Attacks on Digital Watermarks: Classification, Estimation-based Attacks and Benchmarks

S. Voloshynovskiy, S. Pereira, T. Pun, University of Geneva
J.J. Eggers and J.K. Su, University of Erlangen-Nuremberg

ABSTRACT

Watermarking is a potential method for protection of ownership rights on digital audio, image and video data. Benchmarks are used to evaluate the performance of different watermarking algorithms. For image watermarking, the Stirmark package is the most popular benchmark, and the best current algorithms perform well against it. However, results obtained by the Stirmark benchmark have to be handled carefully since Stirmark does not properly model the watermarking process and consequently is limited in its potential for impairing sophisticated image watermarking schemes. In this context, the goal of this article is threefold. First, we give an overview of the current attacking methods. Second, we describe attacks exploiting knowledge about the statistics of the original data and the embedded watermark. We propose a stochastic formulation of estimation-based attacks. Such attacks consist of two main stages: a) watermark estimation; b) exploitation of the estimated watermark to trick watermark detection or create ownership ambiguity. The full strength of estimation-based attacks can be achieved by introducing additional noise, where the attacker tries to combine the estimated watermark and the additive noise to impair watermark communication as much as possible while fulfilling a quality constraint on the attacked data. With a sophisticated quality constraint it is also possible to exploit human perception, e.g., the human auditory system in case of audio watermarks and the human visual system (HVS) in case of image and video watermarks. Third, we discuss the current status of image watermarking benchmarks. We briefly present Fabien Petitcolas' Stirmark benchmarking tool [1]. Next, we consider the benchmark proposed by the University of Geneva Vision Group that contains more deliberate attacks. Finally, we summerize the current work of the European Certimark project, whose goal is to accelerate efforts from a number of research groups and companies in order to produce an improved ensemble of benchmarking tools.

INTRODUCTION

Digital watermarking is a communication method in which the information b is embedded directly and imperceptibly into digital data x, e.g., image, video, or audio signals, also called **original data** or **host data**, to form **watermarked data** y. Loosely analogous to watermarks in paper documents, the embedded information is bound to the watermarked data wherever it goes. The embedded information should still be decodable from the watermarked data, even if the watermarked data is processed, copied, or re-distributed. Potential applications of digital watermarking include copyright protection, distribution tracing, authentication and conditional-access control. Thus, the information b could be for instance a user-ID, a serial number for a certain copy of a document, or authentication information.

We will concentrate our analysis on the copyright protection of still images that is an urgent problem for modern e-commerce. Obviously, the attacks introduced in the paper can be applied to audio and video watermarking algorithms with the safe of generality and with the technical modifications depending on the physics of the considered media.

The watermark can be regarded as an additive signal w, which contains the encoded and modulated watermark message b under constraints on the introduced perceptible distortions given by a mask M, so that

$$y = x + w(M).$$

Note that w must not necessarily be independent from the original data x. The most simple approach to achieve a perceptually indistinguishable watermarked and original signal is to keep the power of the watermark signal very low. Using sophisticated psycho-acoustic or psycho-visual models, more appropriate masks M can be applied to enhance the robustness of the watermarking scheme. Commonly used embedding techniques can be classified into **additive [2], multiplicative [2],** and **quantization-based schemes [3,13].** In additive schemes, there are usually very weak dependencies between w and x, e.g., introduced by choosing w dependent on a data-dependent perceptual mask M. In multiplicative schemes, samples of the original data are multiplied by an independent signal v, so that $w=xv-x$. Here, w and x are of course dependent on each other. Strong local dependencies between the realizations of w and x exist in quantization based watermarking schemes. However, these dependencies are such that statistically x and w appear (almost) independent.

The term "watermark" itself is not always well-defined in the literature. To be precise, we have to distinguish between the watermark signal w, which is the actual signal added to the original data, and the watermark message or information b that is conveyed by the watermark signal. Usually the meaning is clear from context. Coding schemes can be used to achieve reliable watermark communication. In some cases only one bit need to be communicated (on-off signaling), while in other cases a sequence of M-ary watermark symbols is transmitted.

In most watermarking applications, the marked data is likely to be processed in some way before it reaches the watermark receiver. The processing could be lossy compression, signal enhancement or D/A and A/D conversion. An embedded watermark may unintentionally or inadvertently be impaired by such a processing. Other types of processing may be applied with the explicit goal of hindering watermark reception. In watermarking terminology, an **attack** is any processing that may impair detection of the watermark or communication of the information conveyed by the watermark. The processed, watermarked data is then called **attacked data**.

An important aspect of any watermarking scheme is its robustness against attacks. The notion of robustness is intuitively clear: A watermark is robust if it cannot be impaired without also rendering the attacked data useless. Watermark impairment can be measured by criteria such as miss probability, probability of bit error, or channel capacity. For multimedia, the usefulness of the attacked data can be gauged by considering its perceptual quality or distortion. Hence, robustness can be evaluated by simultaneously considering watermark impairment and the distortion of the attacked data. An attack succeeds in defeating a watermarking scheme if it impairs the watermark beyond acceptable limits while maintaining the perceptual quality of the attacked data.

Since the complete theoretical analysis of the watermarking algorithm performance with respect to different attacks is rather complicated, the developers of watermarking algorithms refer to the results of experimental testing performed in the scope of some benchmark. The benchmark combines the possible attacks into a common framework and weights the resulted performances depending on the possible application of the watermarking technology.

This article discusses several state-of-the-art attacks that deliberately attempt to impair the watermark without excessively distorting the attacked data. Where possible, we take a general point of view to highlight attack principles that can be applied in many different circumstances. When it comes to practical examples, we usually refer to image watermarking since the background of the authors is strongest in this application area.

First, we briefly summarize state-of-the-art watermarking attacks and coarsely categorize them. Next, we discuss a relatively new direction: attacks based on watermark estimation. Combining the estimated watermark with additional noise in an optimized way can significantly improve the strength of attacks. Finally, three different approaches for benchmarking image watermarking schemes are presented.

STATE-OF-ART WATERMARKING ATTACKS

One categorization of the wide class of existing attacks contains four classes of attacks: removal attacks, geometric attacks, cryptographic attacks, and protocol attacks. Here, we describe coarsely these four attack types and present some examples. More detailed descriptions can be found in [7, 12].

Removal attacks

Removal attacks aim at the complete removal of the watermark information from the watermarked data without cracking the security of the watermarking algorithm, e.g., without the key used for watermark embedding. That is, no processing, even prohibitively complex, can recover the watermark information from the attacked data. This category includes denoising, quantization (e.g., for compression), remodulation, and collusion attacks. Not all of these methods always come close to their goal of complete watermark removal, but they may nevertheless damage the watermark information significantly.

Sophisticated removal attacks try to optimize operations like denoising or quantization to impair the embedded watermark as much as possible while keeping the quality of the attacked document high enough. Usually, statistical models for the watermark and the original data are exploited within the optimization process.

Collusion attacks are applicable when many copies of a given data set, each signed with a key or different watermark, can be obtained by an attacker or a group of attackers. In such a case, a successful attack can be achieved by averaging all copies or taking only small parts from each different copy. Recent results show that a small number of different copies, e.g., about 10, in the hand of one attacker can lead to successful watermark removal.

Geometric attacks

In contrast to removal attacks, **geometric attacks** do not actually remove the embedded watermark itself, but intend to distort the watermark detector synchronization with the embedded information. The detector could recover the embedded watermark information when perfect synchronization is regained. However, the complexity of the required synchronization process might be too great to be practical.

For image watermarking, the most known benchmarking tools, Unzign and Stirmark [1], integrate a variety of geometric attacks. Unzign introduces local pixel jittering and is very efficient in attacking spatial domain watermarking schemes. Stirmark introduces both global and local geometric distortions. We give a few more details about these attacks later in this paper. However, most recent watermarking methods survive these attacks due to the use of special synchronization techniques. Robustness to global geometric distortions often relies on the use of either a transform-invariant domain (Fourier-Melline) or an additional template or of specially designed periodic watermarks whose autocovariance function (ACF) allows estimation of the geometric distortions. However, as will be discussed below, the attacker can design dedicated attacks exploiting knowledge of the synchronization scheme.

Robustness to global affine transformations is more or less a solved issue. However, resistance to the local random alterations integrated in Stirmark still remains an open problem for most commercial watermarking tools. The so-called random bending attack in Stirmark exploits the fact that the human visual system is not sensitive against local shifts and affine modifications. Therefore, pixels are locally shifted, scaled, and rotated without significant visual distortion. However, it is worth noting that some recent methods are able to resist against this attack.

Cryptographic attacks

Cryptographic attacks aim at cracking the security methods in watermarking schemes and thus finding a way to remove the embedded watermark information or to embed misleading watermarks. One such technique is the brute-force search for the embedded secret information. Another attack in this category is the so-called Oracle attack, which can be used to create a non-watermarked signal when a watermark detector device is available. Practically, application of these attacks is restricted due to their high computational complexity.

Protocol attacks

Protocol attacks aim at attacking the entire concept of the watermarking application. One type of protocol attack is based on the concept of **invertible watermarks [11]**. The idea behind inversion is that the attacker subtracts his own watermark from the watermarked data and claims to be the owner of the watermarked data. This can create ambiguity with respect to the true ownership of the data. It has been shown that for copyright protection applications, watermarks need to be non-invertible. The requirement of non-invertibility of the watermarking technology implies that it should not be possible to extract a watermark from a non-watermarked document. A solution to this problem might be to make watermarks signal-dependent by using one-way functions.

Another protocol attack is the **copy attack**. In this case, the goal is not to destroy the watermark or impair its detection, but to estimate a watermark from watermarked data and copy it to some other data, called **target data** [6]. The estimated watermark is adapted to the local features of the target data to satisfy its imperceptibility. The copy attack is applicable when a valid watermark in the target data can be produced with neither algorithmic knowledge of the watermarking technology nor the knowledge of the watermarking key. Again, signal-dependent watermarks might be resistant against the copy attack.

ESTIMATION-BASED ATTACKS

Here, we consider attacks that take into account the knowledge of watermarking technology and exploit statistics of the original data and watermark signal [4-8,10]. In addition, we emphasize that for the design of attacks against watermarking schemes, the distortion of the attacked document and the success of watermark-impairment has to be considered. Within the scope of these attacks, we present the concept of **estimation-based attacks**. This concept is based on the assumption that the original data or the watermark can be estimated - at least partially - from the watermarked data using some prior knowledge of the signals' statistics. Note that estimation does not require any knowledge of the key used for watermark embedding. Further, knowledge of the embedding rule is not required, but the attack can be more successful when it is.

Depending on the final purpose of the attack, the attacker can obtain an estimate of the original data or of the watermark based on some stochastic criteria such as maximum likelihood (ML), maximum a posteriori probability

(MAP), or minimum mean square error (MMSE). We do not focus here on the particularities of the above estimation but rather concentrate on different ways to exploit the obtained estimates to impair the embedded watermark. Depending on the way the estimate is used, we can classify estimation-based attacks as removal, protocol, or desynchronization attacks.

Estimate of the original data

Considering the watermark as noise in the watermarked data, the attacker can try to estimate the original, unwatermarked data. This attack results in the design of an optimal denoising scheme. Taking into account the results of recent investigation that established the strong connection between denoising and compression for filtering of additive noise from the images, this means in the case of image watermarks that the attacker can easily apply the most recent advanced coders based on wavelet decompositions to remove the watermark. Keeping in mind the design of such coders in terms of an optimal rate-distortion trade-off, the attacker can obtain a considerable gain in resolving the compromise between distortions introduced by the attack and success in removal of the watermark. Note that in both cases of denoising and of optimized compression, both the perceptual quality and the objective quality of the attacked image can be improved significantly. We classify both denoising and optimized compression as removal attacks.

Remodulation attacks

Remodulation attacks aim at modification of the watermark using modulation opposite to that used for watermark embedding. Assuming the estimated watermark is correlated with the actual watermark, meaning a good estimate could be obtained, the estimated watermark can be subtracted from the watermarked data. Subtracting a very inaccurate estimate of the watermark might decrease the document quality without affecting the watermark too much. On the other hand, correlation-based detection can be defeated by subtracting an amplified version of the estimated watermark. For this reason, we introduced a gain factor gamma >= 1, which gives us the possibility to trade off the distortion of the attacked document versus the success of the attack.

There are four basic variations of the remodulation attack. First, when gamma = 1, the attack yields the MMSE/MAP estimate of the original and reduces to the denoising attack. Second, for gamma >1, the quality of the attacked document might be reduced, but correlation based detection might be defeated more successfully. The attack can even drive the correlation to zero so that the detector incorrectly decides that the watermark is not present in the attacked data.

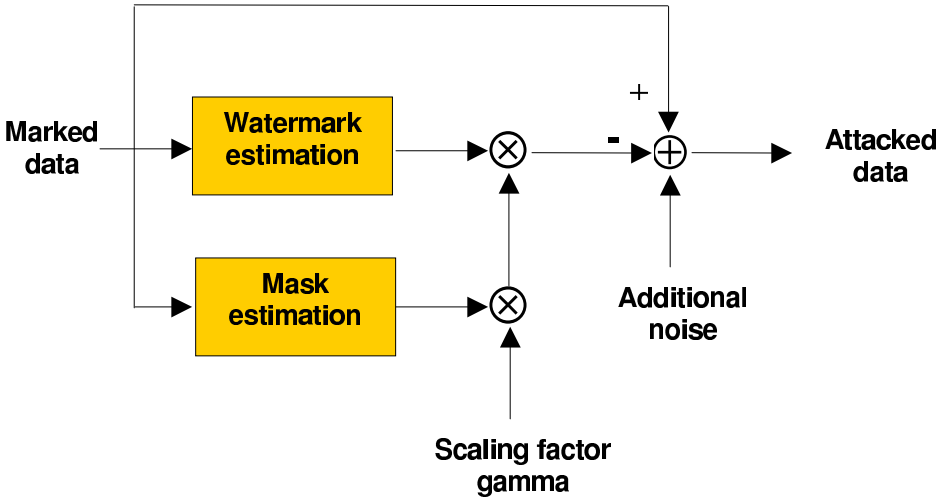

Figure 1. *Perceptual remodulation attack.*

Third, when using a more sophisticated distortion measure than the simple MSE, a better compromise between success of the attack and introduced distortion can be obtained by weighting the remodulated watermark by a perceptual mask. Fourth, the attacker can not only subtract the weighted, estimated watermark, but also add outliers to obtain a non-Gaussian noise distribution, which decreases the performance of correlation-based detection. Moreover, exploiting features of the human perceptual system, the attacker can efficiently embed a large amount of outliers in perceptually less-significant parts of the data. For image data, this approach has been demonstrated to be successful in [4]. We refer to this attack as **perceptual remodulation.**

Copy attack

The estimated watermark can be exploited to implement a copy attack, as already described. Of course, the copied watermark has to be adapted to the target data to keep the quality of the falsely watermarked target data high enough. There are many practical ways to adapt the watermark to the target data based on perceptual models. For images, contrast sensitivity and texture masking phenomena of the HVS can be exploited. The estimation-based copy attack is most successful when the same perceptual model is used as in the original watermarking algorithm.

Note that the copy attack in its described version is mainly applicable to additive watermarking schemes. In the case of quantization based watermarking schemes, even a perfectly estimated watermark signal w cannot be copied since it is highly unlikely that the copied signal w is a valid watermark in the target signal.

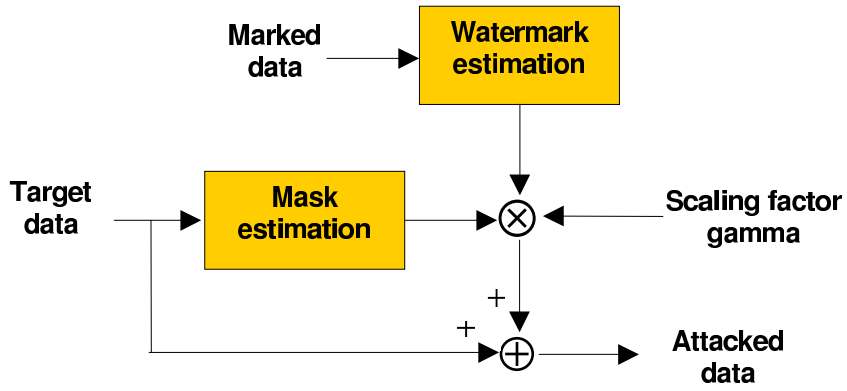

Figure 2. *Copy attack.*

Synchronization removal

Watermark estimation can also be very efficiently applied to attack synchronization mechanisms The basic idea of **synchronization removal** is to detect synchronization patterns, remove them, and then apply desynchronization techniques, e.g., global affine transformation in the case of image watermarking.

Here, we concentrate on synchronization methods for image watermarking based on a template in the magnitude image spectrum or on the ACF of periodic watermarks. In both cases, peaks are generated in the Fourier domain [4,7]. It is obvious that such peaks can be easily detected. Once the peaks have been detected, the next step of the attack is to interpolate the spectrum of the watermarked image or previously attacked image in the locations of spatial frequencies determined by a local peak detector. The generalized block diagram of this attack is shown in Figure 3 where $\Im$ and $\Im^{-1}$ denote direct and inverse Fourier transforms, respectively, $|\cdot|$ is the magnitude, and arg is the phase.

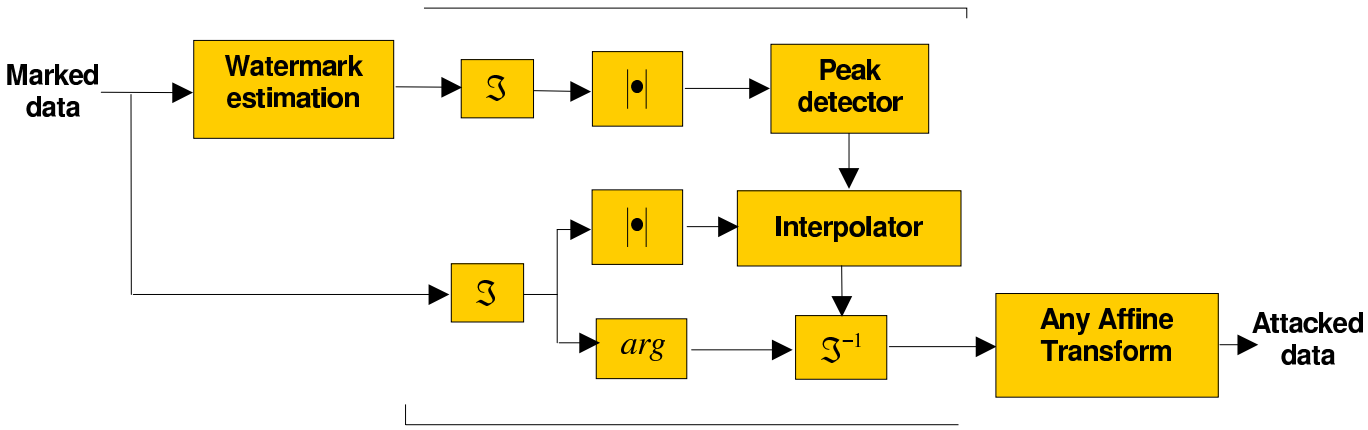

Figure 3. *Synchronization removal attack.*

Countermeasures against estimation-based attacks

To resist estimation-based attacks, the embedder aims at making the watermark difficult to estimate. This approach has been investigated for two different scenarios.

Power-Spectrum Condition (PSC): An idealized theoretical approach [5] for analyzing estimation-based attacks treats the original signal and watermark as independent, zero-mean, stationary, colored Gaussian random processes. The watermarked data is the sum of these two processes. Since the original signal is given, its power spectrum is assumed fixed, but the watermark power spectrum can be varied. The question is, "How should the watermark power spectrum be shaped to resist an estimation-based attack?" For this scenario, the optimal estimate is obtained by a Wiener filter.

The mean-squared error E between the original watermark and the estimated watermark provides a convenient way to measure how well a watermark resists estimation. It can be shown that E is maximized if and only if the watermark power spectrum is directly proportional to the power spectrum of the original signal. This requirement is called the power-spectrum condition (PSC). A watermark whose power spectrum satisfies the PSC is the most resistant against estimation.

If distortion is measured by the mean-squared difference between the attacked data and the unwatermarked, original data, then the PSC has another important consequence: For any output of the matched filter, a watermark that fulfills the PSC causes the above attack to incur the greatest distortion. To drive the correlation to zero, the attack must make the distortion as large as the power of the original data, so the attacked data is unlikely to be useful.

Noise Visibility Function (NVF): The PSC is attractive because it can be proven rigorously and has a convenient mathematical form. For image watermarking, image denoising provides a natural way to develop estimation-based attacks [9] optimized for the statistics of images, although optimality might be difficult to prove. The watermarked image is treated as a noisy version of the original image, and the watermark represents noise that should be eliminated. Thus, the estimated watermark is the same as the estimated noise.

We applied different statistical models for the original images, namely a non-stationary Gaussian process (1), or a stationary, generalized Gaussian process (2). The noise/watermark can be treated as one of these processes, however, here we assume that it is still a stationary Gaussian process. In case (1), the denoising method uses an adaptive Wiener filter, while in case (2) it reduces to the popular denoising methods of hard-thresholding and soft-shrinkage as particular cases.

Both denoising methods produce a **texture masking function** (TMF), which is derived from the image statistics and is therefore image-dependent. The TMF takes on values in [0,1]. To embed a watermark that resists such estimation, the watermark embedding should use the inverted function that is known as a **noise visibility function** (NVF), defined by NVF = 1 - TMF. NVF values near unity indicate flat regions, where the watermark should be attenuated, while NVF values near unity indicate texture or edge regions where the watermark should be amplified. In this way, the watermark is embedded to resist estimation-based attacks derived from image denoising.

A qualitative comparison sheds light on the structure of watermarks produced in this manner. Figure 4a shows the original Cameraman image and the NVFs derived using cases (1) and (2), and Figures 4b depicts the corresponding magnitude spectra. The figure clearly shows that the resulting watermark spectra are closely matched to the power spectrum of the original image. Note that in case of images, the PSC can give only a coarse result since the underlying statistical model does not fit very closely to images. Nevertheless, the results obtained using the PSC agree with those of the NVF. The two approaches complement each other well. These results support the heuristic argument of Cox et al. [2] that the watermark should be placed in the "perceptually significant frequency components".

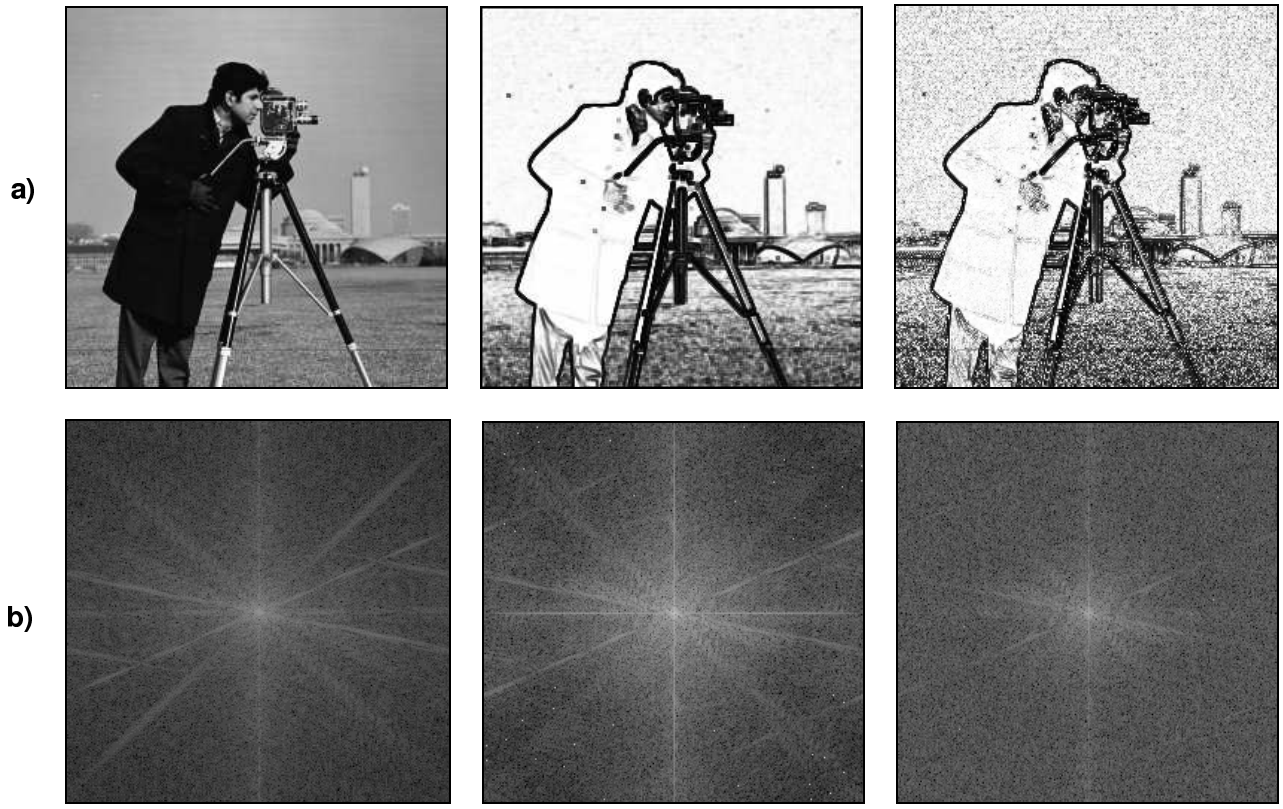

Figure 4. *(a) From left to right: original image, NVF estimated based on the non-stationary Gaussian and stationary Generalized Gaussian image models and (b) corresponding spectra.*

Attacks dependent on local signal statistics

Different estimation-based attacks can be successfully combined depending on local signal statistics. Here we focus on image watermarking. The attacker is motivated to reduce the maximum rate of reliable communication also by exploiting the HVS and the possibility to remove the watermark based on different models of the image. The watermark can be easily predicted and removed from flat image areas rather than from the areas of edges and textures. The stochastic model for the edges and textures is non-stationary and relatively complicated to use for accurate image estimation. Therefore, the attacker will try as much as possible to utilize the advantages of denoising and to remove the watermark from the flat areas without visual distortions and even enhancing the PSNR. In contrast, the attacker will use the remodulation with an increased strength in the edges and texture areas, which again will be masked by the HVS. At the same time, the attacker can use the NVF to automatically determine the flat regions, edges and textures. This attack is schematically shown in Figure 5b. The practical power of this attack was first demonstrated in [4] and a further theoretical analysis can be found in [7, 8].

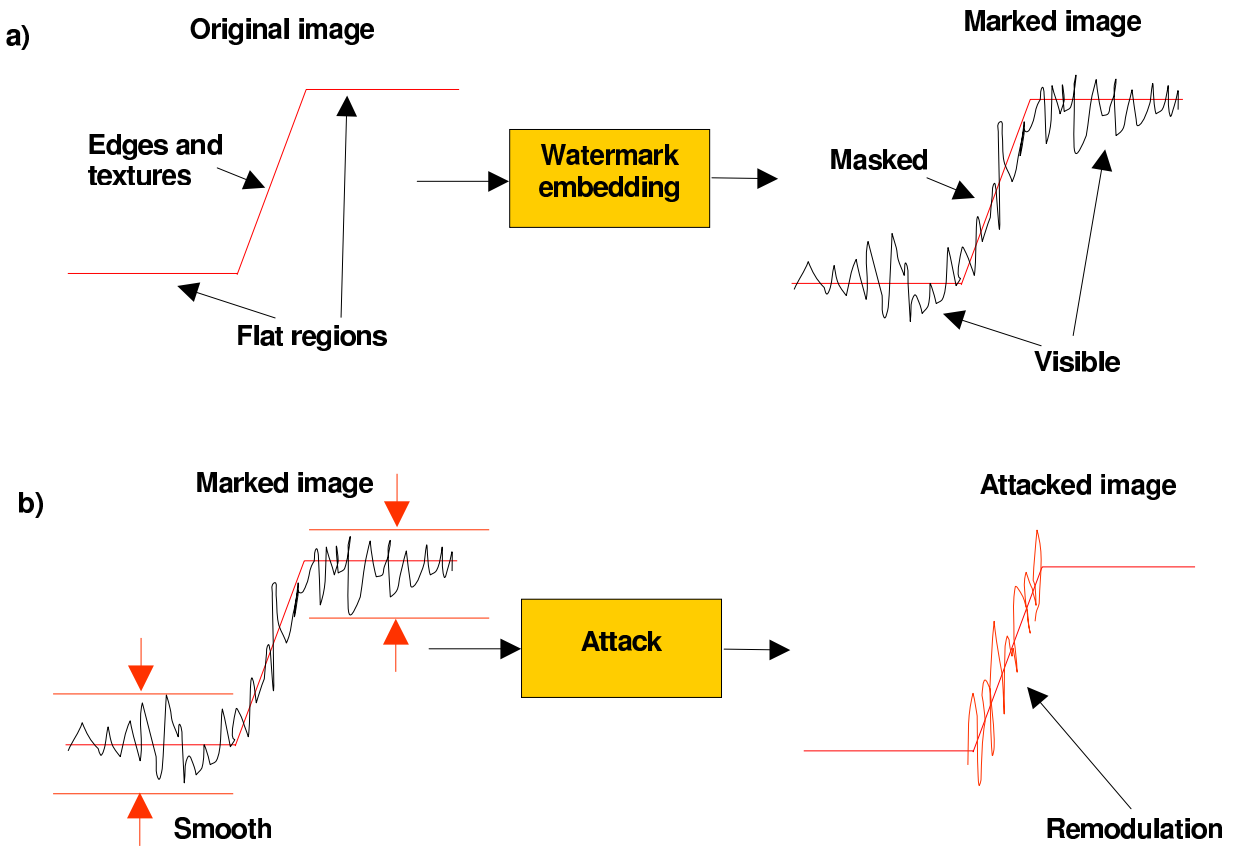

Figure 5. *The data hider strategy exploiting the texture masking function of the HVS (a), the attacker strategy using denoising and perceptual remodulation (b).*

Optimized attacks

So far, we developed different attacks based on watermark estimation and discussed how an embedder can make watermark estimation as difficult as possible. However, the embedder has another chance to react to estimation based attacks, in particular to the remodulation attack. The detector could first estimate the remodulated watermark and try to invert the remodulation, and thus retaining reliable watermark detection. For instance, when watermark estimation is based on Wiener filtering, the detector could apply inverse Wiener filtering. This is of course not wanted by the attacker. Thus, he has to add noise to the attacked data, which would be amplified by the inverse Wiener filtering and thus impair watermark detection. Of course, the additive noise further degrades the attacked signal.

Now, we have the situation that the attacker has to find a good combination of using the estimated watermark and the noise to be added. The embedder no longer simply tries to make watermark estimation as hard as possible, however, he also has to get a power advantage over the additive noise an attacker might introduce. This problem can be formulated in an even more general way: the attacker attempts to minimize the watermark capacity under a constraint on the distortions introduced by the attack. The embedder attempts to maximize the watermark capacity under a constraint on the embedding distortions. This situation can be regarded as a game between the attacker and embedder. To solve this problem, we followed a game-theoretic approach, assuming that the embedder and attacker know each other's behaviour [8,10].

We consider the scenario where the original signal and watermark are drawn from independent, zero-mean, stationary, colored Gaussian random processes. In this case a solution to the described game can be found. It appears that the optimal watermark power allocation for reliable watermarking is dependent on the amount of distortion that can be introduced by an attacker. At low distortion levels, white watermarks perfrom near optimally, where at high distortion, PSC-compliant watermarks are more appropriate. Watermarks embedded into signal components with low variance, e.g. high frequencies or flat regions in images, will be filtered out by the optimized attacks, while watermarks embedded into signal components with high variance are more efficiently disturbed by additive noise. To apply these results to real-world watermarking applications, good signal decompositions have to be used to separate signal components with different statistics. For images, the wavelet decompositon might be a good choice. However, research

in this direction is still inconclusive.

BENCHMARKING

In this section we present three benchmarking initiatives for image watermarking schemes: Stirmark [1], a benchmark including estimastion based attacks as proposed in [7], and the current Certimark initiative.

Stirmark

The Stirmark benchmark [1] divides attacks into the following 9 categories: signal enhancement, compression, scaling, cropping, shearing, rotation, linear transformations, other geometric transformations, and random geometric distortions. In the case of signal scaling, cropping, shearing, rotation, linear transformations, and other geometric transformations, the attacked images are obtained with and without JPEG 90% quality factor compression. In order to produce a score relative to the benchmark, a score of 1 is assigned when the watermark is decoded and 0 when the watermark is not decoded. The average is then computed for each of the 9 sections, and the average of the results is computed to obtain an overall score. The benchmark should also be averaged over several images. In order to ensure a fair comparison, Petitcolas suggests imposing a minimum PSNR of 38dB for the watermarked image. However, this constraint is questionable since PSNR is not a meaningful measurement in the context of geometric distortions.

Benchmark including estimation based attacks

While the Stirmark benchmark is an excellent tool for measuring the robustness of watermarking algorithms, it is heavily weighted towards geometric transformations, which do not take into account prior information about the watermark. Therefore, another benchmark was proposed [7], which contains the following 6 categories of attacks:

- Denoising : Wiener filtering, soft shrinkage and hard thresholding.
- Denoising followed by perceptual remodulation.
- Denoising followed by Stirmark random bending.
- Copy Attack: the watermark is estimated using Wiener filtering and copied onto another image. If the watermark is successfully detected in the new image the algorithm has failed.
- Template removal followed by small rotation.
- Compression based on wavelet decomposition of the image, which is superior to JPEG compression and is integrated to reflect the future appearance of the JPEG2000 standard.

Also included in this benchmark is a new proposal for evaluating the image quality. Rather than using PSNR, two approaches based on weighted PSNR and Watson's metric are proposed. The Watson metric measures local image characteristics and reflects luminance, contrast, and texture masking.

CERTIMARK

The lack of officially accepted benchmarking tools fastered the launch of the European project Certimark (CERTIfication for waterMARKing techniques) that started in May 2000. The project includes 15 academic and industrial partners. The objectives of Certimark are:

- to design, develop and publish a complete benchmark suite for watermarking technologies within promising application scenarios;
- to make this benchmark suite a reference tool for technology suppliers and for technology customer within promising applications scenarios;
- to set up a certification process for watermarking algorithms;
- to concentrate research on pending key issues in watermarking for protection of still images and low bit rate video over the Internet.

The aim of CERTIMARK, based on the benchmark reference, is to make watermarking algorithms labelled with an international certification. This certification process and award will be conducted by a major international rights holder representative. Benchmarking, leading to an internationally recognised reference, will permit customers to assess the appropriateness of a given watermarking technology for their needs. Assessment of technologies in a clear framework will allow competition between technology suppliers while maintaining a given quality standard as measured by the benchmark. In summary, the originality and emphasis of the project will be on the parallel development of objective evaluation tools and robust watermarking techniques to be certified in order to win the confidence of content providers.

CONCLUSIONS

Attacks on digital watermarking systems have been investigated in this paper. First, a categorization of different attacks was given, and popular attacks were briefly described. We pointed out that attacks on digital watermarks must consider both watermark survival and the distortion of the attacked document. Early attacks do not exploit as much knowledge of the watermarking scheme as possible, also they do not consider the distortion of the attacked document. Since attacks can be improved by using knowledge of the watermarking scheme and the signal statistics, we described a new set of attacks which were called estimation-based attacks. The general idea is to estimate the watermark and exploit it to trick the detector. It was shown that this approach is related to denoising, but can be extended to a variety of different attack methods. Next we described how an embedder can try to resist estimation based attacks, which leads to the concept of PSC-compliant watermarks or watermarks adapted to the NVF, depending on the application and signal model at hand. Finally, we explained how considering the watermarking and attacking problem as a game between embedder and attacker can be exploited to find the watermark capacity, when facing an optimized attack with a constrained attack distortion. The theoretical analysis of watermark attacks gives many insight into the watermarking problem, and enables us to show fundamental limits of this technology. However, theory does not render benchmarks of practical watermarking algorithms useless. Contrary, both concepts complement each other. Therefore, we described and compared some recent watermark benchmark initiatives in the last part of the paper.

REFERENCES

[1] M. Kutter and F. Petitcolas, "A fair benchmark for image watermarking systems," *Electronic Imaging 199: Security and Watermarking of Multimedia Content*, Vol. **3657** of SPIE Proceedings, San Jose, California USA, 25-27 January 1999.

[2] I. Cox, J. Kilian, T. Leighton and T. Shamoon, "Secure spread spectrum watermarking for multimedia," IEEE Trans. Image. Processing, 6, No 12, pp. 1673-1687, Dec. 1997.

[3] B. Chen and G. W. Wornell, "Dither modulation: a new approach to digital watermarking and information embedding," In Proc. of SPIE Vol. 3657: Security and Watermarking of Multimedia Contents, San Jose, CA, USA, January 1999.

[4] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgärtner and T. Pun, "Generalized watermark attack based on watermark estimation and perceptual remodulation," In Ping Wah Wong and Edward J. Delp eds., *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, Vol. **3971** of SPIE Proceedings, San Jose, California USA, 23-28 January 2000.

[5] J. K. Su, B. Girod, "Power-spectrum condition for energy-efficient watermarking," In Proc. IEEE Int Conf. Image processing ICIP99, October 1999.

[6] M. Kutter, S. Voloshynovskiy and A. Herrigel, "Watermark copy attack," In Ping Wah Wong and Edward J. Delp eds., *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, Vol. **3971** of SPIE Proceedings, San Jose, California USA, 23-28 January 2000.

[7] 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, 2001.

[8] J. K. Su, J.J. Eggers and B. Girod, "Analysis of digital watermarks subjected to optimum linear filtering and additive noise," Signal Processing, Special Issue on Information Theoretic Issues in Digital Watermarking, 2001.

[9] S. Voloshynovskiy, A. Herrigel, N. B. and T. Pun, "A stochastic approach to content adaptive digital image watermarking," In *International Workshop on Information Hiding*, Vol. **LNCS 1768** of Lecture Notes in Computer Science, pp. 212-236, Springer Verlag, Dresden, Germany, 29 September -1 October 1999.

[10] P. Moulin and J. O'Sullivan, "Information-theoretic analysis of information hiding," submitted to IEEE Trans. Inform. Theory.

[11] S. Craver, N. Memon, B. Yeo, and M. Young, "On the invertibility of invisible watermarking techniques," Proc. Of the IEEE Int. Conf. On Image Processing 1997, Vol. 1, p. 540-543.

[12] F. Hartung, J.K. Su, and B. Girod, "Spread Spectrum Watermarking: Malicious Attacks and Counter-Attacks," Proc. of SPIE Vol. 3657: Security and Watermarking of Multimedia Contents, San Jose, CA, USA, Jan. 1999.

[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] A. B. Watson, " DCT quantization matrices visually optimized for individual images", in Proc. SPIE: Human vision, Visual Processing and Digital Display IV,Vol. 1913, pp.202-216, 1993.

BIOGRAPHIES

SVIATOLSAV VOLOSHYNOVSKIY (svolos@cui.unige.ch) received a Radio Engineer Degree from the Lviv Polytechnic Institute in 1993, and a Ph.D. degree in Radio Engineering and Television Systems from State University "Lvivska Polytechnika", Ukraine, in 1996. He immediately started as Assistant Professor at the above University. In 1998 he has been at the Coordinated Science Lab, University of Illinois at Urbana-Champaign as a visiting scholar. Currently he is a Research Assistant Professor at the University of Geneva, Switzerland, and an Associate Professor at State University "Lvivska Polytechnika". He has over 70 journal and conference papers, and 4 patents in radar imaging, smart antenna arrays, image restoration, steganography and digital watermarking.

SHELBY PEREIRA (shelby.pereira@cui.unige.ch) received the Bachelor's degree in Electrical Engineering in 1995. He then completed a Master's in biomedical engineering at Ecole Polytéchnic in Montreal in 1997 where his research considered image restoration problems in tomography. He received the PhD from the University of Geneva in the computer science in 2000. He is currently a senior researcher at the same University. Research interests include image watermarking, attacks on watermarking systems, spread spectrum communications, detection and estimation theory.

THIERRY PUN (thierry.pun@cui.unige.ch) obtained his Ph.D. in image processing in 1982, at the Swiss Federal Institute of Technology in Lausanne (EPFL). He joined the University of Geneva, Switzerland in 1986, where he is currently full professor at the Computer Science Department. Since 1979 he has been active in various domains of image processing, image analysis and computer vision. He has authored or co-authored over 150 journal and conference papers and 3 patents in these areas, and led or participated to a number of national and European research projects. His current research interest is focused on several aspects of the design of multimedia information systems: image and video content-based information retrieval systems, Web browser for blind users, and image and video watermarking.

JOACHIM J. EGGERS (eggers@LNT.de) received the Dipl.-Ing. in electrical engineering from the University of Technology (RWTH) Aachen, Germany, in 1998. Since 1998 he is a member of the Image Communication Group at the Telecommunications Laboratory of the University of Erlangen-Nuremberg, Germany, where he is working towards his Ph.D. His research interests include digital watermarking of multimedia data, attacks on digital watermarks and information-theoretic aspects of digital watermarking.

JONATHAN K. SU (SU@LNT.DE) received the Ph.D. degree from the Georgia Institute of Technology in 1997. In 1998 he joined the Telecommunications Laboratory at the University of Erlangen-Nuremberg, Germany, as a Member of Research Staff. In October 2000, he joined MIT Lincoln Laboratories in Lexington, Massachusetts, USA. His research interests include digital image and video compression, digital watermarking and information hiding, human visual perception, and optimal estimation. He was co-presenter of the tutorial on image and video watermarking at the 2000 IEEE International Conference on Image Processing in Vancouver, British Columbia, Canada. He served as co-chair of the 1999 DFG (Deutscheforschungsgemeinschaft) V3D2 Watermarking Workshop in Erlangen, Germany, and is co-author of the book "Introduction to Optimal Estimation" (Springer 1999). Dr. Su is a member of Phi Beta Kappa, Sigma Xi, Tau Beta Pi, and Eta Kappa Nu.