StegoWall: Blind statistical detection of hidden data

Sviatoslav Voloshynovskiy[*a], Alexander Herrigel[b], Yuriy Rytsar[b] and Thierry Pun[a]

[a]University of Geneva, 24 rue Général-Dufour, CH-1211 Geneva 4, Switzerland
[b]Digital Copyright Technologies, Stauffacher-Strasse 149, CH-8004, Zurich, Switzerland

ABSTRACT

Novel functional possibilities, provided by recent data hiding technologies, carry out the danger of uncontrolled (unauthorized) and unlimited information exchange that might be used by people with unfriendly interests. The multimedia industry as well as the research community recognize the urgent necessity for network security and copyright protection, or rather the lack of adequate law for digital multimedia protection. This paper advocates the need for detecting hidden data in digital and analog media as well as in electronic transmissions, and for attempting to identify the underlying hidden data. Solving this problem calls for the development of an architecture for blind stochastic hidden data detection in order to prevent unauthorized data exchange. The proposed architecture is called StegoWall; its key aspects are the solid investigation, the deep understanding, and the prediction of possible tendencies in the development of advanced data hiding technologies. The basic idea of our complex approach is to exploit all information about hidden data statistics to perform its detection based on a stochastic framework. The StegoWall system will be used for four main applications: robust watermarking, secret communications, integrity control and tamper proofing, and Internet/Network security.

Keywords: watermarking, stochastic visibility, blind detection, StegoWall, steganalysis.

1. INTRODUCTION

New possibilities of digital imaging and data hiding open wide prospects in modern imaging science, content management and secure communications. However, together with the obvious advantages of digital data hiding technologies and their current progress, these developments bring an associated danger for many domains such as copyright violation, prohibited usage and distribution of digital media, secret communications and network security.

The urgent necessity for copyright protection and data authentication, or rather the lack of adequate laws for digital multimedia protection, is recognized by the multimedia industry and research. Therefore, for example, if the copyright of images or videos is protected by someone using some unknown or proprietary digital watermarking technology, the practical proof of ownership or data origin seems to be very difficult without a third authorized trusted party.

Moreover, it is a well-known fact that many international terrorist organizations, competitive companies (aiming at intellectual property theft), military and industrial bodies, harassers and vandals benefit from steganography. Taking into account the practical difficulties to control and to verify steganographic communications, it is either very difficult or almost impossible to retrieve the hidden information from multimedia such as digital images, audio or video signals based on existing methods of steganalysis.

The drastic increase of public networks usage makes all computer systems more susceptible for attacks than ever before. Unfortunately, the limited capabilities of current protection devices, such as different types of Firewalls and Intrusion Detection sensors do not provide an adequate level of protection. Moreover, a subtle attack may not be immediately detectable. Another related and extremely dangerous “application” of data hiding technologies can appear in the form of novel mobile viruses and ancient “Trojan horses”. In these scenarios, the watermark or the hidden data might act as the virus activator that can destroy personal data or database content, or open an access to the protected or secret information (“trojanizing”) or even start an unauthorized data transfer. The practical danger of hidden activating agents is magnified by several orders of magnitude due to the perceptual invisibility of “Trojan horses” that can be hidden in innocent photos of greeting cards, commercial images or even popular songs.

* svolos@cui.unige.ch; phone: +41 22 7057637; http://watermarking.unige.ch; Department of Computer Science, University of Geneva, 24 rue Général-Dufour, 1211 Geneva, Switzerland

The existing state-of-art methods of steganalysis report certain level of progress in this direction. Westfeld and Pfitzmann were among the first who recognized the fact that the least significant bits (LSB) of images have a certain level of correlation and the data embedding destroys this relationship[1]. In particular it was reported that this effect could be observed as a change in color histograms. Therefore, they claim that at least different versions of known steganographic techniques such as EzStego, Jsteg, Steganos and S-Tools can be recognized based on the above fact.

Provos and Honeyman have extended the original work of Westfeld and Pfitzman to the DCT coefficient statistics instead of colors[2]. They use χ^2 –test to determine whether an image has distortions due to the hidden data embedding. The authors report successful detection of hidden data in Jsteg, JPHide and OutGuest. These authors also present a Stegdetect tool accompaigned by the web crawler and the corresponding parallel architecture.

Farid proposed to use higher-order statistics of natural images to detect hidden information[3]. Fridrich *et al* have demonstrated a possibility to detect image modifications caused by steganography and watermarking in images that were originally stored in the JPEG format[4]. This method investigates the compatibility of stego data with the results of possible DCT-based JPEG quantization. Their technique is able to identify the location of hidden bits and is thus able to estimate the size of hidden message. However, this method is unable to detect messages embedded using the DCT-based quantization index modulation. Fridrich *et al*[5] investigate the statistics of so-called regular and singular groups for the given mask. They have established the very interesting behavior of these statistics for real images that do not contain the hidden data. The data embedding in the LSB randomizes the LSB-plane and consequentially changes these statistics. An appropriate modeling of randomization influence and statistics extrapolation can recognize the fact of hidden data embedding.

It is necessary also to mention the first steganalysis system S-DART coming from industry[6]. Their approach to detect the hidden data based on image quality metrics without the reference to the original image was proposed in[7]. This approach was further extended for the identification of the used watermarking algorithm. Chandramouli and Memon presented a stochastic formulation of LSB steganography detection using a maximum a posteriori probability (MAP) detector[8]. The watermark embedding was considered as an additive antipodal signaling with a given rate of embedding and the cover image was assumed to be zero-mean stationary Gaussian process. That is definitely not the most accurate model for non-stationary processes such as image. However, one can note that in this formulation the problem is reduced to multiple hypotheses testing that is a typical case for M-ary modulation. We will extend this approach in this paper. A similar approach based on the Neyman-Pearson hypothesis test was proposed by the same authors in their previous work for on-off signaling model of watermark[9].

Despite obvious advantages in the design of modern steganalysis systems, a great amount of work still remains to be done to generalize the existing approaches into the strong theoretical framework of blind stochastic hidden data detection. Here we consider only the technical aspects of this problem, trying to initiate an open discussion about possible secure system architecture. Since the multimedia and computer industry have not yet agreed upon a universal standard for digital watermarking, a possible solution consists in international standardization and scientific cooperation, as well as in pushing for the development of national and industrial security systems.

Despite many differences in the design, the architectures and the applications of current data hiding methods, they still possess a number of common features that can be used for the detection, prevention and destruction of hidden invisible data. Briefly summarizing the impact of hidden data embedding on the cover data, one can note that the addition of hidden data sequentially leads to the modification of cover data statistics. It can be observed as:

- increase of the stego image entropy and randomization of LSB planes;
- damage of the correlation between the pixels within LSB-planes and changes of different groups of pixels appearance;
- degradation of the visual image appearance of LSB-planes;
- damage of correlation between different LSB-planes;
- change in the statistics between color image components and transform domain coefficients;
- introduction of “hidden” periodicities;
- “saturation” of the LSB-planes statistics due to randomization that manifests itself by the fact that the additional data embedding does not increase any more entropy. This is true first of all for the high rate embedding;

- damage of image feature integrity and digital image acquisition artifacts of CCD and CMOS-cameras, lossy compression or generally image processing artifacts and statistics (for example after thresholding or shrinkage denoising);
- change of scale-space relationship in wavelet data analysis that can be tracked by zero-trees structures and appropriate inter- and intra-scale stochastic image models;
- change of 1/f law via content-based data hiding technologies that can be identified and estimated;
- change of high-order statistics of cover data;
- difference in the objects statistics when an image is composed of several objects in MPEG-4 or other object manipulations.

The list of possible modifications is definitely not complete and can be extended taking into account particularities of data embedding algorithms and image statistics. However, all these facts prove the possibility to detect the hidden data or the image modifications.

This paper presents a generalized approach to hidden data detection and identification in different applications. The StegoWall goals, objectives and some practical operational scenarios are described in Section 2. Section 3 presents the stochastic detection part of StegoWall system. Section 4 concludes the paper.

2. STEGOWALL OBJECTIVES

The StegoWall can be considered as a multipurpose technology with a dual role. We explain the duality in more details in section 2.1. The StegoWall system aims at fulfilling a gap in modern security systems with respect to unauthorized hidden data exchange, transmission, usage, content authentication and integrity control. The distinctive features of the StegoWall system can be briefly summarized as:

- blind stochastic hidden data analysis and detection;
- identification of hidden data content and classification of the used data hiding technology to establish a possible data origin;
- hierarchical distributed system architecture that can be effectively implemented using parallel computing systems;
- prediction of possible tendencies in the development of advanced data hiding technologies that can be exploited in dual applications from the position of game-theoretical approach (i.e. enhance technology/attack technology).

The commonly accepted requirements for data hiding technologies such as digital watermarking, authentication, tamper proofing, self-recovering watermarks, and document security are perceptual invisibility, capacity, and robustness to certain types of attacks. However, it is not easy to simultaneously satisfy all above contradictory requirements. Therefore, a number of practical data hiding systems either reduce the requirements or completely neglect some of them. Another very important requirement both for watermarking and for steganography systems is the statistical invisibility of the hidden data (Figure 1). This requirement is often neglected by watermarking developers and may have an important impact on the reliability of the whole system in general.

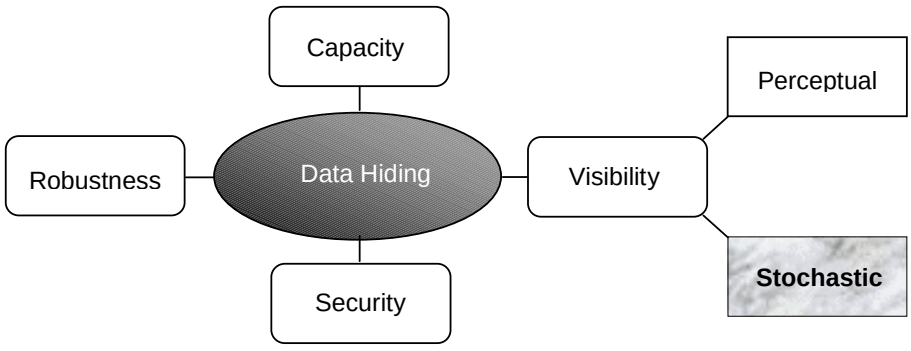


Figure 1. The trade-off that should be satisfied for data hiding technology.

Talking about the main applications of data hiding technologies one can distinguish between robust watermarking and steganography. We assume that the reader is familiar with these applications and therefore we only briefly summarize the main technical differences between them in Table 1.

Table. 1. The main differences between robust watermarking and steganography.

	Robust Watermarking	**Steganography**
Embedding rate, bits/pixel	$10^{-4} \le R < 10^{-1}$	To prevent stochastic detection
Robustness	High	Relatively low
Perceptual Visibility	To preserve commercial data quality	To prevent steganalysis
Stochastic Visibility	To prevent watermark removal	To prevent steganalysis
Security	Protocol security	High enough to prevent steganalysis
Equivalent watermark-to-noise ratio	Very low due to attacks	Relatively high

The main goal of this paper is to extend the apparatus of blind stochastic hidden data detection not only to robust watermarking and steganography, but also to integrity control and tamper proofing (possibly even without embedded hidden data) and to Internet/network security. The kernel of this approach is the StegoWall system that covers the above applications and can be used for the dual purposes, i.e. for the design and for the test of new technologies as well as for the opposite applications, consisting in attacking these technologies (Figure 2).

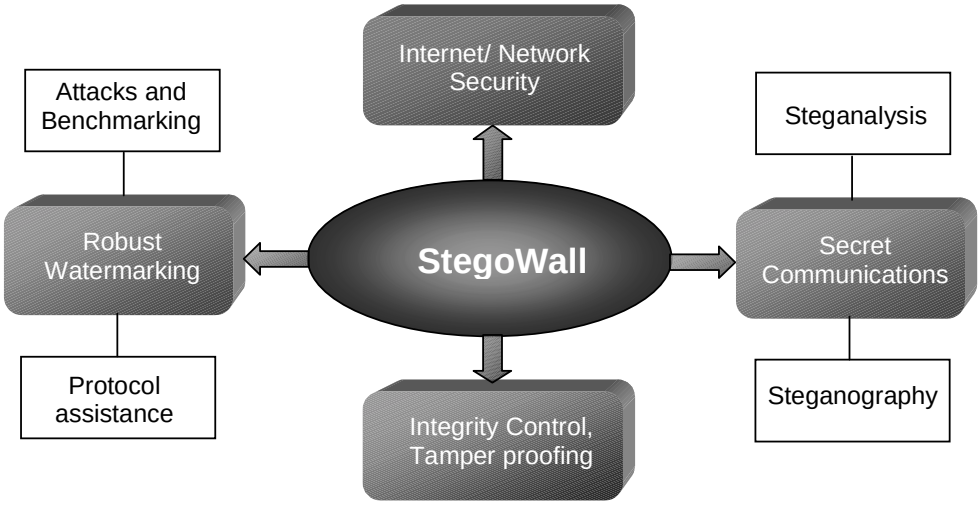


Figure2. The StegoWall operational scenarios.

2.1. Robust watermarking

The role of StegoWall in robust watermarking applications is dual and twofold. First, we consider the StegoWall as a system able to provide protocol assistance to robust watermarking. Second, the dual role of StegoWall consists in obtaining a fair evaluation of watermarking technology robustness and security during benchmarking exploiting a fact that watermarking technology might be identified and an attacking strategy might be selected appropriately.

To better understand the possibilities of the StegoWall system in robust watermarking applications, we briefly summarize the main factors that determine the statistics of watermarks. The main goal of robust watermarking is to satisfy the trade-off between capacity, robustness and visibility. The additional requirements are security (of the protocol and against unauthorized detection) and the ability to detect the watermark without the original or cover data, i.e., oblivious watermarking. Therefore, to satisfy the above requirements, one should properly select:

- watermark embedding/extraction domain (coordinate or transform domain such as Fourier, Cosine, Wavelet or Radon);
- method of message encoding/decoding (M-ary modulation or error correction codes (ECC)).

- method of watermark embedding: additive[10], multiplicative[11] or embedding with side information about cover image state, i.e. schemes such as [12], scalar Costa scheme (SCS)[13], quantization index modulation(QIM)[14].
- method used to recover from geometrical distortions, i.e. synchronization: methods using a transform invariant domain, methods based on an additional template, methods exploiting the self-reference principle based on an auto-correlation function (ACF)[15] or magnitude spectrum of a periodical watermark[16], methods using feature points;
- perceptual masking.

An important issue is the adaptation of the watermark to the properties of the human visual system (HVS), i.e. content-adaptive watermarking. Assuming we are given a masking function of the HVS, we wish to embed the watermark into the cover data keeping it under the threshold of visual perceptibility. Four main factors can be identified that should be taken into account for the design of perceptual masks based on the HVS:

- background luminance sensitivity;
- contrast sensitivity (MTF) depending on the subband or the resolution;
- orientation sensitivity (anisotropy);
- edge and pattern (texture) masking.

2.1.1. Protocol assistance

The protocol assistance to robust watermarking is a possible contribution to existing state-of-art in this area, since no watermarking standard exists and is not planned in the near future. Different industrial companies and academic institutions propose a variety of solutions targeting copyright protection and broadcast monitoring. Therefore, in some practical situations, an inexperienced user will not be able to identify if some image is watermarked at all (i.e., contain some copyright notice) or which watermarking technology is used. Such identification can be important for the reasons explained below. Another aspect of this problem is symmetric key exchange protocol that assumes that the same key used for the watermark embedding is used for the watermark extraction (note that key used for watermark encryption/decryption might be different). Most robust watermarking technologies are based on a symmetric pair of keys to avoid watermark removal. Some commercial solutions use the same key for all users and all images to simplify the task of Internet crawlers/spiders. However, the attackers can use this fact as a powerful hint to destroy a watermark or to violate the protocol.

We would like to emphasize again that the stochastic visibility of robust watermarking technologies is unavoidable due to the many conflicting requirements discussed in Section 1. Therefore, the StegoWall could efficiently resolve the problem of "asymmetric" watermarking detecting the presence of watermarks and determining the kind of used watermarking technology. The protocol assistance function of StegoWall can be divided on two parts:

- assistance on the user side;
- assistance on the service side.

In the framework of assistance on the user side, the StegoWall can serve as a:

- ***warning system*** that provides a message in Internet browsers or image processing tools (such as Photo Shop or Paint Shop Pro) or video tools (such as Adobe Premiere) that an image or video might contain a watermark and which most likely watermarking technology is used. This message could prevent many users from further usage of the data and copyright violation;
- ***prevention system*** that avoids multiple watermark embedding (time priority copyright protection);
- ***copyright acknowledgement system*** that can automatically deliver on user request an image of interest and corresponding URL from which this image was downloaded to the technology provider registration service or to the third authorized party depending on the used protocol to get information about buying conditions or further information, or to contact a content owner.

The possible scenarios we present here are not limited by these examples and can be extended depending on the particular protocol used.

The StegoWall usage on the service side can consist of:

- ***crawler assistance system*** that creates a priority in verification of images downloaded from some website depending on the results of the StegoWall detection. The images with the detected watermark and corresponding URL are sent to database where an appropriate key is supplied to perform the extraction and decoding of the watermark. If the watermark is successfully decoded with this key, it is assumed that this posting URL has the

appropriate rights. Oppositely, the complete spider verification process is activated. This assistance service could considerably save time and resources of crawler engines;

- ***broadcast monitoring assistance system*** that could provide a reporting service about time, place, URL, broadcaster based on the StegoWall detection results. This assistance will be especially useful in the advertisement applications and broadcasting with a restricted number of broadcasts.

2.1.2. Attack development and fair benchmarking

It is our strong belief that is almost impossible to design a robust watermarking system that is secure against unauthorized detection. The reason for that is the complex relationship between the different conflicting requirements to robust watermarking. Our analysis indicates that the stochastic visibility and consequentially watermarking security sacrify the most to satisfy this trade-off. Moreover, according to Kerckhoff's principle, a technology is supposed to be secure, if it is known except for one or more secret keys. In practice, an attacker (benchmarker) might not know all details of the watermarking algorithm under testing. Therefore, the power of his attack is restricted. The application of "blind attacks" such as in Stirmark[17] might not be always a fair evaluation. To fulfill this gap a second generation benchmarking was proposed[18] that was later implemented in the Checkmark benchmarking tool[19]. However, even in this enhanced benchmark, the information about watermarking technology is not completely used. Therefore, the StegoWall can be of great assistance in this application regarding estimation-based attacks.

The main idea of our approach consists in the fact that the stochastic visibility of a watermark can be a good evidence of the used technology or even the estimation of parameters of new technology that is not yet in the list of known techniques. Therefore, using this information one can specify an attacking strategy that tries to kill all those features or "fingerprints" that were detected by the StegoWall and uniquely describe the technology. For example, a "security leakage" of many technologies could be found in a template, periodicity of watermark, specific feature points detected without knowledge of the key and used for synchronization, known signal shape coding, method of watermark embedding and perceptual masking. Our statement is based on a simple motivation. Since these "security leakages" cannot be avoided the watermarking technology is using it without reference to the key. Therefore, according to Kerckhoff's principle, one can successfully attack them. The possible countermeasures describing how to decrease the stochastic visibility are considered in the next section.

2.2. Secret communications

The StegoWall also has a dual usage for secret communications applications. First, the StegoWall could be used as a tool for evaluation and design of new steganographic methods. Second, the StegoWall could be directly applied to steganalysis.

2.2.1. Steganography

The steganography, originally designed for cover or hidden communications, should provide a certain level of security for public communications. While most existing steganographic tools can provide perceptually invisible data hiding, the stochastic visibility or unauthorized detectability of hidden data still remains a challenging task. Therefore, to be secure, the steganographic system should satisfy a set of requirements. The main requirement consists in providing the statistical indistinguishability between the cover data and the host data. A possible information-theoretic measure of stochastic closeness is the relative entropy or *Kullback-Leiber distance* (KLD) between two distributions under test, which was first propose by Cachin[20]. More generally, the stochastic visibility can be considered as a possibility of unauthorized detection to differentiate between the cover and the host data based on a hypothesis testing.

We revise here the main countermeasures that can be used to decrease the stochastic visibility of watermark. The main idea of these countermeasures consists in the preservation of the statistical properties of the cover data, i.e., in the design of a data hiding technology with minimum possible stochastic distortions. We briefly summarize the main countermeasures aiming at reducing stochastic visibility of hidden data:

- Reduce the amount of modifications in the cover data, i.e., embedding. This will decrease the embedding capacity;
- Reduce the amount of modifications by applying some error correction codes with the error correction possibility corresponding to the amount of unchanged inappropriate data;

- Reduce the amount of distortions by applying data hiding in some transform domain where the amount of zero and non-zero coefficients might not be equiprobable due to decorrelation and energy compaction properties of the applied transform. Use the ECCs that produce the resulting codewords with corresponding statistics;
- apply encoding that uses the prior information about the host data as a side information at encoder;
- apply a transform that corrects the statistics of stego data but preserves capacity (for example preserve relative entropy or p.d.f. or bit-plane relationship);
- use content-based embedding assuming non-stationary Gaussian model of cover data, i.e., the model with locally smoothly changing variance. This will provide a possibility to embed more data in the textured areas which both perceptually and stochastically are less predictable and better preserve original content and hide modifications;
- choose the cover image from the set of images that provides the highest level of stochastic invisibility for a given message;
- synthesize a composite cover data for the same purpose.

The role of StegoWall in a steganographic perspective is considered as a feedback or benchmark that should evaluate the level of steganographic security.

2.2.2. Steganalysis

We consider the steganalysis part of StegoWall from two perspectives: passive and active. The passive steganalysis is reduced to a reporting function providing a binary decision about the presence or absence of hidden data in the given cover image. The active steganalysis is broader and includes detection, decryption, destroying, tracking and reporting to authorities. In fact, active steganalysis exploits the state-of-the art from cryptographic steganalysis for data decryption. It can also use the developed attacks against watermarking technologies that preserve image content[21] for hidden data destroying. The blind stochastic detection is presented in more details in Section 3.

2.3. Integrity control and tamper proofing

The StegoWall architecture is well fitted for integrity control and tamper proofing. We intend to exploit following data for the above verification:

- hidden data (watermarking or steganography): the knowledge of hidden data statistics can provide a reliable basis for the detection of modifications in images;
- natural statistics or artifacts of image acquisition devices (CCD or CMOS);
- statistics and common artifacts of image processing tools: compression, denoising, restoration;
- features integrity verification.

This information makes it possible to detect and to identify modifications, tampering, object compositions, applied image processing operations, the origin of data (type of image acquisition device).

2.4. Internet/Network security

The stego-viruses can be classified in three broad categories, i.e., those that attempt at:

- fooling the protocol;
- overloading the servers;
- activating unauthorized actions.

The viruses that fool communication/watermarking protocol are:

- the viruses that are targeting data hiding protocol violation. They can create the wrong pointers (both physical and URLs), date stamps, copyright (in applications such as Media Bridge (http://www.digimarc.com). They can change the annotations and the indexes for content-based retrieval assisting systems for images on the Internet or in databases of hospitals, museums, military applications, e-commerce, libraries and schools. For example, one can add instead of useful URL link to the e-library or educational program in Digimarc Media Bridge some links to pornography sites, violence, etc.;
- the viruses that create problems in e-commerce protocols via embedding of pseudo-watermarked data or data with hidden communications that might lead to refuse of system administrator or the other partners to post, to sell or to send data via Internet (since some hidden data is detected);

- the viruses that fool broadcast monitoring or Internet advertising systems based on protocol attacks that can copy a watermark labels in completely different content or create copyright ambiguity for law-enforcement institutions.

The viruses that overload the server are the viruses that overload crawler servers, steganalyzer or stegoscanner server work. For example, it can be accomplished by stegonoise distribution in multimedia data througth some innocent PhotoShop plug-ins that can add minor modifications to every image during processing or simply viewing. Stegonoise can contain either wrong watermark, some hidden message or just simple noise that looks like hidden info. Another source of stegonoise in multimedia may be image acquisition devices such as photo-, videocameras and scanners. It can be caused by the construction failure in design, or by the manufacturing failures in CCD or CMOS, or based on economic motivations of simplification producing or cost. It can be done of cuase by purpose to introduce some distortions that can be interpreted as stegonoise. In several years, all multimedia data produced by this sort of imaging devices might be infected by "stegonoise".

The viruses that activate unauthorized actions in some new plug-ins for image processing tools in PhotoShop, Paint Shop Pro, Netscape and Java plug-ins might cause the background processes aiming at:

- creating protocol ambiguity (removing of watermarks, or vice versa adding wrong watermarks);
- creating stegonoise in all data that goes through these software tools;
- activating "time bomb" viruses, or all kind of known viruses.

3. STOCHASTIC DETECTION

The basic idea of our complex approach is to exploit all information about hidden data statistics to perform its detection based on a Bayesian framework. Once the presence of hidden data is detected, the system performs data verification, display and steganalysis or finally destroys or modifies the data depending on its content or destination. Let $x[n]$ be a two-dimensional sequence representing the luminance of the original image, where $n=(n_1,n_2)$.

A watermark w is created by some key-dependent function $\mathbf{w}=\varepsilon(\mathbf{c},\mathbf{p},M,Key)$ that ensures the necessary spatial allocation of the watermark based on a key-dependent projection function $\mathbf{p}$, and according to HVS features as expressed by a perceptual mask M in order to improve the watermark. The resulting watermark is a linear combination of a set of L orthogonal funstions $\{p_i[n]\}, i=\{0,...,L-1\}$

$$w[n]=\sum_{i=0}^{L-1} c_i\, p_i[n] M[n] \tag{1}$$

where $\{p_i[n]\}$ satisfies $\langle p_i, p_j\rangle = \|p_i\|^2 \delta_{ij}$ and $\{c_i\}$ is either directly the message or previously encrypted and encoded data. Let $S=\{S_0,...,S_{L-1}\}$ be the sets of points where the pulses $\{p_i\}$ take nonzero values:

$$S_i \equiv \{[n] \mid p_i[n]\neq 0\}, i=\{0,...,L-1\}. \tag{2}$$

In the most cases, these sets assume non-overlapping pulses, i.e., $S_i \cap S_j$, $for\ \forall i\neq j$. This assumption guarantees that the pulses will always be orthogonal:

$$p_i[n]=\begin{cases} s[n], if\ \ n\in S_i \\ 0, otherwise. \end{cases} \tag{3}$$

where $s[n]$ is a key dependent zero-mean uncorrelated pseudorandom sequence with unit variance at each pixel. $s[n]$ is i.i.d. to provide maximum uncertainty (entropy). In most cases, the spreading function is chosen such that $s[n]\in\{-1,+1\}, \forall n$. The projection function performs a "spreading" of the data over the image area. It can be also considered as a diversity communication problem with parallel channels. Moreover, the projection function can have a particular spatial structure with the given correlation properties that can be used for the recovery of affine geometrical transforms as well as for the unauthorized detection of the watermark.

Therefore, the resulting stego image is:

$$y[n] = x[n] + w[n] \tag{4}$$

and the p.d.f. of the stego image is $p_{\mathbf{y}}(y) = p_{\mathbf{x}}(x) * p_{\mathbf{w}}(w)$. We consider further different stochastic models for the cover image and the watermark.

3.1. Cover image models

We consider here only the simplest class of independent identically distributed (i.i.d.) image models assuming that either the local mean is subtracted in the coordinate domain or the image is transformed in some domain such as DCT or wavelet. The simplest model that is very often used in steganalysis is i.i.d. stationary Gaussian model $x[n] \sim N(0, \sigma_x^2)$. Although this model leads to the tractable results in detection theory, it is very poorly reflects the statistics of real images. Therefore, we propose to use either i.i.d. non-stationary variance Gaussian model $x[n] \sim N(0, \sigma_{x_n}^2)$ or i.i.d. stationary Generalized Gaussian model (sGG) $x[n] \sim GGD(0, \gamma, \sigma_x^2)$. The relationship between these two models was demonstrated in our previous work[21].

3.2. Watermark models

The particular stochastic models of watermark depend on three factors according to the equation (1): statistics of encrypted and encoded message (mostly equiprobable), the projection function and the perceptual mask. We summarize in Table 2 the most common method of watermark generation and corresponding statistics. The embedding capacity in table 2 is denoted as q bits/pixels, $q \in (0,1)$.

Table 2. Different watermark generation methods and their pixel-wise and marginal statistics

Method of message embedding	Watermark	Hypothesis	Pixel-wise statistics p.m.f.	Marginal statistics p.m.f.
LSB, BPSK, constant mask	$w \in \Omega = \{-1,0,+1\}$	$H \in \Gamma = \{H_1, H_2, H_3\}$	$p_w(w_n) = \begin{cases} 0, \text{ with probability } 1-q \\ -1, \text{ with probability } q/2 \\ +1, \text{ with probability } q/2 \end{cases}$	The same
Generalized LSB. BPSK, constant mask	$w \in \Omega = \{-T,0,+T\}$	$H \in \Gamma = \{H_1, H_2, H_3\}$	$p_w(w_n) = \begin{cases} 0, \text{ with probability } 1-q \\ -T, \text{ with probability } q/2 \\ +T, \text{ with probability } q/2 \end{cases}$	The same
BPSK, content adaptive mask	$w \in \Omega = \{-T[n],0,+T[n]\}$	$H \in \Gamma = \{H_1, H_2, H_3\}$	$p_w(w_n) = \begin{cases} 0, \text{ with probability } 1-q \\ -T[n], \text{ with probability } q/2 \\ +T[n], \text{ with probability } q/2 \end{cases}$	Depends on the mask
M-ary PSK, content adaptive mask	$w \in \Omega = (-T[n];+T[n])$	$H \in \Gamma = \{H_1,...,H_M, H_{M+1}\}$	$p_w(w_n) = U\left(0, \frac{T[n]^2}{3}\right)$	Depends on the mask
Quantization (high-rates approximation)	$w \in \Omega = \left(-\frac{\Delta}{2};+\frac{\Delta}{2}\right)$	M hypothesis, depending on delta	$p_w(w_n) = U\left(0, \frac{\Delta^2}{12}\right)$	$w \sim U\left(0, \frac{\Delta^2}{12}\right)$

3.3. Pixel-wise watermark detection

The watermark detection can be considered based on the available statistics about the watermark and the cover image. We present here three main approaches to watermark detection investigated within the StegoWall project. The simplest watermark detection problem can be considered using multiple hypothesis testing. It is a common practice to use for this purpose a MAP detector that minimizes the probability of error (although a NP detector can be used as well) assuming known signaling. In this case, the problem is similar to the M-ary pulse amplitude modulation (PAM). The hypothesizes under test are:

$$\begin{aligned} H_0 &: y[n] = x[n] - T \\ H_1 &: y[n] = x[n] \\ H_2 &: y[n] = x[n] + T \end{aligned} \tag{5}$$

The MAP detector takes the decision:

$$H = \arg\max_j p(H_j) p(y_n | H_j) \text{ for all } n. \tag{6}$$

We have investigated this type of detection for different cover image models. The probability of error $P_e = \frac{4}{3} Q\left(\sqrt{\frac{NT^2}{4\sigma_x^2}} \right)$ is a well known expression for M-ary PAM (LSB watermark with a constant mask) and the stationary Gaussian cover image model. In our case *M*=3 and we have six types of errors. *N*=1 for the detection on the sample level (pixel-wise detection) and *N* can be increased, if the watermark bit allocation scheme is known and the watermark is repeated. Therefore, the reliable detection is only possible for relative small image variance that is a case for the flat image regions and/or relatively large sample space *N*.

In the case of a random watermark, i.e. a watermark weighted by an unknown perceptual mask, we can assume that the watermark is a zero mean Gaussian random process with a known covariance (i.e., some information about periodicity might be available). Then the presence of watermark is detected according to the next hypothesis testing:

$$\begin{aligned} H_0 &: y[n] = x[n] \\ H_1 &: y[n] = x[n] + w[n] \end{aligned} \tag{7}$$

In this case, a NP detector decides about the presence of the watermark, if the likelihood ratio exceeds a threshold or if:

$$L(\mathbf{y}) = \frac{p(\mathbf{y}; H_1)}{p(\mathbf{y}; H_0)} > \gamma \tag{8}$$

The typical assumptions can be $\mathbf{y} \sim N(\mathbf{0}, \sigma_x^2 \mathbf{I})$ under H_0 and thus $\mathbf{y} \sim N(\mathbf{0}, \sigma_x^2 \mathbf{I} + \mathbf{C}_w)$ under H_1. The resulting detector will consist of so-called *estimator-correlator structure*. The watermaking community often uses this kind of detection.

Finally, one can consider the problem of known deterministic signal detection (both antipodal or on-off signaling) in i.i.d. Generalized Gaussian noise. That is the typical case for the transformed domain detection when the image is essentially decorrelated and its energy is packed in several non-zero coefficients. Since the watermark energy is much lower that the energy of the cover image, one can extend this problem to the detection of a weak signal (i.e. watermark) in i.i.d. Generalized Gaussian noise. The resulted detector can be expressed in the form of a locally optimum detector (LOD)[22].

3.4. Watermark estimation

One can also first estimate the parameters of watermark and then make the decision or more generally classify the watermark. The MAP estimator is well suited for this purpose:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p_{\mathbf{x}}(\mathbf{y}|\mathbf{w}) p_{\mathbf{w}}(\mathbf{w}) \tag{9}$$

If we assume $\mathbf{x} \sim N(\bar{\mathbf{x}}, \mathbf{C}_{\mathbf{x}})$ and $\mathbf{w} \sim N(\mathbf{0}, \sigma_w^2 \mathbf{I})$, where $\mathbf{C}_{\mathbf{x}}$ is diagonal, we receive a well-known Wiener filter as the solution of the MAP estimation problem:

$$\hat{w}[n] = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_{x_n}^2} (y[n] - \bar{y}[n]) \tag{10}$$

The variance of this estimate is:

$$\mathbf{Var}[\hat{w}[n]] = \frac{\sigma_w^2 \sigma_{x_n}^2}{\sigma_w^2 + \sigma_{x_n}^2} \tag{11}$$

that again supports the conclusion that the accurate estimation of a watermark is only possible for relativly small σ_x^2, i.e., in flat image areas.

3.5. Watermark classification

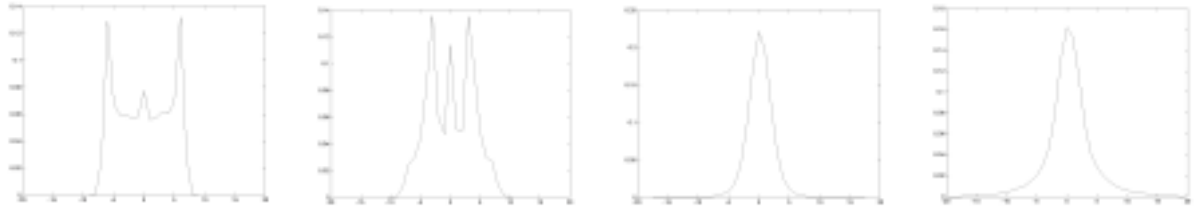
The critical issue of the classification problem is the selection of appropriate watermark features that are well distinguished in the classification space. Due to the limited paper space we only present here two statistics that can be used for

classification problem. As an example, the watermark distribution can be used to uniquely classify watermarking technologies. To demonstrate this statement we have computed the watermark distributions for three industrial technologies Digimarc (http://www.digimarc.com), EverSign (http://www.dct-group.com) and SysCop (http://www.mediasec.com), and Berkut developed at University of Geneva (http://watermarking.unige.ch) (Figure 3). One can see the obvious differences between them.

If w and $\hat{w}$ are two finite random watermarks taking values in the same set Ω, then the *Kullback-Leiber distance* $D(p_{\hat{w}} \| p_w)$[23], from the probability distribution $p_{\hat{w}}$ to the probability distribution p_w is defined to be

$$D(p_{\hat{w}} \| p_w) = \sum_{\substack{w \in \Omega \\ p_{\hat{w}}(w) \neq 0}} p_{\hat{w}}(w) \log \frac{p_{\hat{w}}(w)}{p_w(w)}, \tag{12}$$

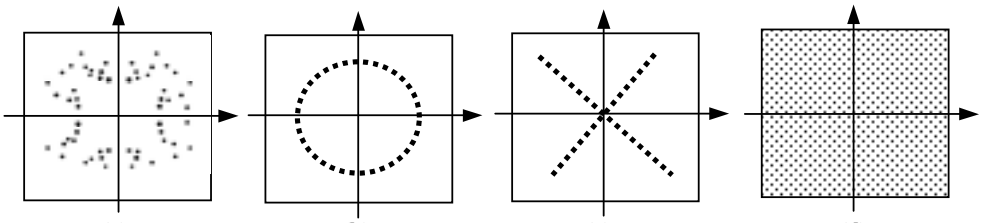
where p_w is computed from the pseudo-watermark (noise) and the mask directly computed from the stego-image and $p_{\hat{w}}$ is computed from the predicted watermark directly.



(a) (b) (c) (d)

Figure 3. Watermark distributions: (a) Digimarc from PhotoShop 5 with durability 4, (b) EverSign (Digital Copyright Technologies), (c) SysCop (Mediasec with GGD approximation and shape parameter 1.1), (d) Berkut (University of Geneva with GGD approximation and shape parameter 0.92).

The second distinguished feature of robust watermarking is the kind of used synchronization. One can see different synchronizations presented as the magnitude spectra (Figure 4).



(a) (b) (c) (d)

Figure 4. Synchronization detection in the DFT and the watermark parameters estimation: (a) Digimarc template, (b) and (c) common circular and diagonal template patterns, (d) typical pattern produced by the periodical watermarks using self-synchronization.

4. CONCLUSIONS

We have presented in this paper the StegoWall system dedicated for the detection of hidden data based on a stochastic framework. The StegoWall system covers four main applications: robust watermarking, secret communications, integrity control and tamper proofing, and Internet/Network security. The stochastic framework is based on multiple hypothesis testing and utilizes available prior information about the cover image and the watermark. We have demonstrated the urgent necessity in the development of systems such as StegoWall. However, we believe that the development of advanced steganalysis systems requires strong international collaboration between different academic, industrial, governmental and law-enforcement institutions and open public discussion of possible countermeasures against “electronic” terrorist activity. Our next steps consist in the extension of distributed computing to the available computational resources of CUI (University Center of Informatics) at the University of Geneva and several industrial companies.

ACKNOWLEDGMENT

We are grateful to Prof. Jessica Fridrich for interesting discussions on this subject. This work was partially supported by the European Certimark project.

REFERENCES

1. A. Westfeld and A. Pfitzmann, "Attacks on steganographic systems", *In Proc. of 3rd International Workshop on Data Hiding*, Springer Verlag, September 1999.
2. N. Provos and P. Honeyman, "Detecting Steganographic Content on the Internet", *ISOC NDSS'02*, San Diego, CA, February 2002, to appear. *(available from* http://www.citi.umich.edu/u/provos/).
3. H. Farid, "Detecting Steganographic Messages in Digital Images", *Technical Report*, TR2001-412, Dartmouth College, Computer Science. (http://www.cs.dartmouth.edu/~farid/publications/tr01.html).
4. J. Fridrich, M. Goljan and R. Du, "Steganalysis Based on JPEG Compatibility", *SPIE Multimedia Systems and Applications IV*, Denver, CO, August 20-24, 2001.
5. J. Fridrich, M. Goljan and R. Du, "Detecting of LSB Steganography in Color and Gray-scale Images", *Magazine of IEEE Multimedia*, Special Issue on Multimedia and Security, pp. 22-28, October-December 2001.
6. Steganography Detection & Recovery Toolkit (S-DART), http://www.wetstonetech.com/sdart.htm, December 5, 2001.
7. N.D. Memon, I. Avcibas and B. Sankur, "Steganalysis techniques based on image quality metrics", *Proc. of SPIE, Security and Watermarking of Multimedia Contents III*, 4314, San Jose, USA, January 2001.
8. R. Chandramouli and N.D. Memon, "Analysis of LSB Based Image Steganography Techniques", *Int. Conference on Image Processing ICIP2001*, Thessaloniki, Greece, 2001.
9. R. Chandramouli and N.D. Memon, "A distributed detection framework for watermark analysis", *ACM Multimedia Workshop on Multimedia and Security*, CA, USA, 2000.
10. J. Hernandez and F. Perez-Gonzalez, "Statistical analysis of watermarking schemes for copyright protection on images", *Proc. IEEE*, Vol. 87, No 7, pp. 1142-1166, July 1999.
11. M. Barni, F. Bartolini, V. Cappellini, A. Piva, F. Rigacci, "A MAP identification criterion for DCT-based watermarking", *Proceedings of the IX European Signal processing Conference EUSIPCO'98*, pp. 1513-1516, Island of Rhodes, Greece, September 8-11, 1998.
12. M. Ramkumar, A.N. Akansu, "Self-Noise Suppression Schemes for Blind Image Steganography", *SPIE Special Session on Image Security*, Vol. 3845, pp 55-65, Boston, MA, September 99.
13. J.J. Eggers, J.K. Su and B. Girod, "A Blind Watermarking Scheme Based on Structured Codebooks," *IEE Colloquium: Secure Images and Image Authentication*, London, UK, April 2000.
14. B. Chen and G.W. Wornell, "Provably robust digital watermarking", *In Proc. of SPIE*, Vol. 3845, pp. 43-54, Boston, USA, September 1999.
15. M. Kutter, "Watermarking resisting to translation, rotation, and scaling", *Proc. of SPIE, Multimedia systems and applications*, Vol. 3528, pp. 523-531, San Jose, USA, November 1998.
16. S. Voloshynovskiy, F. Deguillaume, T. Pun, "Content adaptive watermarking based on a stochastic multiresolution image modeling", *submitted EUSIPCO'2000*.
17. F.A. Petitcolas, R.J. Anderson, M.G. Kuhn. Attacks on copyright marking systems, *In Information Hiding: Second International Workshop IH'98*, Portland, Oregon, USA., April 15-17, 1998, Proceedings, LNCS 1525, Springer-Verlag, ISBN 3-540-65386-4, pp. 219-239.
18. S. Voloshynovskiy, S. Pereira, V. Iquise and T. Pun, "Attack modelling: Towards a second generation benchmark", *Signal Processing*, Special Issue: Information Theoretic Issues in Digital Watermarking, 2001. V. Cappellini, M. Barni, F. Bartolini, Eds., **81**, **6**, pp. 1177-1214, June 2001.
19. S. Pereira, S. Voloshynovskiy, M. Madueño, S. Marchand-Maillet and T. Pun, "Second generation benchmarking and application oriented evaluation", *In Information Hiding Workshop*, Pittsburgh, PA, USA, April 2001.
20. C. Cachin, "An information-Theoretic Model for Steganography", *In Information Hiding: Second International Workshop IH'98, Preprocedings*, Portland Oregon, USA, April 15-17 1998.
21. S. Voloshynovskiy, S. Pereira, V. Iquise and T. Pun, "Attack modelling: Towards a second generation watermarking benchmark," *Signal Processing*, Special Issue on Information Theoretic Issues in Digital Watermarking, **81**, pp. 1177-1214, 2001.
22. S. Kassam, *Signal detection in non-Gaussian noise*, Springer-Verlag, New York, 1998.
23. T. Cover and J. Thomas, *Elements of Information Theory*, Wiley series in telecommunications, John Wiley&Sons, Inc., New York, 1991.