

DETECTION OF ± 1 LSB STEGANOGRAPHY BASED ON THE AMPLITUDE OF HISTOGRAM LOCAL EXTREMA

Giacomo Cancelli, Gwenaël Doërr, Ingemar J. Cox and Mauro Barni

ABSTRACT

Recently Zhang *et al* described an algorithm for the detection of ± 1 LSB steganography based on the statistics of the amplitudes of local extrema in the greylevel histogram. Experimental results demonstrated performance comparable or superior to other state-of-the-art algorithms. In this paper, we describe improvements to this algorithm to (i) reduce the noise associated with border effects in the histogram, and (ii) extend the analysis to amplitudes of local extrema in the 2D adjacency histogram.

Experimental results on a composite database of 7125 images, averaged over a 20-fold cross validation, with classification based on Fisher linear discriminant analysis, demonstrate that the improved algorithm exhibits significantly better performance. The experimental results are reported in the form of receiver operating characteristic (ROC) curves and summarized by computing the area under the ROC curve (AUC). The new algorithm, using 10 features derived from the 1D and 2D histograms, has an AUC value of 0.77 compared to 0.57 for the original algorithm. It also significantly outperforms other state-of-the-art steganalysers.

Index Terms— Steganography, steganalysis.

1. INTRODUCTION

Steganography seeks to provide a covert communication channel between two parties [1]. The secret message is hidden in a *cover Work*, which is often an innocuous image, video or text document. The combination of cover Work and secret message is referred to as the *stego Work*. The goal of all steganographic algorithms is to ensure *undetectability* [2], i.e. a third party, referred to as the Warden, should be unable to distinguish between a cover Work and a stego Work. On the other hand, the Warden's task is that of steganalysis, i.e. the discrimination of stego Works from cover works.

In this paper we are concerned with a common steganographic algorithm known as ± 1 embedding or LSB matching, in which the least significant bit of each sample is compared to its corresponding secret message bit, and the sample is randomly incremented or decremented if the LSB is not equal to the message bit. This is a variation on the simpler algorithm of LSB flipping, in which the least significant bit bitplane is replaced by the secret message bits.

Almost all steganalysis algorithms rely in detecting statistical anomalies between cover and stego Works. In the case of LSB flipping steganography, these statistical anomalies are obvious and thus easy to detect. To illustrate this, let us consider grayscale images with pixels values in the range $0 \dots 255$. When LSB flipping is used, an even-valued pixel will either retain its value or be incremented by one. However, it will never be decremented. The converse is true for odd-valued pixels. This asymmetry introduces a statistical anomaly into the intensity histogram: pairs of intensity values, specifically 0-1, 2-3 etc., will, on average, exhibit the same frequency if the image

is a stego Work. This can be exploited for steganalysis purposes, as described in [3, 4, 5].

LSB matching or ± 1 embedding addresses this issue. Rather than simply replacing the LSB with the desired message bit, the corresponding pixel value is randomly incremented or decremented whenever the LSB value needs to be changed. By so doing, the asymmetry present in LSB flipping is eliminated. However, other statistical anomalies are created that still permit discrimination between cover and stego Works, though these anomalies are more subtle and discrimination accuracy is significantly lower than for LSB flipping.

In this paper, we describe a new steganalysis algorithm that significantly improves upon previous results. It is based on work by Zhang *et al*, which is described in Section 2 and is based on the statistical properties of the amplitudes of local extrema (ALE). The performance of this algorithm has previously been compared with other state-of-the-art techniques [6], and shown to have comparable or superior performance. Extensions to this algorithm are then described in Section 3. Specifically, we first describe a modification to the algorithm that reduces noise associated with border effects, i.e. pixel values with intensities of either 0 or 255. Section 3.2 then describes extension of the amplitudes of local extrema to 2D adjacency histograms. These enhancements result in a collection of 10 features whose classification performances are evaluated in Section 4 through extensive experimental validation. The results clearly demonstrate significantly improved classification compared to the original steganalyser of Zhang *et al* [6], and to other state-of-the-art steganalysers [9, 10] Finally, Section 5 draws some conclusions before highlighting current challenges to be taken up in ± 1 steganalysis.

2. PREVIOUS WORK

The previously mentioned LSB matching algorithm can be formally described as follows:

$$p_s = \begin{cases} p_c + 1, & \text{if } b \neq \text{LSB}(p_c) \text{ and } (\kappa > 0 \text{ or } p_c = 0) \\ p_c - 1, & \text{if } b \neq \text{LSB}(p_c) \text{ and } (\kappa < 0 \text{ or } p_c = 255) \\ p_c, & \text{if } b = \text{LSB}(p_c) \end{cases} \quad (1)$$

where p_s (resp. p_c) denotes a pixel value in the stego image (resp. cover image), b is the message bit to be hidden, and κ is an i.i.d. random variable with uniform distribution on $\{-1, +1\}$ ¹. This process can be applied to all pixels in the image or only for a pseudo-randomly chosen portion, when the embedding rate, ρ , is less than one, i.e. the length of the hidden message is less than the number of pixels in the image.

In [7], the authors noted that LSB matching steganography induces a low-pass filtering of the intensity/colour histogram \mathbf{h}_1 of the image². Indeed, it is easy to show that, when looking at the intensity histogram, ± 1 steganography reduces to a filtering operation with the kernel:

G. Cancelli and M. Barni are with Università di Siena, Italy. G. Doërr and I. J. Cox are with University College London, UK. This work has been conducted during a 6 month research visit of G. Cancelli at UCL Adastral Park, which was financially supported by the Erasmus Programme from the European Commission.

¹Note that this strategy may affect bit-planes other than the LSB plane. For example, if the secret bit is a "0", and the original 8-bit pixel value is 01111111, then incrementing this value results in 10000000.

²In this article, all histograms will be considered to be implicitly normalized by the total number of samples.

$$\begin{bmatrix} \frac{\rho}{4} & 1 - \frac{\rho}{2} & \frac{\rho}{4} \end{bmatrix}$$

where ρ is the embedding rate. This implies that the histogram of a stego Work contains less high-frequency power than the histogram of the corresponding cover image.

Based on this, Zhang *et al* proposed to observe what happens in the vicinity of local extrema of the histogram [6]. Since LSB matching is equivalent to low pass filtering the intensity histogram, it will therefore reduce the amplitude of local extrema (ALE). This motivated the introduction of a new feature, which is basically the sum of the amplitudes of local extrema in the intensity histogram, as defined below:

$$A_1(\mathbf{h}_1) = \sum_{n \in \mathcal{E}_1} |2\mathbf{h}_1(k) - \mathbf{h}_1(k-1) - \mathbf{h}_1(k+1)| \quad (2)$$

where $\mathcal{E}_1 \subset [1, 254]$ is the set of local extrema in the histogram given by:

$$k \in \mathcal{E}_1 \Leftrightarrow (\mathbf{h}_1(k) - \mathbf{h}_1(k-1))(\mathbf{h}_1(k) - \mathbf{h}_1(k+1)) > 0 \quad (3)$$

Experimental results reported in [6] confirmed that this feature A_1 is statistically larger for original cover Works than for stego Works. Moreover, thresholding this feature results in a classifier with in much better classification performances compared with other state-of-the-art steganalysers, such as WAM [9] or HCF-COM [7, 10]

3. FEATURES BASED ON AMPLITUDES OF LOCAL EXTREMA (ALE)

In this section, we modify the design of the ALE steganalyser by incorporating additional complementary features, also based on the amplitude of some local extrema. Subsection 3.1 reduces border effects from the features used in [6], whereas subsection 3.2 introduces 2D adjacency histograms to compute additional features.

3.1. Removing Interferences at the Histogram Borders

Embedding based on Equation (1) introduces a small asymmetry: 0-valued pixels will *always* be changed to 1 if their LSB needs to be modified. Similarly, 255-valued pixels will *always* be changed to 254. This asymmetry in the histogram can cause interferences with the extracted feature, as demonstrated in Section 4.2. To avoid this, Equation (2) is modified, as follows, to remove this border effect:

$$A_1(\mathbf{h}_1) = \sum_{n \in \mathcal{E}_1} |2\mathbf{h}_1(k) - \mathbf{h}_1(k-1) - \mathbf{h}_1(k+1)| \quad (4)$$

where the set of local extrema \mathcal{E}_1 is now reduced to be within [3, 252]. In other words, the positions $\{1, 2, 253, 254\}$ are simply not considered as potential local extrema. Nevertheless, to account for this border effect, the following additional feature is defined:

$$d_1(\mathbf{h}_1) = \sum_{k \in \mathcal{E}_1^*} |2\mathbf{h}_1(k) - \mathbf{h}_1(k-1) - \mathbf{h}_1(k+1)| \quad (5)$$

where $\mathcal{E}_1^* \subset \{1, 2, 253, 254\}$ is a set of local extrema as defined by Equation (3).

3.2. Considering 2D Adjacency Histograms

Inspired by [10], the analysis of local extrema has been extended to 2D adjacency histograms, $h_2(k, l)$, which tabulate how often each pixel intensity is observed next to another in the horizontal direction:

$$\mathbf{h}_2(k, l) = \left| \{(i, j) \in \mathcal{I} \mid \mathbf{p}(i, j) = k, \mathbf{p}(i, j+1) = l\} \right| \quad (6)$$

where $\mathbf{p}(i, j)$ is the pixel value at location (i, j) in the input image, and \mathcal{I} is a bidimensional index which runs through all pixel locations in the image. Since adjacent pixels have, in general, close intensity values, this histogram is sparse off the diagonal. It should be noted that the histogram defined by Equation (6) can be slightly modified

1	$A_1(\mathbf{h}_1)$
2	$d_1(\mathbf{h}_1)$
3	$A_2(\mathbf{h}_2)$ (horizontal direction)
4	$A_2(\mathbf{h}_2)$ (vertical direction)
5	$A_2(\mathbf{h}_2)$ (main diagonal direction)
6	$A_2(\mathbf{h}_2)$ (minor diagonal direction)
7	$d_2(\mathbf{h}_2)$ (horizontal direction)
8	$d_2(\mathbf{h}_2)$ (vertical direction)
9	$d_2(\mathbf{h}_2)$ (main diagonal direction)
10	$d_2(\mathbf{h}_2)$ (minor diagonal direction)

Table 1. Table of ALE features

to obtain 3 other adjacency histograms for other directions (vertical, main diagonal, and minor diagonal).

LSB matching steganography also reduces to low-pass filtering the adjacency histogram with the following kernel:

$$\begin{bmatrix} \left(\frac{\rho}{4}\right)^2 & \frac{\rho}{4}\left(1 - \frac{\rho}{2}\right) & \left(\frac{\rho}{4}\right)^2 \\ \frac{\rho}{4}\left(1 - \frac{\rho}{2}\right) & \left(1 - \frac{\rho}{2}\right)^2 & \frac{\rho}{4}\left(1 - \frac{\rho}{2}\right) \\ \left(\frac{\rho}{4}\right)^2 & \frac{\rho}{4}\left(1 - \frac{\rho}{2}\right) & \left(\frac{\rho}{4}\right)^2 \end{bmatrix}$$

Consequently, it should also be possible to distinguish between cover and stego Works by examining local amplitude extrema in the 2D adjacency histogram. The set of local extrema in an adjacency histogram $\mathcal{E}_2 \subset [0, 255]^2$ is defined as:

$$\mathbf{p} = (k, l) \in \mathcal{E}_2 \Leftrightarrow \begin{cases} \exists \epsilon \in \{-1, 1\}, \forall \mathbf{n} \in \mathcal{N}_+ \\ \text{sign}(\mathbf{h}_2(\mathbf{p}) - \mathbf{h}_2(\mathbf{p} + \mathbf{n})) = \epsilon \end{cases} \quad (7)$$

where $\mathcal{N}_+ = \{(-1, 0), (1, 0), (0, -1), (0, 1)\}$ is used to define a cross-shaped neighborhood. However, many of these extrema have a small amplitude and are thus highly sensitive to changes to the cover Work. To achieve higher stability, this set is further reduced to:

$$\mathbf{p} = (k, l) \in \mathcal{E}'_2 \Leftrightarrow (k, l) \in \mathcal{E}_2 \text{ and } (l, k) \in \mathcal{E}_2 \quad (8)$$

In other words, only pairs of extrema symmetrical with respect to the main diagonal are retained. Empirical observations have revealed that such extrema have significantly higher amplitude and are thus more stable. The resulting feature is defined by,

$$A_2(\mathbf{h}_2) = \sum_{\mathbf{p} \in \mathcal{E}'_2} \left| 4\mathbf{h}_2(\mathbf{p}) - \sum_{\mathbf{n} \in \mathcal{N}_+} \mathbf{h}_2(\mathbf{p} + \mathbf{n}) \right| \quad (9)$$

which is the sum of the amplitude of extrema located at positions in \mathcal{E}'_2 .

In addition to this feature, empirical experiments have demonstrated that the sum of all the elements on the diagonal of a 2D adjacency histogram, defined as follows:

$$d_2(\mathbf{h}_2) = \sum_{k=0}^{255} \mathbf{h}_2(k, k) \quad (10)$$

could also be exploited to improve classification results. Indeed, ± 1 steganography decreases the value of this feature and its variations can be used in the decision process.

Altogether, this results in a collection of 10 alternative ALE features which are listed in Table 1.

4. EXPERIMENTS

In this Section we describe a number of experiments to investigate the impact of the various features on classification performance. Subsection 4.1 describes the experimental setup used in this study and Subsection 4.2 then presents the classification results achieved for different sets of ALE features.

4.1. Setup

The experiments were run on a database composed of images originating from three different sources. Specifically:

- 2,375 images from the NRCS Photo Gallery [13]. The photos are of natural scenery, e.g. landscape, cornfields, etc. There is no indication of how these photos were acquired. This database has been previously used in [10].
- 2,375 images captured using 24 different digital cameras (Canon, Kodak, Nikon, Olympus and Sony) previously used in [9]. They include photographs of natural landscapes, buildings and object details. All images have been stored in a raw format i.e. the images have *never* undergone lossy compression.
- 2,375 images from the Corel database [14]. They include images of natural landscapes, people, animals, instruments, buildings, artwork, etc. Although there is no indication of how these images have been acquired, they are very likely to have been scanned from a variety of photos and slides. This database has been previously used in [6].

This results in a composite database of 7125 images. Where necessary, all images have been converted to grayscale. Moreover, a central cropping operation of size 512×512 was applied to all images to obtain images of the same dimension across all three source databases. Cropping was preferred over resampling with interpolation, in order to avoid any interference with the source signal.

The motivation for using more than one source database is to account for the known variability in steganalizers' performances across different databases [11]. It is hoped that this set of databases will become a reference for subsequent works in steganalysis research³.

Using this composite database, several investigations have been conducted which all follow the procedure described below:

1. Select a set of ALE features to be used for classification;
2. Apply LSB embedding, with embedding rate ρ , to all images in the database \mathcal{D} to obtain a database of stego images \mathcal{D}^* ;
3. Separate both databases into a training set, $\{\mathcal{D}(\mathcal{I}), \mathcal{D}^*(\mathcal{I})\}$, and a test set, $\{\mathcal{D}(\bar{\mathcal{I}}), \mathcal{D}^*(\bar{\mathcal{I}})\}$, where \mathcal{I} is a subset of the image indexes and $\bar{\mathcal{I}}$ is its complement. The size of the training set was set to be equal to 20% of the database size;
4. Compute the selected features for all images in the training set and perform FLD analysis to obtain the trained projection vector \mathbf{u}^\dagger ;
5. Compute the selected features for all images in the test set, and project the feature vector onto \mathbf{u} ;
6. Compare the resulting scalar values to a threshold τ and record the probabilities of false positives and true positives for different values of the threshold in order to obtain the Receiver Operating Characteristic (ROC) curve of the system.

Steps 2 to 5 are repeated 20 times for cross-validation [8] and the ROC curves are vertically averaged to obtain the mean performance of the system. The overall performance of the steganalyser is then measured by computing the area under the ROC curve (AUC) [15]. An AUC value close to 1 indicates excellent discrimination, while a value close to 0.5 indicates poor discrimination.

4.2. Results

Although similar results were observed for various embedding rates, we will only report classification results for $\rho = 0.5$ in this paper due to space limitation.

³To encourage the use of this database, the authors have made it accessible on their website [12].

⁴It should be noted that, when a single feature is used, the FLD-based classification degenerates to thresholding the scalar feature.

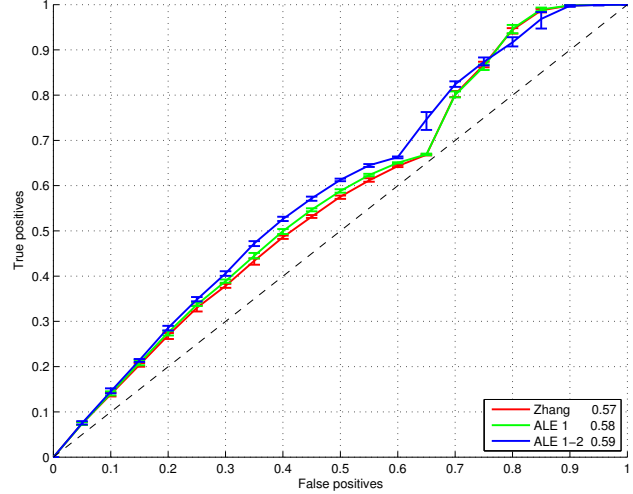


Fig. 1. Analysis of the impact of the border effect described in Subsection 3.1 on classification results.

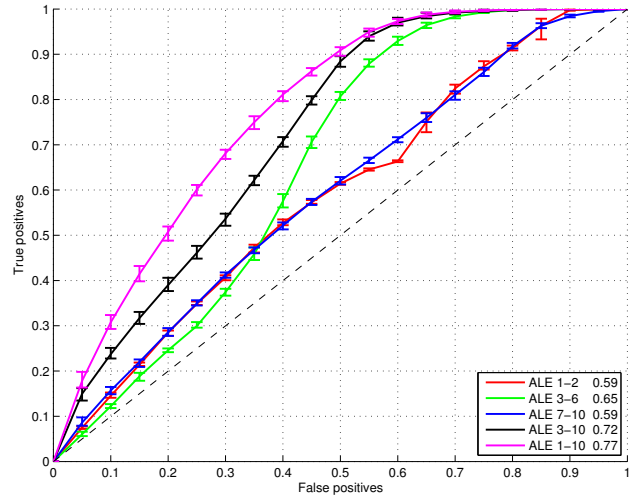


Fig. 2. Analysis of the impact of ALE features selection on classification results.

Figure 1 shows the improvements in classification resulting from the elimination of border effects. The original algorithm of Zhang *et al* is compared with using feature 1 of Table 1 (ALE 1), and using features 1 and 2 (ALE 1-2). The error bars on each plot indicate the minimum and maximum values observed during the 20 cross-validation runs. First of all, we note the unexpectedly poor performances of all three algorithms, i.e. the ROC curves are very close to the diagonal. This is due to the wide variety of images present in the composite database. Unfortunately, the performance variation across databases is beyond the scope of this study but the interested reader is directed to [11] for further information. Despite the poor performance of all three algorithms, the two algorithms based on new ALE features (ALE 1 and ALE 1-2) exhibit a slight improvement in classification performances. The system using the first two ALE features (ALE 1-2) achieves the highest performances based on area under the ROC curve (AUC) scores, with a score of 0.59, and is therefore used as a reference in the next experiment.

Figure 2 reports the classification performances achieved when using ALE features computed from the 2D adjacency histogram.

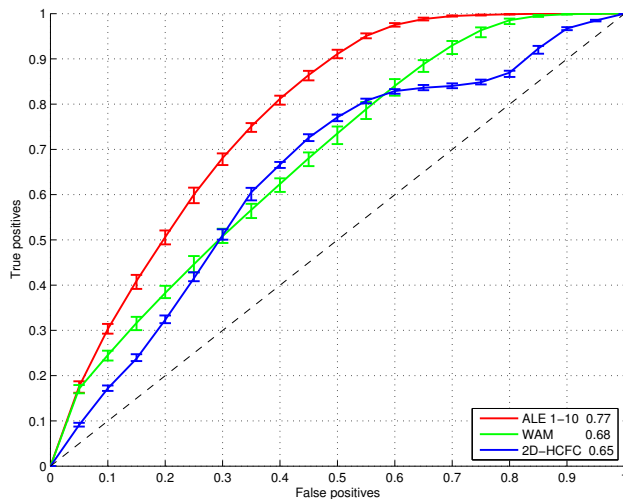


Fig. 3. Classification performances comparison between ALE, WAM [9] and 2D-HCFC [10].

Four sets of ALE features are investigated:

- ALE 3-6 i.e. the amplitude of the local extrema in the adjacency histograms,
- ALE 7-10 i.e. the amplitude of the diagonal in the adjacency histograms,
- ALE 3-10 i.e. all features from the adjacency histograms,
- ALE 1-10 i.e. all features from the intensity histogram and the adjacency histograms.

All 4 systems perform at least as well as the reference classification system obtained above (ALE 1-2). ALE 3-6 features perform significantly better than ALE 7-10 features. Nevertheless, when these two sets of features are combined (ALE 3-10), the resulting steganalyser outperforms both systems that rely on a single set of features computed from adjacency histograms. However, the best classification performance is achieved when all ALE features are combined (ALE 1-10). Compared to the original steganalyser [6], the area under the ROC curve (AUC) value increases from 0.57 to 0.77, which is a very significant improvement.

As a final sanity check, the final ALE steganalysis system has been compared to other state-of-the-art steganalysers, namely WAM [9] and 2D-HCFC [10]. The classification results are reported in Figure 3 and clearly demonstrated the superior performance of the proposed system. Nevertheless, comparing with Figure 1, it looks that claiming that Zhang’s steganalyser outperforms WAM and 2D-HCFC was a bit overstated. This reflects the high variability of steganalysis systems to the used database. The claim is true when considering only the Corel database used in [6] but no longer holds for the composite database which we are using in this paper and which we do think better reflects reality. This aspect is further analyzed in [11].

5. SUMMARY

The algorithm of Zhang *et al* was modified to deal with (i) border effects associated with the 1D intensity histogram, and (ii) extended to include statistics associated the amplitude of local extrema in the 2D adjacency histogram.

Experimental results demonstrated the impact of eliminating the border effects and very substantial improvements in classification when features derived from the 2D adjacency histogram were also included. Using the area under the ROC curve as a figure of merit, the

new ALE algorithm improved performance from 0.59 to 0.77. Moreover, the proposed steganalysis system proved to outperform other state-of-the-art steganalysers such as WAM [9] and 2D-HCFC [10].

While outside the scope of this paper, experimental results not reported here, revealed that the performance of many steganalysis algorithms [6] varied considerably depending on the source of imagery. Our composite database attempts to address this issue by including images from a variety of sources, including RAW never-compressed imagery and images derived from scanners, the latter exhibiting more high-frequency noise. This issue needs to be studied further and is examined in detail in [11].

6. ACKNOWLEDGEMENTS

The authors would like to thank J. Fridrich and M. Goljan from SUNY Binghamton (USA) for providing (i) the image databases they are using and (ii) their implementation of the WAM steganalyser.

7. REFERENCES

- [1] G. J. Simmons, “The prisoners’ problem and the subliminal channel,” in *Advances in Cryptology: Proceedings of CRYPTO’83*. 1984, pp. 51–67, Plenum Pub Corp.
- [2] I. Cachin, “An information-theoretic model for steganography,” in *Proceedings of the 2nd International Workshop on Information Hiding*, April 1998, vol. 1525 of *LNCS*, pp. 306–318.
- [3] J. Fridrich, M. Goljan, and D. Soukal, “Higher-order statistical steganalysis of palette images,” in *Security and Watermarking of Multimedia Contents V*, January 2003, vol. 5020 of *Proceedings of SPIE*, pp. 178–190.
- [4] O. Dabeer, K. Sullivan, U. Madhow, S. Chandrasekaran, and B. S. Manjunath, “Detection of hiding in the least significant bit,” *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 3046–3058, October 2004.
- [5] A. D. Ker, “A general framework for structural analysis of LSB replacement,” in *Proceedings of the 7th Information Hiding Workshop*, June 2005, vol. 3727 of *Lecture Notes in Computer Science*, pp. 296–311.
- [6] J. Zhang, I. J. Cox, and G. Doërr, “Steganalysis for LSB matching in images with high-frequency noise,” in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, October 2007.
- [7] J. Harmsen and W. Pearlman, “Steganalysis of additive noise modelable information hiding,” in *Security and Watermarking of Multimedia Contents V*, January 2003, vol. 5020 of *Proceedings of SPIE*, pp. 131–142.
- [8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, 2nd edition, 2001.
- [9] M. Goljan, J. Fridrich, and T. Holotyak, “New blind steganalysis and its implications,” in *Security, Steganography, and Watermarking of Multimedia Contents VIII*, January 2006, vol. 6072 of *Proceedings of SPIE*, pp. 607201–1.
- [10] A. D. Ker, “Steganalysis of LSB matching in grayscale images,” *IEEE Signal Processing Letters*, vol. 12, no. 6, pp. 441–444, June 2005.
- [11] G. Cancelli, G. Doërr, I. J. Cox, and M. Barni, “Comparing steganalysers: A case study with ± 1 steganography,” *IEEE Transactions on Information Forensics and Security*, submitted for publication.
- [12] G. Doërr, “Image database for steganalysis studies,” [online] <http://www.adastral.ucl.ac.uk/gwendoer/steganalysis>, (soon).
- [13] United States Department of Agriculture, “Natural resources conservation service photo gallery,” [online] <http://photogallery.nrcs.usda.gov>, 2002.
- [14] Corel Corporation, “Corel stock photo library 3,” Ontario, Canada.
- [15] T. Fawcett, “ROC graphs: Notes and practical considerations for researchers,” Tech. Rep., HP Laboratories, March 2004.