A NEW STEGANALYSIS METHOD FOR ±1EMBEDDING BASED ON AMPLITUDE OF LOCAL EXTREMA

Giacomo Cancelli

University of Siena

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 grey-level 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)extendthe analysisto amplitudesof localextremainthe2D adjacencyhistogram. The new algorithm, using 10 features derived from the 1D and 2D histograms, also significantly outperforms other state-of-the-art steganalyzers.

INTRODUCTION

Steganographyis the art of invisible communication. The term invisible is not linked to the meaning of the communication, as in cryptographyin which the goal is to secure communications from an eavesdropper, on the contrary it refers to hiding the existence of the communication channel itself. On the other side, we refer to steganalysis as the science which goal is to discover the presence of secret communication channels (secret messages) established by steganography.

In this paper we describe a new steganalysis algorithm which works on images in the pixel domain. To prove the performances of the proposed scheme, we use 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 of a sample is forced (or flipped) to the value of the corresponding secret message bit. This light variation is able to make the ±1embedding extremely less detectable than the classical LSB.

The ±1embedding algorithm can be formally described as follows:

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

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.

IMPROVING PREVIOUS WORK ON HISTOGRAM DOMAIN

In [2], the authors noted that ±1 embedding steganography induces a low-pass filtering of the intensity/colour histogram h1 of the image. Indeed, it is easy to show that, when looking at the intensity histogram, ±1steganography is equivalent to a filtering operation with the kernel:

$$\left| \frac{\rho}{4} \right| 1 - \frac{\rho}{2} \left| \frac{\rho}{4} \right|$$

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 idea, Zhang et al. [1] proposed to observe what happens in the surrounding of local extrema of the histogram [1]. Since±1 embedding is equivalent to lowpass filtering the intensity histogram, thenthefilteringoperationwillreducetheamplitudeoflocalextrema(ALE).Thismotivatedthe introduction ofanew feature,whichis basicallythesumofthe amplitudesoflocalextremainthe intensity histogram,as defined below:

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

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

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

Experimental results reported in [1] confirmed that the feature A_1 is statistically larger for original cover Works than for stego Works. Moreover, using this feature in conjunction with a classifier based on Fisher linear discriminant (FLD) [3] analysis, resulted in much better classification results compared with other state-of-the-art steganalyzers, such as WAM [4] or HCF-COM [2, 5].

Removing Interferences at the Histogram Borders

Embedding based on Equation (1) introduces a minor 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 in eq. (2). To avoid this problem, Equation (2) is modified, as follows:

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

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 not considered as potential local extrema. Nevertheless, to account the bound values of the histogram, the following additional feature is defined:

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

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

Considering 2D Adjacency Histograms

Inspired by [5], the analysis of local extrema has been extended to 2D adjacency histograms, h2 (k, l), which tabulates how often each pixel intensity is observed next to another in the horizontal direction:

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

where 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 to obtain 3 other adjacency histograms for other directions (vertical, main diagonal, and minor diagonal). For clarity we will use the apex *h, v, D, d,* respectively for horizontal, vertical, main diagonal, minor diagonal, to the adjacency function $h_2(k, l)$ in order to specify, if necessary, the kind of adjacency, otherwise $h_2(k, l)$ is referred to a generic kind of adjacency matrix. In particular, we define again the four kinds of adjacency matrix:

$$\mathbf{h}_2^h(k,l) = \left| \left\{ (i,j) \in \mathcal{I} \,|\, \mathbf{p}(i,j) = k,\ \mathbf{p}(i,j+1) = l \right\} \right| \tag{7}$$

$$\mathbf{h}_2^v(k,l) = \left| \left\{ (i,j) \in \mathcal{I} \,|\, \mathbf{p}(i,j) = k,\ \mathbf{p}(i+1,j) = l \right\} \right| \tag{8}$$

$$\mathbf{h}_2^D(k,l) = \left| \left\{ (i,j) \in \mathcal{I} \,|\, \mathbf{p}(i,j) = k,\ \mathbf{p}(i+1,j+1) = l \right\} \right| \tag{9}$$

$$\mathbf{h}_2^d(k,l) = \left| \left\{ (i,j) \in \mathcal{I} \,|\, \mathbf{p}(i,j) = k,\ \mathbf{p}(i+1,j-1) = l \right\} \right| \tag{10}$$

where 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.
Moreover, we can extend previous considerations about the ±1 embedding artefacts on the histogram domain by using the adjacency matrix. In this case, by using ±1 embedding with payload ρ, we obtain a 2-D low pass filtering with the following kernel:

$\left(\frac{\rho}{4}\right)^2$	$\frac{\rho}{4}\left(1-\frac{\rho}{2}\right)$	$\left(\frac{\rho}{4}\right)^2$
$\frac{\rho}{4}\left(1-\frac{\rho}{2}\right)$	$\left(1-\frac{\rho}{2}\right)^2$	$\frac{\rho}{4}\left(1-\frac{\rho}{2}\right)$
$\left(\frac{\rho}{4}\right)^2$	$\frac{\rho}{4}\left(1-\frac{\rho}{2}\right)$	$\left(\frac{\rho}{4}\right)^2$

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:

$$\mathbf{p} = (k,l) \in \mathcal{E}_2 \Leftrightarrow \begin{cases} \exists \epsilon \in \{-1, 1\},\ \forall \mathbf{n} \in \mathcal{N}_+ \\ \operatorname{sign}\left(\mathbf{h}_2(\mathbf{p}) - \mathbf{h}_2(\mathbf{p}+\mathbf{n})\right) = \epsilon \end{cases} \tag{11}$$

where $\mathcal{N}_+$ = {(−1, 0), (1, 0), (0, −1), (0, 1)} is used to define a cross-shaped neighborhood and $h_2(\cdot)$ is the generical adjacency matrix. However, many of these extrema have a small amplitude and are thus highly sensitive to changes of the cover Work. To achieve higher stability, this set is further reduced to:

$$\mathbf{p} = (k,l) \in \mathcal{E}_2' \Leftrightarrow (k,l) \in \mathcal{E}_2 \text{ and } (l,k) \in \mathcal{E}_2 \tag{12}$$

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 generical feature is defined by,

$$A_2(\mathbf{h}_2) = \sum_{\mathbf{p} \in \mathcal{E}_2'} \left| 4\mathbf{h}_2(\mathbf{p}) - \sum_{\mathbf{n} \in \mathcal{N}_+} \mathbf{h}_2(\mathbf{p}+\mathbf{n}) \right| \tag{13}$$

which is the sum of the amplitude of extrema located at positions in $\mathcal{E}_2'$.
In addition to eq. 13 feature, empirical experiments have demonstrated that the sum of all the elements on the diagonal of a 2D adjacency histogram, defined as follows:

$$d_2(\mathbf{h}_2) = \sum_{k=0}^{255} \mathbf{h}_2(k,k) \tag{14}$$

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, the above observations result in a collection of 10 features features which are listed in Table 1.

Table 1: Table of ALE features

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

Performances of ALE

In this Section we describe a number of experiments that we carried out to investigate the impact of the various features on classification performance.

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 [6].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 [5].

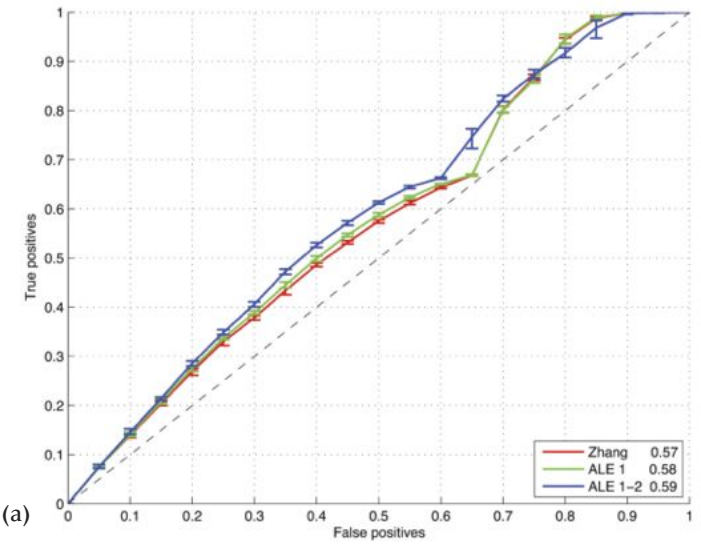
- 2,375 images captured using 24 different digital cameras (Canon, Kodak, Nikon, Olympus and Sony) previously used in [4]. They include photographs of natural landscapes, buildings and object details. All images have been stored in a raw format i.e. the images have never undergone lossy compression.

- 2,375 images from the Corel database [7]. 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 [1].

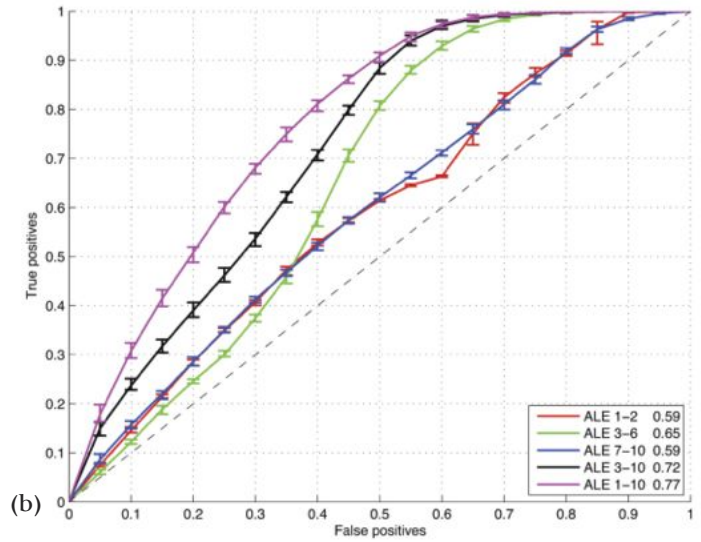
The above image sets result in a composite database of 7125 images. Where necessary, all images have been converted to grayscale. Moreover, a central cropping operation of size 512 × 512 was applied to all images to obtain images of the same dimension across all three source databases. Cropping was preferred 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 variability in steganalyzers' performances across different databases [8].

Given the composite database, the stego images are built by using ±1 embedding at 0.5 bpp of payload, thus obtaining the stego database. Then, for every image ALE features are extracted and we randomly separated the cover-features database D_{ALE} and stego features database D^*_{ALE} into a training set (20% of the database size), and a test set (the remaining 80% of the database) and we built a ROC curve by using Fisher Discriminant classifier on a training set and by projecting all the test feature vectors onto the trained projection vector u. To apply a cross validation on the obtained results, we repeat 20 times the above procedure with a different randomization of the train and test datasets. At the end we joined the 20 ROCs by the vertical averaging scheme and we show the average curve and the minimum and maximum bound of 20



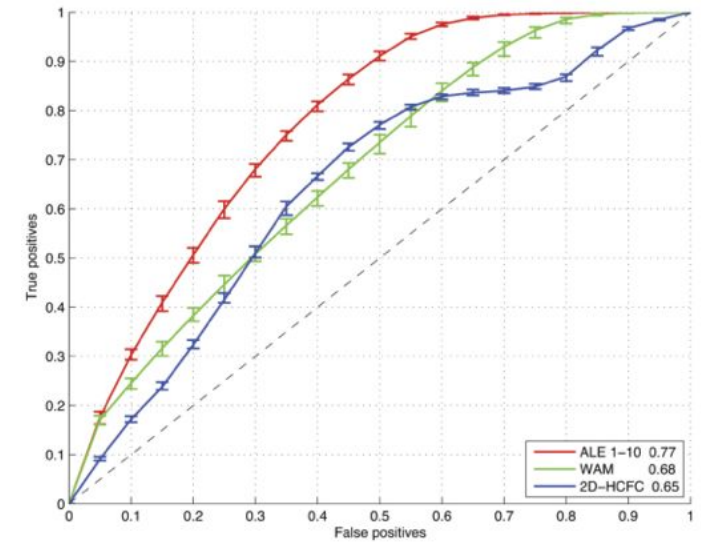
(a)



(b)

Fig.1 (above):
(a) Analysis of ()a the impact of the border effect and
(b) the impact of ALE features selection on classification results.

Fig.2: Classification performances comparison between ALE, WAM [4] and 2D-HCFC [5].



ROCs.

RESULTS

Since similar results were observed for various embedding rates, we only report classification results for ρ=0.5.

Figure 1(a) shows the improvements in classification resulting from elimination of border effects. The original algorithm of Zhang et al. is compared with a system based on feature 1 of Table 1 (ALE 1), and 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 of composite database.

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), with a score of 0.59, and is therefore used as a reference in the next experiment.

Figure 1(b) reports the classification performances achieved when using ALE features computed from the 2D adjacency histogram. Four sets of ALE features are investigated:

- ALE 3-6 i.e. the amplitude of the local extrema in the adjacency histograms,
- ALE7-10 i.e. the amplitude of the diagonal in the adjacency histograms,
- ALE3-10 i.e. all features from the adjacency histograms,
- ALE1-10 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 considered 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 steganalyzer outperforms the 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 steganalyzer [1], the area under the ROC curve (AUC) value increases from 0.57 to 0.77, which is a significant improvement.

As a final sanity check, the final ALE steganalysis system has been compared to other state-of-the art steganalyzers, namely WAM [4] and 2D-HCFC [5]. The classification results are reported in Figure 2 and clearly demonstrated the superior performance of the proposed system. Nevertheless, comparing with Figure 1(a), it looks that claiming that Zhang's steganalyzer outperforms WAM and 2D-HCFC was a bit overstated. This reflects the high variability of steganalysis systems to the used database.

CONCLUSION

In this work 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 steganalyzers such as WAM [4] and 2D-HCFC [5].

Even though ALE seems to be have very well, anappropriate comparison procedures hould be designed to compare ALE behavior against state-of-art classifiers. Specifically, we should investigate as future work how ALE performance vary by changing the experimental conditions by changing both the image databaseand the payload. Due to the importance of experimental settings and comparison with other steganalyzers likeWAM and 2D-HCFC, we willinvestigate the ALE performance and comparisonin the next chapter.

REFERENCES

[1] J. Zhang, I. J. Cox, and G. Doërr,"Steganalysis for LSB Matching in images with high-frequencynoise," IEEE 9thWorkshop on Multimedia SignalProcessing (MMSP), 2007, pp. 385–388, 2007.

[2] J. Harmsen, "Steganalysis of additive noise modelable information hiding," Ph.D. dissertation, Ph.d. thesis at Rensselaer Polytechnic Institute, 2003.

[3] R. Duda,P. Hart, andD. Stork, Pattern classification. Wiley-Interscience, 2000.

[4] M. Goljan, J. Fridrich, andT. Holotyak, "New blind steganalysis and its implications," Proceedings of SPIE, vol. 6072, pp. 1–13, 2006.

[5] A. D. Ker, "Steganalysis of LSB matching in grayscale images," IEEE Signal Processing Letters, vol. 12, no. 6, pp. 441–444, 2005.

[6] U. S. D. of Agriculture, "Natural Resources Conservation Service photo gallery," 2002. [Online]. Available: http://photogallery.nrcs.usda.gov

[7] C. Corporation, "Corel Stock Photo Library 3," Ontario, Canada.

[8] G. Cancelli, G. Doërr, M. Barni, and I. J. Cox, "A comparative study of ±1steganalyzers,"Multimedia SignalProcessing, 2008 IEEE 10thWorkshop on, pp. 791–796, Oct. 2008.