A Novel Active Warden Technique for Image Steganography

Igor Bisio, Fabio Lavagetto, Giulio Luzzati and Andrea Sciarrone
University of Genoa - DITEN Department
Via all’Opera Pia 13
16145, Genova Italy
Email: {igor.bisio@unige.it, giulio.luzzati@edu.unige.it, fabio.lavagetto@unige.it, andrea.sciarrone@unige.it}

Abstract—This paper focuses on countering data hiding in images. Specifically, we refer to the active warden problem, a process which aims at disrupting the covert communications that are possibly taking place within host media. Although this has been extensively studied in the literature, typical state of the art approaches treat the hidden message as a noise-like component, proposing de-noising techniques, lossy compression or noise addition as a countermeasure. In this paper we present novel approach called Lesser Components Distortion (LCD) that aims at disrupting the covert communication while minimizing distortion on the host medium. Although stemming from insights gained from notable works in the spread spectrum steganography field, we prove through extensive testing that it shows remarkable effectiveness against steganographic techniques in general, compared to common attacks.

I. INTRODUCTION

Multimedia contents are subject to human interpretation, and are thus processed by the Human Visual System (HVS), which is extremely efficient in filtering and extrapolating the semantic elements out of perceptions. This makes them particularly robust to minor distortions and, for this reason they represent a good choice for Information Hiding (IH) techniques, which in fact exploit small distortions in the host media to insert hidden information.

Two are the main IH applications: steganography and watermarking. The former is primarily focused on the communication aspect, and thus on i) maximizing the probability of correctly receiving the embedded message while ii) minimizing the distortion introduced in the carrier signal, in order to avoid arousing suspicions.

Watermarking shares many aspects with steganography, but on the other hand it is less concerned with undetectability and maximal information transfer while more on the robustness of the embedded message [1], [2]: typically, watermarking techniques are employed in copyright protection and traitor tracing problems.

In general, in addition to the entities communicating through a covert channel, a third element called warden is considered. As described in the prisoner’s problem [3], the warden, is the entity that controls the communication medium, and it can be either i) passive or ii) active.

While the former warden acts like a completely transparent channel, i.e., it can only analyse the content sent through without modifying it, the active warden is a model for a scenario in which it is allowed to act upon the transmitted medium. These two models entail two very different classes of countermeasures to covert communications: passive warden models, differently from eavesdropping problems [5], [6], have no prior on whether actual communication is taking place, and employ solutions from the widely discussed topic of steganalysis [7], which studies the statistical properties of images with the aim to detect the presence of hidden communications. Although widely investigated, steganalysis techniques typically perform poorly outside of laboratory conditions [8]. Conversely, active warden models tackle the problem from a different perspective, aiming to trade small distortion on possibly genuine multimedia content for the assured disruption of any subliminal communication.

The problem of the active warden for images has been studied extensively in the literature. Referring to the scenario depicted in Fig. 1, Alice (A) wants to secretly communicate with Bob (B) through a cover image, but Alice’s communications have to transit through Wendy (W). Wendy is the entity who relays communication from Alice to Bob and has no a priori knowledge on whether the image contains secret messages, but it possesses the ability to actively modify the image prior to delivering it to Bob.

Ideally, the goal of Wendy is to i) be able to deliver an image which retains the semantics of the image and its visual quality, and at the same time ii) to disrupt hidden communications.

Practically, when Alice hides her secret message $M$ in an image that goes through Wendy, Bob subsequently recovers a possibly different message $M'$. Wendy’s goal is to minimize the mutual information $I(M; M')$, to the point that they are uncorrelated.

This problem has been studied within the two main branches of data hiding, watermarking and steganography. In the first case it is more notably known as watermark “removal” or
“attack”. Many works proposed a set of approaches or attacks [2, 1, 9, 10, 11], the majority of which considers simple image processing techniques, such as low-pass filtering, Gaussian noise addition [12] (in general, noising and denoising), and lossy compression. For what concerns steganography, the “active warden” topic has received somewhat less attention. Specifically, many works highlighted the employment of recompression [13], [14], [15] as a countermeasure to steganography, especially in Online Social Networks (OSNs) [16], [17], [18]. Some have proposed the employment of a quantizer [19]. The general problem of active warden has been addressed explicitly in [14], although only trivial attacks are evaluated, such as bit deletion, JPEG compression and denoising techniques, which typically can be loosely grouped under the low-pass filtering category. Such approaches that treat the hidden message like noise are regarded as not very effective [20].

[15] first introduces the term scrub, referring to techniques which are aimed at disrupting existing covert communications; specifically, it proposes a proof of concept framework for systematic scrubbing of stego images, but it does not introduce any novel element for the scrubbing process.

In this paper we introduce and describe a novel scrubbing technique, called Lesser Component Distortion (LCD), and we compare it to other scrubbing techniques commonly found in the literature. To the best of the authors’ knowledge, LCD is not related to any current state of the art approach. It is a method which targets the lesser components (i.e., the ones that minimize the host’s statistical variance along their direction) of the image, differently from other state of the art approaches, that in general

- target least perceptible components (compression methods);
- target high frequency components (e.g., low-pass filtering, de-noising);
- cause indiscriminate degradation (AWGN);

The rest of this article is structured as follows: in Section II we describe the data hiding techniques against which we test the scrubbing methods, in Section III we specify the rationale that drove our work, we describe LCD scrubbing and our implementation details. In Section IV we evaluate our technique’s performance, comparing it to three simple scrubbing methods commonly employed in the literature. Finally, conclusions are drawn.

II. THE CONSIDERED EMBEDDING TECHNIQUES

In this paper we consider three popular data hiding techniques, which we will try to attack using the methods described in Section III and IV.

1) Least Significant Bit (LSB): LSB embedding is a simple method for image steganography. It exploits the fact that the level of precision in many image formats is much greater than the one perceivable by the HVS. As a consequence, slight variations in the luminance (or color) values of the pixels act as a minor noise component which is too weak for the HVS to perceive. Practically, LSB steganography is carried out by substitution of one or more of the least significant bit planes of the image. In this work we consider 1-LSB plane substitution, which allows the insertion of one bit of information for each image pixel.

2) Scalar Costa Scheme (SCS): SCS is a popular quantization-based steganography method originally proposed in [21], and it is based on a pixel-by-pixel modification designed to insert one embedding symbol per pixel. In its most simple form, the general SCS embedding is done as in (1)

\[ y_i = x_i + \alpha (Q_\Delta \{x_i - m_i \Delta\} - (x_i - m_i \Delta)) \]  

where \( y_i \) and \( x_i \) respectively denote the \( i \)-th pixel of the stego and original image, \( \alpha \) is a scalar modulating the embedding strength, \( Q_\Delta \{ \cdot \} \) is a scalar uniform quantizer with stepsize \( \Delta \) and \( m_i \) is the \( i \)-th message symbol.

3) Spread Spectrum (SS): While both LSB and SCS insert one message bit within a single pixel, SS techniques spread the energy of a symbol of the secret message over a wide range of pixels.

In this paper we consider additive spread spectrum embedding, which generally can be modeled as

\[ Y = Ams + X \]  

where \( Y \) is the stego image, obtained by summing to the original image \( X \) a carrier \( s \) which is modulated by a message \( m \) and an amplitude scalar \( A \). Detection in additive spread spectrum steganography is done by matched filtering.

The shape of \( s \) can be critical for the embedding performance. Under the hypothesis of Gaussian distribution of the host pixel values, it is possible to find the optimal carrier given the specific image statistics. In more details, considering blockwise processing (likewise of how JPEG 8 × 8 block processing works), it is proven [22] that optimal carrier design is achieved using the lowest eigenvalue eigenvector of the autocorrelation matrix of the DCT components of the blocks. A filter which is matched to the lowest component will be mostly orthogonal (on average) to the components of a given image, so that communication carried by such components will suffer minimally from interference by the image, thus requiring minimal amount of energy to meet the desired Bit Error Rate (BER) on the embedded message.

III. THE PROPOSED LCD TECHNIQUE

Optimal signature design in additive spread spectrum embedding indeed allows major gains in terms of covert channel capacity; however, as a consequence of the design method, given a specific stego-image it is easy to locate where the secret message energy is distributed.

This technically equals to assume that the \( Ams \) component in Equation (2) must not be strong enough to significantly change the image statistics, thus making the secret message unnoticeable: in other words, if Alice hides a message, it is in her best interest to limit the embedding strength to a bare minimum.
This insight suggests that a properly designed attack can "surgically" remove the energy of covert communication embedded through OSS. This can be especially effective if we consider that the secret message energy is located in the lesser components of the image, thus causing minimal image detriment.

In general, while this may seem an ad-hoc solution, we argue that it proves effective even when targeting different embedding techniques. This is in accordance to what other state of the art scrubbing methods aim to, which essentially is to induce the greatest possible pixel values change without affecting the semantic information.

A. LCD implementation

Given an image, we perform $8 \times 8$ blockwise Discrete Cosine Transform (DCT) and reshape all DCT coefficients excluding the zero frequency component, thus obtaining $X$ in accordance with (2), which is a $63 \times N$ matrix, where each row corresponds to a single block.

The $63 \times 63$ autocorrelation matrix is computed with

$$R_x = \frac{XX^T}{N} \quad (3)$$

Its eigenvectors form a $63 \times 63$ basis $S$ where each column is a single eigenvector $s_j = [s_{j1}, \ldots, s_{jk}, \ldots, s_{j63}]$, corresponding to the $\lambda_j$ eigenvalue, with $j = 1, \ldots, 63$.

Once sorted the eigenvectors according to their eigenvalues (i.e., $\lambda_1$ is the smallest), we use the first $V$ eigenvectors $s_j$, $j = 1, \ldots, V$ (with $V < 63$) to build the $63 \times 1$ "map" vector $p$, whose $k$-th component is

$$p_k = \frac{\sum_{j=1}^{V} |s_{jk}|}{V}, \quad k = 1, \ldots, 63 \quad (4)$$

which locates the frequency coefficient related to the lesser components of the image. Denoting $Z$ as the LCD scrubbed image and the vector $z_j$ as the $j$-th row, the $z_{jk}$-th element is computed as

$$z_{jk} = \begin{cases} n_k & \text{where } p_k = \theta \\ y_{jk} & \text{otherwise} \end{cases} \quad (5)$$

Where $y_{jk}$ is an element of $y_j$, a row of $Y$, and $\theta$ is a threshold over which the corresponding $k$-th DCT component is replaced with a random realization $n_k$ of a random variable whose Probability Density Function (PDF) is $\mathcal{N}(0, \sigma_n)$, which is Gaussian with mean equal to 0 and standard deviation $\sigma_n$. This serves the purpose of distorting the frequencies which are maximally sensitive for optimal SS steganography. $\theta$ is the parameter we vary in the experiments presented in the next Section, and it controls the strength of the LCD scrub. The results in this paper are obtained by setting $\sigma = 5$ and $V = 20$.

Henceforth, we denote the LCD scrub operation on the $Y$ image as $Z_{LCD} = \tilde{j}(Y)_{\sigma}$

IV. RESULTS

We used all the 512 $\times$ 512 images of the miscellaneous USC-SIPI database (26 images), and compared our proposed LCD method with three image scrubbing approaches considered in the literature, namely

- Low Pass Filtering (LPF) implemented by convolution of the image with a gaussian kernel $K_{LPF}$ with variable variance $\sigma$, thus $Z_{LPF} = Y \ast K$;
- JPEG compression, varying the coder quality parameter;
- the injection of AWGN with variance $\sigma_{AWGN}$, thus $Z_{AWGN} = Y + n$, with $n$ being a random variable drawn from a Gaussian PDF $\mathcal{N}(0, \sigma_{AWGN})$

Such scrubbing methods are employed to disrupt the hidden message embedded using the techniques described in Section II. To sum up, for each image, we perform

1) embedding of a random secret binary message $m$ into the image $X$ using one of $\{SCS, SS, OSS\}$ thus producing $Y$
2) scrubbing of the obtained stego image $Y$ with variable scrubbing strength parameter thus obtaining $Z$; the actual parameter depends on the specific scrub method (i.e., LCD, LPF, JPEG, AWGN)
3) BER measuring between $m$ and the message obtained from the scrubbed image, $m'$

In order to assess the degradation introduced by the scrubbing process we consider the CPA metric [23], which, unlike the well known Peak Signal to Noise Ratio (PSNR), represents a measure of distortion (two identical images would score CPA = 0), and it is particularly effective in predicting the Mean Opinion Score (MOS), and thus the actual distortion perceived by the human visual system [24].

As mentioned in Section I, the goal of an active warden is to minimize the mutual information between Alice’s message $M$ and what Bob receives $M'$. Ideally, Wendy aims to make $I(M;M') = 0$, meaning that $M$ and $M'$ are completely uncorrelated. This implies a minimization of the covert channel capacity $C$, which is given by

$$C = \max \{ I(M;M') \} \quad (6)$$

Assuming binary coding, the BER will approach 0.5 as $C \to 0$. We thus consider the BER between the $M$ and $M'$ as a measure of how a particular scrub method fares in terms of effectiveness, with the aim to bring it as close to 0.5 as possible.

In Fig. 2 we report the results against LSB embedding. LSB allows to insert one symbol per pixel; considering binary coding the payload is $262144$ bits per image. The trends of the four considered scrubbing techniques are all very similar: the BER reaches the maximal values of 0.5 even at very low distortion. This is due to the fact that LSB embedding the message is contained only in the least significant bit plane which is extremely fragile. As a consequence, even slight modifications in the image are likely to produce variations in the LS bitplane. In general, if the LSB steganography
is considered, almost every scrubbing technique is able to efficiently destroy the embedded message.

Fig. 2: Smart Scrub technique versus LPF, JPEG and AWGN injection with the LSB embedding method.

With SCS embedding (Fig. 3) the scrubbing methods now exhibit different behaviour. As with LSB embedding, the payload is 262144 bits per image. The LCD curve lies above all other ones, reaching higher induced BERs at lower distortion levels with respect to all the other approaches. Moreover, we can note that in this case JPEG compression and lowpass approaches substantially give the same results, and this is likely due to the SCS embedding working in a pixel-by-pixel manner, which means that the secret message energy will reside, for a significant portion, in the high frequency part of the spectrum, which are either cut off by lowpass filtering or heavily distorted by JPEG compression. The AWGN scrub is an attack which is essentially spectrally uniform and thus fails to privilege important components of the image, leading to higher distortions per given BER target.

Fig. 3: Smart Scrub technique versus LPF, JPEG and AWGN injection with the SCS embedding method.

Finally, in Fig. 4 we tested LCD against SS steganography. Specifically we consider the optimal carrier SS embedding as introduced in [22]. Differently from LSB and SCS, SS embedding spread one information symbol over a wide range of frequencies. Even though SS allows adding more than one carrier thus increasing the total payload, for sake of generality we consider only a single carrier, for a total of 4096 bits per image using $8 \times 8$ pixels blocks. This grants SS steganography higher robustness to AWGN, although resulting in a lower payload.

In this case LPF and JPEG scrubbing perform differently, with JPEG showing a small advantage over LPF: this happens because LPF preserves the lower frequencies components, while JPEG distorts a less strictly defined region of frequency coefficients with quantization tables. Interestingly, our LCD approach exhibits a clear advantage, and the curves suggest that $\sim 0.5$ BER can be obtained on average by allowing just over 2.75 CPA, which constitutes a virtually unnoticeable degradation on the image.

The data gathered from the three experiment is summed up for convenience in TABLE I, and some example images are reported in Fig. 5 to give an idea of the distortions introduced by the scrub process.

As a concluding remark, we highlight a very interesting feature of LCD scrubbing: differently from other common techniques, because of how it is designed, it targets a set of frequencies which is in general not exclusively high-pass. This important point suggests that the employment of LCD as complementary scrub mechanism in combination with other techniques could prove extremely effective, considering for example that JPEG compression represents a plausible final step in any practical active warden implementation.

Fig. 4: Smart Scrub technique versus LPF, JPEG and AWGN injection with the SS (implemented as in [22]) embedding method.

V. Conclusions

In this paper we presented a novel approach to image scrubbing, a process which aims at disrupting the covert communications that possibly take place within images. Our approach stems from insights taken from notable works in the
The original image with no message embedded (CPA=0).

The JPEG-scrubbed image

BER = 0.46, CPA~2.54.  The AWGN-scrubbed image

BER = 0.47, CPA~7.3.

The LPF-scrubbed image

BER = 0.45, CPA~3.0.  The LCD scrubbed image

BER = 0.49, CPA~2.5.

Fig. 5: Sample images from the scrubbing experiments.

TABLE I: A comparison of the measured BER achieved by the different scrubbing techniques.

<table>
<thead>
<tr>
<th>CPA</th>
<th>MEASURED BER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LSB</td>
</tr>
<tr>
<td>2.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>3.5</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
</tbody>
</table>

field of spread spectrum steganography, but we proved through extensive testing against common scrubbing methods, that it shows interesting performance against data hiding techniques in general, given the small price it requires in terms of introduced distortion, it may complement as a very useful element in a wider chain of an active warden approach which encompasses more than one technique.

REFERENCES
