over all DCT coefficients, which is called "randomised Jsteg." While there exist several powerful steganalytic techniques for the spatial domain [10,11,12,14], which *can* detect such straddled messages and even estimate their length, there are only a few techniques for randomised Jsteg, which consider merely the marginal statistics of the DCT coefficients, but can also estimate lengths.

In the following subsections we will consider three known detection methods for randomised Jsteg as far as it is essential for the implementation used here. The Method of Zhang and Ping divides the histogram into two interleaved groups, the attack of Yu et al. estimates the density function of the quantised values on a steganographically invariant basis, and the category attack compares equalising and complementary pairs of values with each other. All three can estimate the length of the embedded message. For further details please refer to the description in the respective original publications. The methods start the detection based on the histogram $h(k)$ of the quantised DCT coefficients $k \in \mathbb{Z}$ from a JPEG image.

3.1 Method by Zhang and Ping

We define two sets of values

$$L = \{i > 0 | i \text{ even}\} \cup \{i < 0 | i \text{ odd}\} \text{ and} \tag{1}$$

$$R = \{i > 0 | i \text{ odd}\} \cup \{i < 0 | i \text{ even}\}. \tag{2}$$

The proportion p of the capacity used by randomised Jsteg is estimated by

$$p = \frac{1}{h(1)} \cdot \left(\sum_{r \in R} h(r) - \sum_{l \in L} h(l) \right). \tag{3}$$

3.2 Method by Yu et al.

Inspired by Sallee [21], Yu et al. model the density of quantised JPEG coefficients by a modified, generalised Cauchy distribution

$$f(k, \pi, s) = \frac{\pi - 1}{2s} \left(\left| \frac{k}{s} \right| + 1 \right)^{-\pi}. \tag{4}$$

Its scale parameter s and location parameter π are computed for all DCT coefficients (globally, for all subbands together) by a maximum likelihood estimation over $h(0)$, $h(1)$, and the low precision bins $H(i) = h(2i) + h(2i+1)$ for $i \neq 0$. First we estimate the histogram of the carrier image:

$$\hat{h}(i) = \begin{cases} h(i) & \text{for } i \in \{0, 1\}, \\ (h(2i) + h(2i+1)) \frac{\int_{i-0.5}^{i+0.5} f(x,\pi,s) \mathrm{d}x}{\int_{i-0.5}^{i+1.5} f(x,\pi,s) \mathrm{d}x} & \text{otherwise.} \end{cases} \tag{5}$$

The proportion p of the capacity used by randomised Jsteg is estimated by

$$p = \frac{\sum_{i \neq 0} (h(2i) - \hat{h}(2i)) \cdot (\hat{h}(2i) - \hat{h}(2i+1))}{\sum_{i \neq 0} (\hat{h}(2i) - \hat{h}(2i+1))^2}. \tag{6}$$

3.3 Category Attack

We renumber the bins and create an adjusted version of the histogram that excludes the values 0 and 1, which are not changed by Jsteg:

$$h'(k) = \begin{cases} h(k) & \text{for } k < 0, \\ h(k-2) & \text{for } k > 1. \end{cases} \tag{7}$$

The proportion p of the capacity used by randomised Jsteg is estimated using the statistics s for *shifted* categories that are not equalised and *induced* categories that are equalised by Jsteg embedding:

$$s(k) = \frac{(h'(k) - h'(k+1))^2}{h'(k) + h'(k+1)}, \tag{8}$$

$$p = \frac{\sum_i s(2i) - \sum_i s(2i-1)}{\sum_i s(2i) + \sum_i s(2i-1)}. \tag{9}$$

4 Attacks Derived from Prior Spatial Domain Methods

While there have been several proposals to use higher order statistics in the spatial domain, no specific method exists to the best of our knowledge for the DCT domain. There are only some blind detection methods, which work less reliably than the best histogram based attack [6], require training, and furthermore are no precise estimators for the length of the embedded message.

In the following we will apply several techniques that are known from the spatial domain to (selected) DCT coefficients, which have been scanned along different paths (intrablock and interblock).

Firstly we describe how to prepare DCT coefficients "palatable" for attacks that have only been applied to brightness or colour values so far. In the subsequent subsections we describe the adaptation of the spatial domain attacks RS, Pairs, SPA, and WS to DCT coefficients, as far as this is essential for the implementation used here. For further details please refer to the detailed description in the respective original publications.

4.1 Preprocessing of Coefficients

Let $\mathbf{x}$ be a sequence of DCT coefficients, scanned from a JPEG file in one of the orders described in Sect. 2.2. To ease the application of spatial domain attacks, we delete all elements in $\mathbf{x}$ that are 0 or 1 (i.e., steganographically unused by Jsteg), yielding $\mathbf{x}'$. Then we change all coefficients to be non-negative and get the normalised sequence $\mathbf{d}$ with the elements:

$$d_i = \begin{cases} x'_i - x_{\min} & \text{if } x'_i < 0, \\ x'_i - x_{\min} - 2 & \text{if } x'_i > 1, \end{cases} \tag{10}$$

$$\text{where} \quad x_{\min} = \begin{cases} \min \mathbf{x} & \text{if } \min \mathbf{x} \text{ is even}, \\ \min \mathbf{x} - 1 & \text{otherwise.} \end{cases} \tag{11}$$

Lee et al. reported that the CA and the ZP attack showed increased reliability when only AC coefficients were used in the histogram [6]. Therefore, we derive another sequence $\mathbf{d}_{\text{AC}}$ from the sequence $\mathbf{d}$ of all normalised coefficients, in which we delete all DC coefficients.

4.2 JRS

The RS attack by Fridrich et al. [10] counts the number of regular and singular groups of pixels before and after flipping operations. There are several types of flipping operations: One flips induced pairs like LSB embedding, $f_1\colon 0 \leftrightarrow 1$, $2 \leftrightarrow 3$, $4 \leftrightarrow 5$, …, another one flips shifted pairs, $f_{-1}\colon -1 \leftrightarrow 0$, $1 \leftrightarrow 2$, $3 \leftrightarrow 4$, …, the third one is the identity $f_0(x) = x$. These flipping functions add typical steganographic noise (f_1) and noise that is atypical for steganography (f_{-1}). JRS, the derived version of the attack for JPEG, works in the same manner as RS except that it considers normalised coefficients d_k instead of pixels. The noise is measured in a group of coefficients $\mathbf{d}_k = (d_{nk}, d_{nk+1}, \ldots, d_{nk+n-1})$ with a default length $n = 4$ for $k \in \mathbb{N}$:

$$r(\mathbf{d}_k) = |d_{nk+1} - d_{nk}| + |d_{nk+2} - d_{nk+1}| + \cdots + |d_{nk+n-1} - d_{nk+n-2}|. \tag{12}$$

In a typical image, the noise will increase rather than decrease. We define sets of regular groups R and singular groups S for mask m before and after flipping the LSB of all normalised coefficients:

$$\begin{aligned}
R_m &= \{k | r(\mathbf{f}_m(\mathbf{d}_k)) > r(\mathbf{d}_k)\}, && (13)\\
S_m &= \{k | r(\mathbf{f}_m(\mathbf{d}_k)) < r(\mathbf{d}_k)\}, && (14)\\
R'_m &= \{k | r(\mathbf{f}_m(\mathbf{d}'_k)) > r(\mathbf{d}'_k)\}, && (15)\\
S'_m &= \{k | r(\mathbf{f}_m(\mathbf{d}'_k)) < r(\mathbf{d}'_k)\}, && (16)\\
\mathbf{d}'_k &= (f_1(d_{4k}), f_1(d_{4k+1}), f_1(d_{4k+2}), f_1(d_{4k+3})), && (17)\\
m &= (0, 1, 1, 0), && (18)\\
\mathbf{f}_m(\mathbf{d}_k) &= (d_{4k}, f_1(d_{4k+1}), f_1(d_{4k+2}), d_{4k+3}). && (19)
\end{aligned}$$

The cardinality of these sets is used to formulate the following quadratic equation:

$$2(q_0 + q'_0)x^2 + (q_1 - q'_1 - q'_0 - 3q_0)x + q_0 - q_1 = 0, \tag{20}$$

where $q_0 = |R_m| - |S_m|$, $q'_0 = |R'_m| - |S'_m|$, $q_1 = |R_{-m}| - |S_{-m}|$, and $q'_1 = |R'_{-m}| - |S'_{-m}|$. The estimated proportion p of the capacity used by randomised Jsteg is derived from the root x with the smaller absolute value:

$$p = \frac{x}{x - \frac{1}{2}}. \tag{21}$$

We studied all possible masks for group size $n = 2 \ldots 10$ in the spatial domain using a set of 500 images. Smaller masks lead to higher detection reliability and produced less failures.[1] Therefore two versions of the attack will be evaluated: One with the standard mask (cf. Eq. 18) and another one with the mask $(0, 1)$ and group size $n = 2$.

4.3 JPairs

The Pairs Analysis by Fridrich et al. [14] evaluates the number of homogeneous (00 or 11) and inhomogeneous pairs (01 or 10) in a binary sequence. After preprocessing (cf. Sect. 4.1) these binary sequences can be derived also from a sequence of DCT coefficients. Let $d_{\max}$ be the maximum even coefficient increased by one:

$$d_{\max} = \begin{cases} \max \mathbf{d} + 1 & \text{if } \max \mathbf{d} \text{ is even,} \\ \max \mathbf{d} & \text{otherwise.} \end{cases} \tag{22}$$

[1] Failures occur if Eq. 20 has no real solution.

The JPairs attack creates a binary sequence z for all pairs of normalised DCT values $(0,1),(2,3),\ldots,(d_{\max}-1,d_{\max})$. To construct these sequences, sequence $\mathbf{d}$ is scanned $\lceil d_{\max}/2\rceil$ times. In the first scan, 0 is appended to the initially empty sequence z whenever the value 0 is encountered in $\mathbf{d}$, and 1 if the value 1 is encountered. In the next scan, 0 is appended to z whenever the value 2 occurs in $\mathbf{d}$, and 1 if the value 3 is encountered. This is continued for all pairs. Likewise another sequence z' is created for the shifted pairs $(1,2),(3,4),\ldots,(d_{\max},0)$. Let ℓ be the length of $\mathbf{d}$. The proportion p of the capacity used by randomised Jsteg is estimated by

$$p = 1 - \sqrt{1 - \frac{q-q'}{\frac{\ell-1}{2} - \bar{q}}}\quad , \tag{23}$$

$$\text{where}\quad q = \sum_{i=1}^{\ell-1} |z_i - z_{i+1}|, \tag{24}$$

$$q' = \sum_{i=1}^{\ell-1} |z'_i - z'_{i+1}|, \quad \text{and} \tag{25}$$

$$\bar{q} = \sum_{j=1}^{\ell} \frac{1}{2^j} \sum_{i=1}^{\ell-j} |z'_i - z'_{i+j}|. \tag{26}$$

In our implementation, Eq. 26 is evaluated for $j = 1\ldots 24$ only, because there is no practical contribution for $j > 24$ due to the limited precision of the CPU.

4.4 JSPA

The sample pairs attack (SPA), developed by Dumitrescu et al. [12,13], is directly applicable to a sequence of normalised DCT coefficients. Let $d_{\max}$ be the maximum even coefficient, increased by one (cf. Eq. 22). The following two sets classify the sample pairs (u,v), which originate (non overlapping) from each two consecutive elements of the sequence of coefficients $\mathbf{d}$:

$$A = \{(u,v)|u \geq v,\ u \text{ even}\} \cup \{(u,v)|u < v,\ u \text{ odd}\}, \tag{27}$$

$$B = \{(u,v)|u < v,\ u \text{ even}\} \cup \{(u,v)|u \geq v,\ u \text{ odd}\}. \tag{28}$$

We define two histograms

$$h_0(|u_0 - v_0|) \text{ for } (u_0,v_0) \in A \quad \text{and} \tag{29}$$

$$h_1(|u_1 - v_1|) \text{ for } (u_1,v_1) \in B. \tag{30}$$

The estimated proportion p of the capacity used by randomised Jsteg is the larger of the two roots of the following quadratic equation:

$$0 = ap^2 + bp + c, \quad \text{where} \tag{31}$$

$$c = \sum_{i=0}^{j-1} h_0(2i) - h_1(2i+1), \tag{32}$$

$$b = h_0(0) + h_1(0) - \frac{h_0(2j+2) + h_1(2j+2)}{2} + \frac{c}{2}, \quad \text{and} \tag{33}$$

$$a = \sum_{i=0}^{1} \frac{h_0(i) + h_1(i)}{2} - \frac{h_0(2j+2+i) + h_1(2j+2+i)}{4}. \tag{34}$$

The upper limit j was determined experimentally. In the implementation used here, it was $j = 30$.

4.5 JWS

The WS attack by Fridrich and Goljan [11] is also directly applied to a sequence $\mathbf{d}$ of ℓ normalised JPEG coefficients. The proportion p of the capacity used by randomised Jsteg is estimated by

$$p = \frac{2\sum_{i=3}^{\ell-2} q_i}{\sum_{i=3}^{\ell-2} \frac{1}{1+v_i}} \quad \text{with} \tag{35}$$

$$v_i = \frac{1}{3}((d_{i-2} - \mu_i)^2 + (d_{i-1} - \mu_i)^2 + (d_{i+1} - \mu_i)^2 + (d_{i+2} - \mu_i)^2), \tag{36}$$

$$\mu_i = \frac{1}{4}(d_{i-2} + d_{i-1} + d_{i+1} + d_{i+2}), \quad \text{and} \tag{37}$$

$$q_i = \begin{cases} \frac{1}{1+v_i}(\mu_i - d_i) & \text{if } d_i \text{ is even,} \\ \frac{1}{1+v_i}(d_i - \mu_i) & \text{otherwise.} \end{cases} \tag{38}$$

The JWS attack estimates the original value of the coefficient q_i as the mean μ_i of a local environment of four values (cf. Eq. 37). Original values could be even more precisely estimated from steganographically unchanged neighbours. However, all those unchanged values have been removed in the sequence $\mathbf{d}$ of normalised JPEG coefficients.

4.6 Weighted Nonsteganographic Borders Attack (WB)

WB is a variant of the JWS attack, which estimates the original values directly from the sequence $\mathbf{x}$ of ℓ_0 JPEG coefficients (cf. Eq. 40). The values 0 and 1 are not used by randomised Jsteg. If they occur among the four elements forming the environment used for estimation, we can expect a higher precision of the estimated value. Consequently, this estimated value should have a larger weight in the attack. Apart from that, the mean distance of two consecutive coefficients is smaller in the sequence $\mathbf{x}$ since no elements are skipped. Consequently, the mean dependency between consecutive elements in $\mathbf{x}$, which is exploited by the attack, is larger than in $\mathbf{d}$. Note that the weight w_i (cf. Eq. 41) consists of three factors with the following features: The first factor excludes all $x_i \in \{0, 1\}$ ($w_i = 0$ for values that are not used by Jsteg), the second factor is the weight based on the variance as proposed for WS [11], and the third factor increases the weight if there are steganographically unused values in the estimation environment ($x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2} \in \{0, 1\}$). The proportion p of the capacity used by randomised Jsteg is estimated by

$$p = \frac{3 + 4\sum_{i=3}^{\ell_0-2} w_i(\mu_i - x_i)(-1)^{x_i}}{5\sum_{i=3}^{\ell_0-2} w_i} \quad \text{with} \tag{39}$$

$$\mu_i = \frac{1}{4}(x_{i-2} + x_{i-1} + x_{i+1} + x_{i+2}), \tag{40}$$

$$w_i = (1 - f(x_i)) \cdot \frac{1}{1 + v_i} \cdot \frac{1}{1 - \frac{1}{9}\sum_{j=-2}^{2} f(x_{i+j})(2 - x_{i+j})}, \tag{41}$$

$$f(x) = \begin{cases} 1 & \text{if } x \in \{0, 1\}, \\ 0 & \text{otherwise, and} \end{cases} \tag{42}$$

$$v_i = \frac{1}{3}((x_{i-2} - \mu_i)^2 + (x_{i-1} - \mu_i)^2 + (x_{i+1} - \mu_i)^2 + (x_{i+2} - \mu_i)^2). \tag{43}$$

5 Results and Discussion

5.1 Reliability

For meaningful experimental validation we use scanned images from a public database, which never have been lossy compressed before. We follow the tradition to apply the attacks to greyscale images. Of course the attacks can be applied to colour images as well by treating the colour components the same way as the brightness component. On the one hand it is not really advisable to steganographically use the colour components of a JPEG file since there are mutual dependencies between the components. On the other hand steganalysis could be improved to exploit these dependencies for more reliable detection of steganography in colour components. Steganographic messages with the same embedding rate are harder to detect in small images than in larger ones. Firstly, we choose an embedding rate with strongest possible distinction between the particular attacks in terms of reliability.

For the comparison in Fig. 3 we downloaded 2300 large TIFF images (2100 × 1500 pixels) from the NRCS database [23]. These images have been downsized to 840 × 600 pixels by pnmscale's default method, converted to greyscale, and JPEG compressed with quality $q = 0.8$. We applied randomised Jsteg to use 1 % of the image capacity. Figure 3 shows 39 bars that stand each for two readings of the area under the ROC curve (AUC). The solid lines give the result for all coefficients (AC and DC), while the dashed, which end with a small ring, indicate the case where DC coefficients—though used by randomised Jsteg—have been excluded (cf. coefficient sequence $\mathbf{d}_{\mathrm{AC}}$ in Sect. 4.1). The first three bars from the left represent the known Jsteg attacks with first order statistics (cf. Sect. 3). The

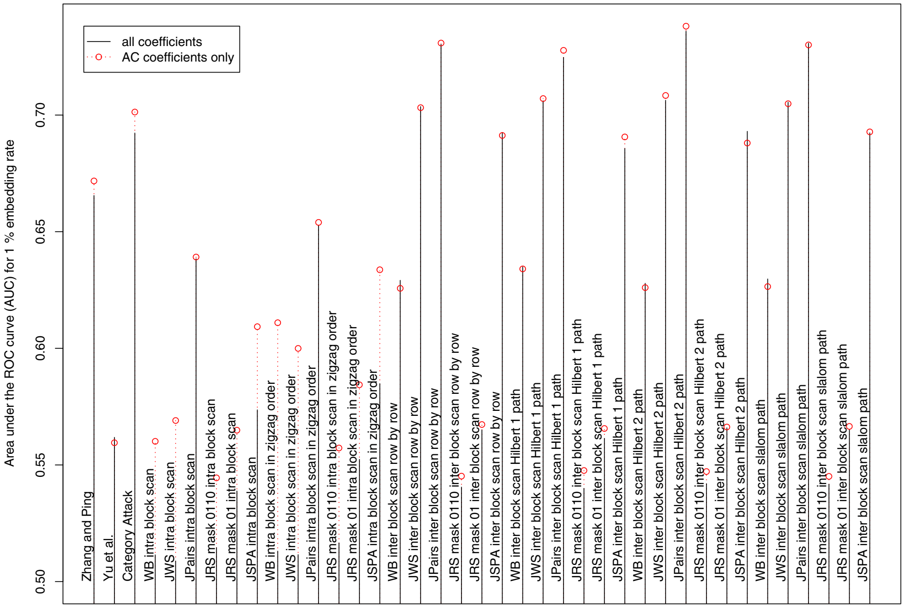

Fig. 3. Area under the ROC curve to compare the reliability of attacks for 3 known and 36 proposed attacks, each in two variants: with and without exclusion of DC coefficients

adjacent bars to the right give the two values for 36 combinations of 6 fundamental methods and 6 scanpaths.

Altogether the results of 78 different steganalytic methods are shown for a very low embedding rate, i.e., for adverse and meaningful, differentiating conditions (for maximum embedding rate all methods that are presented here separate perfectly). The longer the line in the diagram, the larger the area under the ROC curve (receiver operating characteristic). An area 1 means perfect separation. If the area is 0.5, the detector is not better than a random decision. In general, the area under the curve is more meaningful than a single point on the ROC curve (i.e., a pair of false positive rate and detection rate, which is valid for a particular threshold only). As mentioned earlier, the category attack has a small gain in reliability if only AC coefficients are considered. Apart from JPairs, this gain is more pronounced for the methods with intrablock scanpath. However, the intrablock methods perform worse than the leftmost histogram based methods (ZP and CA). Almost independent on the scanpath, the interblock methods score best. Merely the RS attack improves for the intrablock scan of AC coefficients. The JPairs attack is the obvious favourite for low embedding rates in small images. It is solely the reliability of JPairs that clearly tops the histogram based category attack by Lee et al. However, this result is only a snapshot for a particular pair of size and embedding rate. As we will see in Sect. 5.2, the ranking could severely change for smaller image sizes.

It is already known that recursive scanpaths (Hilbert 1, Hilbert 2) lead to noticeable improvements of the reliability for the Pairs Analysis in the spatial

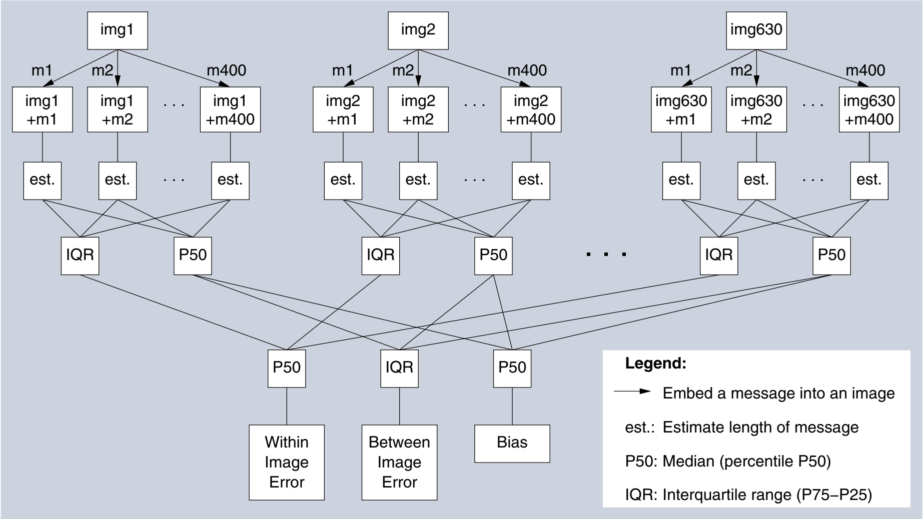

Fig. 4. Separation of within-image error, between-image error, and bias

domain [16]. One could argue that the advantage in the DC subband is equalised by a disadvantage in other subbands. Future work can clarify if different scan-paths in different subbands can still achieve an overall advantage.

5.2 Precision

All proposed methods estimate the length of the embedded message. In this section we will evaluate the precision of these DCT domain methods. Böhme and Ker evaluated the precision for spatial domain methods [24]. They separated different sources of errors by averaging. We will apply this principle also here, to evaluate three error-measures. A *cell* is a set of steganograms, produced by embedding different messages of the same length in one particular carrier medium. To separate the error sources we measure the statistical dispersion (interquartile range, IQR, difference between 0.75- and 0.25-quantile) and the central tendency (median, 0.5-quantile) of length estimations for each cell. From these two measures we derive three kinds of error (cf. Fig. 4).

The *within-image error* is the median of all cell IQRs. This kind of error is induced by the message and the secret key that is used for embedding. The distribution of the cell IQRs passed tests of normality. The *between-image error* is the IQR of all cell medians. This is an image-specific error. The cell medians are rather Student-t distributed with $\nu = 1 \ldots 10$ degrees of freedom for the proposed set of attacks. Finally, the *bias* is the median of the cell medians. This error can be pre-computed for a given source of images and a particular length estimating attack.

We downloaded another 630 large TIFF images (1500 × 2100 pixels) from the NRCS database [23]. These images have been downsized (using pnmscale) to five different sizes (600×840, 400×560, 200×280, 80×112, and 40×56 pixels), converted to greyscale, and JPEG compressed with quality $q = 0.8$. The medium sized images (200×280) have been compressed at seven different qualities ($q = 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99$). We applied randomised Jsteg at 8 different rates (1 %, 5 %, 10 %, 20 %, 40 %, 60 %, 80 %, and 100 % of the image capacity). We repeat the embedding with 400 random messages for over 600 images, at 11 combinations of size and quality, and at 8 embedding rates, detecting the result with 78 attacks. This results in about 1700 million length estimations.

Figures 5 and 6 present some views on these results. In abundance of results we highlight only the best representative for each fundamental estimation method. This selection is based on the between-image error at embedding rate 0.01 (600×840, JPEG quality 0.8, cf. Fig. 5(b)), which is related to the detection power, since this kind of error dominates at low embedding rates, while the within-image error is negligible (cf. Fig. 5(a)) and the bias is without influence to reliability. The representatives are the JPairs attack applied to AC coefficients only, scanned interblock along a Hilbert 1 path (cf. Sect. 4.3 and Fig. 2(b)), the JSPA attack applied to AC coefficients only, scanned interblock row by row (cf. Sect. 4.4 and Fig. 2(a)), the attack by Zhang and Ping [3], the category attack by Lee et al. [5], the JWS attack applied to AC coefficients only, scanned interblock row by row (cf. Sect. 4.5 and Fig. 2(a)), the WB attack applied to both AC and DC coefficients, scanned interblock along a slalom path (cf. Sect. 4.6 and Fig. 2(d)), the JRS attack with mask $(0, 1)$ applied to AC coefficients only, scanned intrablock in zigzag order (cf. Sect. 4.2 and Fig. 1(b)), and finally the attack by Yu et al. [4]. The pale curves in the diagrams show the remaining 70 attack variants. Apparently, the 8 representatives that have been selected based on their small between-image error have also a small within-image error. The within-image error grows with the embedding rate, but also for high JPEG qualities above 0.9 (cf. Fig. 5(c)). It decreases with increasing image size (not shown). This is also true for the between-image error (cf. Fig. 5(d)). While the between-image error of the JPairs attack is smallest for images larger than 200×280, there are more suitable candidates for thumbnail images (40×56): WB attack, category attack, and the attack by Yu et al. Interestingly, the JWS attack, which is closely related to the front runner WB, has the biggest problem with thumbnail images. One possible explanation could be the increased variance in downscaled images. The anonymous curves with the insignificantly lower error rates for small images belong to other variants of the WB attack.

Figure 5(e) shows the between-image error as a function of the JPEG quality. Apart from the category attack and the attack by Zhang and Ping, this kind of error is increased for qualities above 0.9.

The detector specific bias is shown in Fig. 5(f). We can distinguish several types of bias: (1) linearly decreasing (e.g., attack by Yu et al.), (2) arc with zero ends (e.g., JWS and category attack), (3) arc with zero left and negative right end (e.g., JRS, JPairs and JSPA), and (4) bias-free and bias-corrected types, which are close to zero over the whole range of embedding rates (e.g., the attack

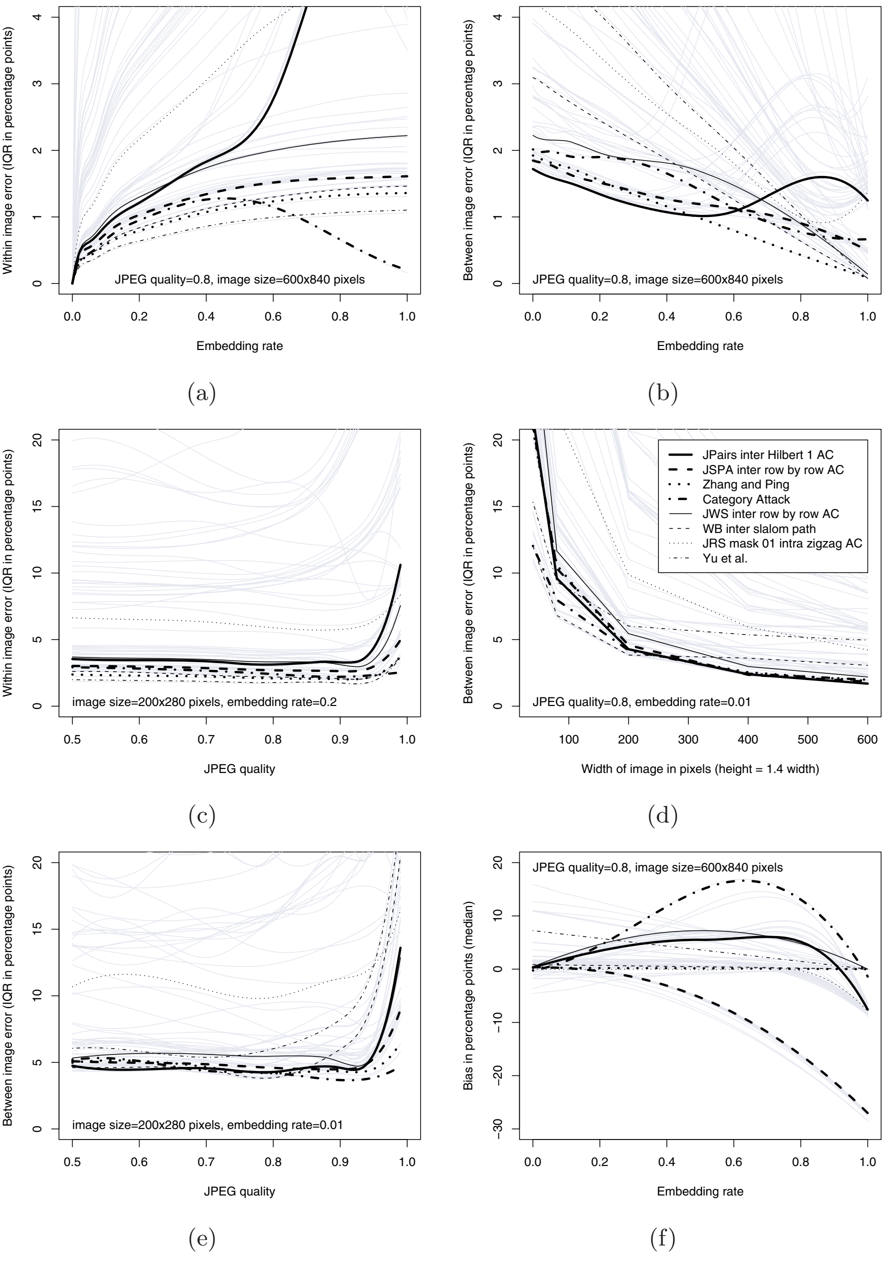

Fig. 5. Separated within- (a) and between-image error (b), within-image error as a function of quality (c), between-image error as a function of size (d) and quality (e), and detector-specific offset (f)

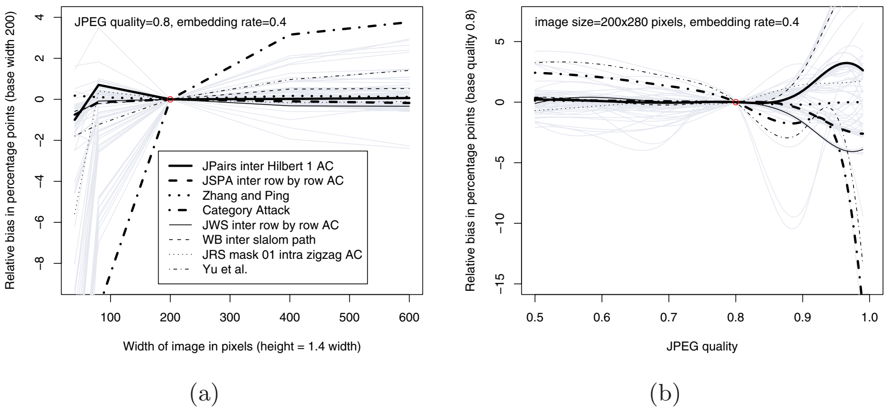

Fig. 6. Bias as a function of image size (a) and JPEG quality (b)

by Zhang and Ping as well as the WB attack). Note that the constants in Eq. 39 are adjusted based on a small, disjoint set of images in order to correct the bias that appears for the JWS attack. In Fig. 6 the bias of both, JWS and WB, appears invariant to changes of the image size. For JPEG qualities above 0.8, the relative bias curves of JWS and WB run in opposite directions. Despite its simple structure (or even because of it) the method of Zhang and Ping remains bias-free also for changes in size and JPEG quality.

6 Conclusion and Further Work

We developed a methodology to apply higher order steganalytic attacks from the spatial domain in the transformed domain. Based on 1700 million attacks, we evaluated the performance of the proposed attacks, and determined the most advisable schemes (WB for small images, JPairs for larger ones). Further work is needed regarding subband-specific scanpaths and an improved model for the estimation of original carrier values.

Acknowledgements

The work on this paper was supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under the research grant number FA8655-06-1-3046. The U. S. Government is authorised to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. Travel to the ACM Multimedia and Security Workshop was supported in part by the European Commission through the IST Programme under contract IST-2002-507932 ECRYPT.

References

1. Westfeld, A., Pfitzmann, A.: Attacks on steganographic systems. In: Pfitzmann, A. (ed.) IH 1999. LNCS, vol. 1768, pp. 61–76. Springer, Heidelberg (2000)
2. Upham, D.: Jsteg (1993), http://www.nic.funet.fi/pub/crypt/steganography/jpeg-jsteg-v4.diff.gz
3. Zhang, T., Ping, X.: A new approach to reliable detection of LSB steganography in natural images. Signal Processing 83, 2085–2093 (2003)
4. Yu, X., Wang, Y., Tan, T.: On estimation of secret message length in Jsteg-like steganography. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), pp. 673–676 (2004)
5. Lee, K., Westfeld, A., Lee, S.: Category Attack for LSB steganalysis of JPEG images. In: Shi, Y.Q., Jeon, B. (eds.) IWDW 2006. LNCS, vol. 4283, pp. 35–48. Springer, Heidelberg (2006)
6. Lee, K., Westfeld, A., Lee, S.: Generalised Category Attack—improving histogram-based attack on JPEG LSB embedding. In: Furon, T., Cayre, F., Doërr, G., Bas, P. (eds.) IH 2007. LNCS, vol. 4567, pp. 378–391. Springer, Heidelberg (2008)
7. Fridrich, J.: Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes. In: Fridrich, J. (ed.) IH 2004. LNCS, vol. 3200, pp. 67–81. Springer, Heidelberg (2004)
8. Shi, Y.Q., Chen, C., Chen, W.: A Markov process based approach to effective attacking JPEG steganography. In: Camenisch, J.L., Collberg, C.S., Johnson, N.F., Sallee, P. (eds.) IH 2006. LNCS, vol. 4437, pp. 249–264. Springer, Heidelberg (2007)
9. Pevný, T., Fridrich, J.: Merging Markov and DCT features for multi-class JPEG steganalysis. In: Delp III, E.J., Wong, P.W. (eds.) Security, Steganography and Watermarking of Multimedia Contents IX (Proc. of SPIE), San Jose, CA (2007)
10. Fridrich, J., Goljan, M., Du, R.: Detecting LSB steganography in color and grayscale images. IEEE Multimedia 8(4), 22–28 (2001)
11. Fridrich, J., Goljan, M.: On estimation of secret message length in LSB steganography in spatial domain. In: Delp III, E.J., Wong, P.W. (eds.) Security, Steganography and Watermarking of Multimedia Contents VI (Proc. of SPIE), San Jose, CA (2004)
12. Dumitrescu, S., Wu, X., Wang, Z.: Detection of LSB steganography via sample pair analysis. In: Petitcolas, F.A.P. (ed.) IH 2002. LNCS, vol. 2578, pp. 355–372. Springer, Heidelberg (2003)
13. Dumitrescu, S., Wu, X., Wang, Z.: Detection of LSB steganography via sample pair analysis. IEEE Trans. of Signal Processing 51, 1995–2007 (2003)
14. Fridrich, J., Goljan, M., Soukal, D.: Higher-order statistical steganalysis of palette images. In: Delp III, E.J., Wong, P.W. (eds.) Security, Steganography and Watermarking of Multimedia Contents V (Proc. of SPIE), San Jose, CA, pp. 178–190 (2003)
15. Latham, A.: JPhide and JPseek (1998), http://packetstormsecurity.org/crypt/stego/jphs/
16. Westfeld, A.: Space filling curves in steganalysis. In: Delp III, E.J., Wong, P.W. (eds.) Security, Steganography and Watermarking of Multimedia Contents VII (Proc. of SPIE), San Jose, CA, pp. 28–37 (2005)
17. Tu, C., Tran, T.D.: Context-based entropy coding of block transform coefficients for image compression. IEEE Transactions on Image Processing 11, 1271–1283 (2002)
18. Deever, A., Hemami, S.S.: What's your sign? Efficient sign coding for embedded wavelet image coding. In: Data Compression Conference, pp. 273–282 (2000)

19. Westfeld, A.: F5—a steganographic algorithm: High capacity despite better steganalysis. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, pp. 289–302. Springer, Heidelberg (2001)
20. Provos, N.: Outguess (2001), `http://www.outguess.org`
21. Sallee, P.: Model-based steganography. In: Kalker, T., Cox, I., Ro, Y.M. (eds.) IWDW 2003. LNCS, vol. 2939, pp. 154–167. Springer, Heidelberg (2004)
22. Fu, D., Shi, Y.Q., Zou, D., Xuan, G.: JPEG steganalysis using empirical transition matrix in block DCT domain. In: IEEE 8th Workshop on Multimedia Signal Processing, pp. 310–313 (2002)
23. NRCS: Photo gallery of the USDA Natural Resources Conservation Service (2006), `http://photogallery.nrcs.usda.gov/`
24. Böhme, R., Ker, A.D.: A two-factor error model for quantitative steganalysis. In: Delp III, E.J., Wong, P.W. (eds.) Security, Steganography and Watermarking of Multimedia Contents VIII (Proc. of SPIE), San Jose, CA, vol. 6072, pp. 59–74 (2006)