Proceedings of SPIE:
Security and Watermarking of Multimedia Content II,
Vol. 3971, January 2000,
San Jose, CA, USA.

The Watermark Copy Attack

Martin Kutter[a] and Sviatoslav Voloshynovskiy[a,b] and Alexander Herrigel[a]

[a]Digital Copyright Technologies, Stauffacher Strasse 149, 8004 Zürich, Switzerland

[b]CUI-University of Geneva, 1211 Geneva, Switzerland

ABSTRACT

Research in digital watermarking has progressed along two paths. While new watermarking technologies are being developed, some researchers are also investigating different ways of attacking digital watermarks. Common attacks to watermarks usually aim to destroy the embedded watermark or to impair its detection. In this paper we propose a conceptually new attack for digitally watermarked images. The proposed attack does not destroy an embedded watermark, but copies it from one image to a different image. Although this new attack does not destroy a watermark or impair its detection, it creates new challenges, especially when watermarks are used for copyright protection and identification. The process of copying the watermark requires neither algorithmic knowledge of the watermarking technology nor the watermarking key. The attack is based on an estimation of the embedded watermark in the spatial domain through a filtering process. The estimate of the watermark is then adapted and inserted into the target image. To illustrate the performance of the proposed attack we applied it to commercial and non-commercial watermarking schemes. The experiments showed that the attack is very effective in copying a watermark from one image to a different image. In addition, we have a closer look at application dependent implications of this new attack.

Keywords: digital watermarking, attack, robustness

1. INTRODUCTION

About 700 years after the invention of the well known paper watermarks in Fabriano,[3] Italy, a similar concept was applied to digital media, such as images, audio and video. Research efforts in the field of digital watermarking are increasing fast all over the world. Furthermore, people such as content providers and intellectual property owners are more and more in need for efficient solutions to protect, track, and monitor digital media.

Since the beginning of the digital watermarking area the technologies have evolved in an impressive way, resulting in very efficient commercial solutions. Furthermore, besides the idea of hiding information in digital data in a robust manner, the concept of fragile digital watermarks for the verification of data integrity emerged. For an introduction to digital watermarking and an overview of different technologies the reader is referred to Hartung and Kutter.[3]

Similar to the paper watermark industry, the invention of digital watermarks not only triggered worldwide research on digital watermarking technologies, but also ways of attacking, counterfeiting and falsifying digital watermarks with the goal of making illegal profit or bypassing laws. The research on watermark attacks has the positive side effect that it encourages people doing research on digital watermarking technologies to develop new methods which are resilient to the attacks. In this article we focus on attacks on robust digital watermarking schemes. We will introduce a new attack, called the watermark copy attack, which has an important impact on current overall digital watermarking solutions for a variety of applications such as copyright protection, monitoring and tracking.

Before we have a closer look at the proposed attack, we give an overview in Section 2 of current attacks. In Section 3 we present the new attack and then outline the conceptual parts of the attack, that is watermark prediction

Further author information:
M.K.: E-mail: martin.kutter@kutter.ch
S.V.: E-mail: svolos@cui.unige.ch
A.H.: E-mail: alexander.herrigel@dct-ch.ch

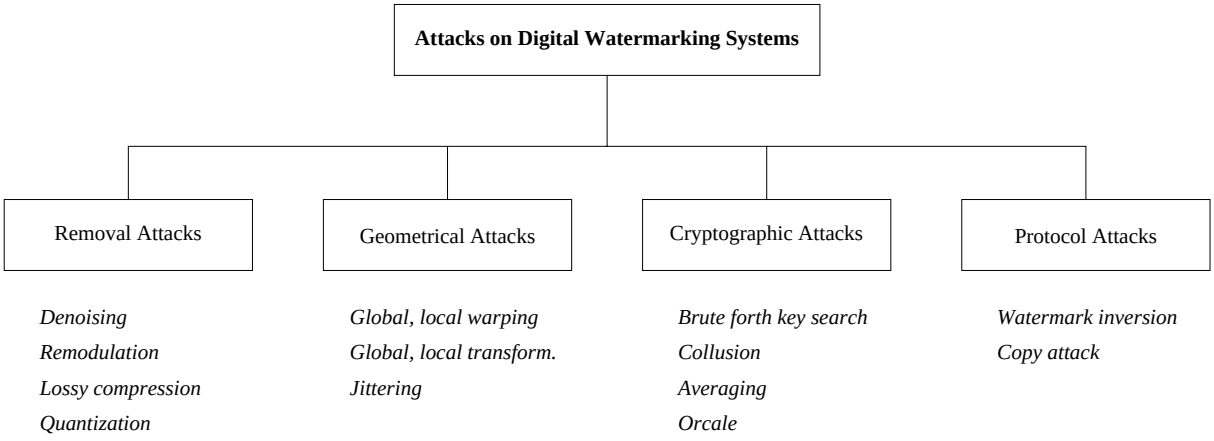


Figure 1. Types of attacks on digital watermarking systems.

in Section 4, and copy insertion in Section 5. To test the efficacy of the attack we applied it to two publicly available watermarking tools. The results of these attacks are presented in Section 6. Before drawing the final conclusions in Section 8 we will discus some implications of the attack in Section 7.

2. ATTACKS ON WATERMARKING SYSTEMS

A closer look at malicious attacks on robust digital watermarking schemes shows that there are mainly four inherently different attacking concepts: removal attacks, geometrical attacks, cryptographic attacks, and protocol attacks. We will now briefly look at each concept. In addition, Figure 2 shows a summary of the different attacks.

Removal attacks aim at completely removing a watermark from the cover data. These approaches consider the inserted watermark as noise with a given statistic and try to estimate the original, non watermarked cover data from the stego, that is watermarked, data. Efficient removal attacks based on denoising have for example been proposed by Langelaar et al.[8] They propose to apply a sequence operations to the watermarked image, including median filtering, highpass filtering, and non-linear truncation. Voloshynovskiy et al.[15] propose a spatial watermark prediction trough a filtering process based on a maximum a posteriori (MAP) watermark estimation with following remodulation to create the least favorable noise distribution for the watermark detector.

In contrast to removal attacks, geometrical attacks intend not to remove the embedded watermark, but distort it through spatial or temporal alterations of the stego data. The attacks are usually such that the watermark detector looses synchronization with the embedded information. The result is the failure of the watermark detection process, and this although the watermark is still in the data. First attacks based on geometrical alterations have been proposed for digital images. Here mainly two utilities are to mention, Unzign and Stirmark. Unzign[13] introduces local pixel jittering and is very efficient in attacking spatial domain watermarking schemes. Stirmark[5,6] introduces local geometrical bending in addition to a global geometrical transformation. For natural images this attack introduced barely noticeable artifacts and at the moment of writing this article no known watermarking technology is resilient to it.

Cryptographic attacks are very similar to attacks used in cryptography and may be of different nature. There are the brute forth attacks which aim at finding a secret through an exhaustive search. Since many watermarking schemes use a secret key it is very important to use keys with a secure length. Another attack is the so called Oracle[10,2,1] attack which can be used to create a non-watermarked image when a watermark detector device is available. Other attacks in this group are statistical averaging, and collusion attacks. The former describes an attack in which many instances of a given data set, each time signed with a different key or different watermark, are averaged to compute the attacked data. If the number of data sets is large enough, the embedded watermark may no be detected anymore. In the collusion attack, again many instance of the same data are available, but this time the attacked data set is generate by tacking only a small part of each data set and rebuilding an new attacked data set from these parts.

The attacks in the last group, the protocol attacks, do neither aim at destroying the embedded information nor disable the detection of the embedded information through local or global data manipulation. The goal of these

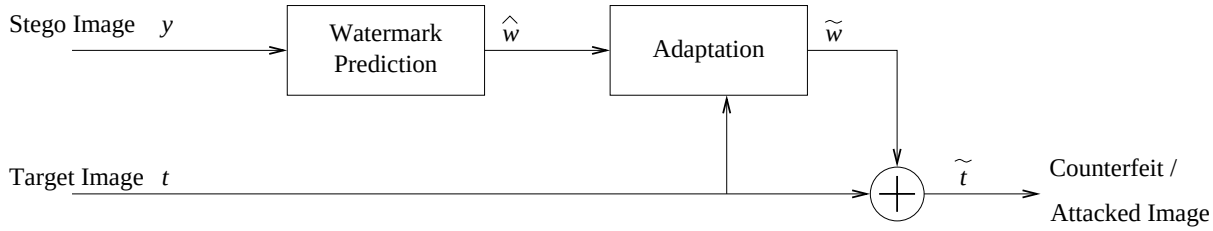


Figure 2. The building blocks of the watermark copy attack.

attacks is to attack the concept of the watermarking application. The first protocol attack was proposed by Craver et al.[16] They introduced the concept of invertible watermarks and showed that for copyright protection purpose watermarks need to be non-invertible. This requirement on the watermarking technology means that it should not be possible to extract a watermark from a non-watermarked image.

Although the presented classification allows for a clear separation between the attacks, it should be noted that very often a malicious attacker applies not only one single attack at the moment, but rather a combination of two or more attacks.

The watermark copy attack proposed in this article belongs to the last group, the protocol attacks. The goal of the attack it to copy a watermark from stego data to the target data without having any specific knowledge about the watermarking technology. As we will see in Section 7, depending on the application of the digital watermarking technology, this attack may have very serious implications.

3. THE WATERMARK COPY ATTACK

As mentioned above, the idea of the watermark copy attack belong to the group of protocol attacks, which means the goal of the attack is not to destroy the embedded watermark, but jeopardize the application for which digital watermarks are used. The basic idea of the attack is to copy a watermark from one image to another image, and this without an prior information about the watermarking technology and additional information such as the secret key.

Figure 2 shows the functional blocks of the attack. The inputs to the system are the stego image, which contains the watermark to be copied, and the target image, into which the watermark form the stego image is to be copied. From a technical point of view the attack consist of three main steps. In the first step the watermark in the stego image is predicted, resulting in $\hat{w}$. The prediction is then processed in the next step. The goal of this processing is to adapt the watermark to the target image in order to maximize its energy under the constraint of keeping it imperceptible after insertion in the target image. In the last step, the predicted and processed image is added to the target image.

4. WATERMARK PREDICTION

4.1. Introduction

Predicting the embedded watermark is the key operation in the watermark copy attack and to a large extend influences the effectiveness of the attack. Predicting the embedded watermark can be performed in two ways: 1) direct prediction, 2) denoising. In this work we focus on watermark prediction through denoising as it allows us to use the well developed apparatus of the denoising theory.

Independent of the watermarking technology employed, we can model the watermarking process as an addition of a watermark to the cover image in the spatial domain:

$$y = x + w, \tag{1}$$

where y is the stego image, x the original/cover image and w the watermark. Considering the stego image as a noisy image, then the watermark is the noise and we can compute an estimate of the noise/watermark $\hat{w}$ by taking the difference between the estimate $\hat{x}$ of the cover image and the stego image:

$$\hat{w} = y - \hat{x}. \tag{2}$$

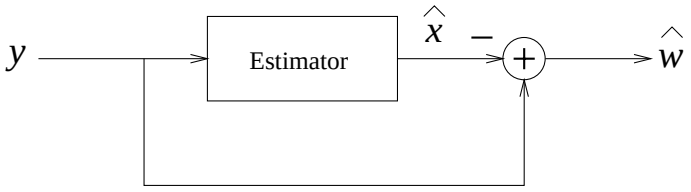


Figure 3. Watermark prediction through denoising.

This approach of predicting the embedded watermark through denoising is illustrated in Figure 3.

If assume to have no prior information about the statistics of the stego image we can use a *maximum likelihood (ML)*-estimate of the watermark. On the other hand, a *Maximum a posteriori Probability (MAP)* estimate can be used if we assume to have prior information about the image statistics.[4] Depending on the statistics, closed form solutions exist in certain cases for both approaches. An overview of the solutions for special cases is shown in Figure 4.

4.2. ML-estimate

If we assume to have no prior information on the statistics of the image and prior information about the statistics of the noise/watermark, we can use a ML-estimator. The estimator is given by:

$$\hat{x} = \arg\max_{\tilde{x}\in\mathbb{R}^N} \left\{\ln p_w\left(y|\tilde{x}\right)\right\}, \tag{3}$$

where $p_w(.)$ is the probability density function of the watermark.

The ML-estimate has a closed form solution for the two cases when the watermark has either a Gaussian or a Laplacian distribution. If the watermark has a Gaussian distribution the ML-estimate is given by the local mean of y:

$$\hat{x} = localmean(y); \quad \text{Gaussian watermark.} \tag{4}$$

On th other hand, if the watermark features a Laplacian distribution the solution of the ML-estimate is given by the local median:

$$\hat{x} = localmedian(y); \quad \text{Laplacian watermark.} \tag{5}$$

Computing the ML-estimate of the cover image therefore reduces to computing the local mean or local median. To do so, several approaches exist such as simply computing the average or median in a square window. However, if we assume to work with natural images then we can compute more accurate estimates of the local mean or median by considering only pixel in a cross-shaped neighborhood. This is due to the fact that natural images feature a higher correlation in horizontal and vertical direction.

4.3. MAP-estimate

If we assume to have a prior information about the statistics of both the cover image and the watermark/noise, then we can use an MAP-estimator. The MAP-estimator is given by:

$$\hat{x} = \arg\max_{\tilde{x}\in\mathbb{R}^N} \left\{\ln p_w\left(y|\tilde{x}\right) + \ln p_x(\tilde{x})\right\}, \tag{6}$$

where $p_x(.)$ is the probability density function of the cover image.

It is not possible to find a closed form solution for the MAP-estimate in all cases. We model the image as stationary generalized Gaussian distribution and the watermark as Gaussian. The generalized Gaussian model is given by:

$$p_x(x) = \left(\frac{\gamma\eta(\gamma)}{2\Gamma(\frac{1}{\gamma})}\right)^{\frac{N}{2}} \cdot \frac{1}{|\det R_x|^{\frac{1}{2}}} \cdot \exp\{-\eta(\gamma)(|x-\bar{x}|^{\frac{\gamma}{2}})^T R_x^{-\frac{\gamma}{2}} |x-\bar{x}|^{\frac{\gamma}{2}}\}, \tag{7}$$

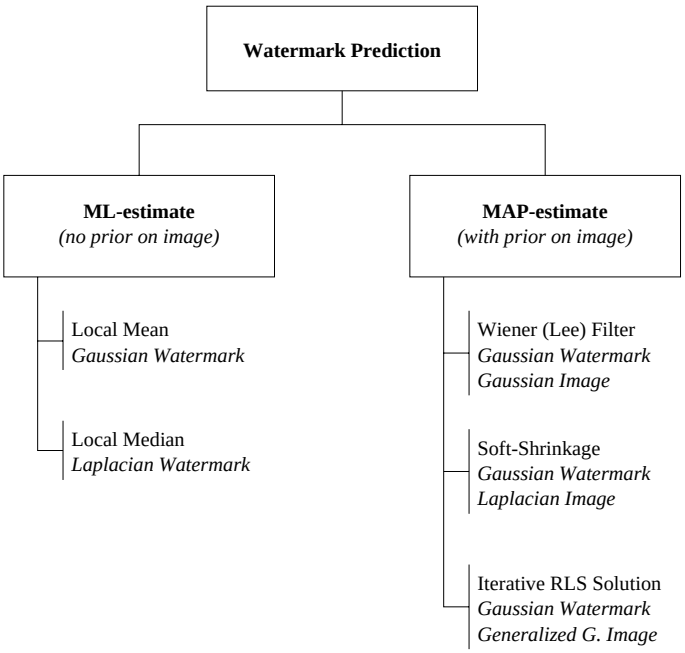


Figure 4. Principles for watermark prediction.

where $\eta(\gamma) = \sqrt{\frac{\Gamma(\frac{3}{\gamma})}{\Gamma(\frac{1}{\gamma})}}$ and $\Gamma(t) = \int_0^\infty e^{-u}u^{t-1}du$ is the gamma function. R_x is a diagonal autocovariance function of dimension N. For the stationary model the elements of R_x are equal and constant. γ is called the *shape parameter* of the generalized Gaussian distribution. For $\gamma = 2$ the generalized Gaussian distribution reduces to the normal Gaussian distribution and for $(\gamma = 1)$ it reduces to the Laplacian distribution. For real images the shape parameter is in the range $0.3 \leq \gamma \leq 1$.

For our approach, that is Gaussian watermark and stationary generalized Gaussian cover image, there exists no closed form solution for the MAP-estimate of the cover image. Therefore, we propose to reformulate it as a *reweighted least squares (RLS)* problem, which allows us to derive an iterative optimal solution.[14] Then equation (6) is reduced to the following minimization problem:

$$\hat{x}^{k+1} = \arg\min_{\tilde{x}\in\mathbb{R}^N} \{\frac{1}{2\sigma_n^2}\|y - \tilde{x}^k\|^2 + w^{k+1}\|r^k\|^2\}, \tag{8}$$

where

$$w^{k+1} = \frac{1}{r^k}\rho'(r^k), \tag{9}$$

$$r^k = \frac{x^k - \bar{x}^k}{\sigma_x^k}, \tag{10}$$

$$\rho'(r) = \gamma[\eta(\gamma)]^\gamma \frac{r}{|r|^{2-\gamma}}, \tag{11}$$

where k is the number of iterations, and γ is again the shape parameter of the generalized Gaussian distribution. In this case, the penalty function is quadratic for a fixed weighting function w.

Assuming w is constant for a particular iteration step, one can write the general RLS solution in the next form:

$$\hat{x} = \frac{w\sigma_n^2}{w\sigma_n^2 + \sigma_x^2}\bar{x} + \frac{\sigma_x^2}{w\sigma_n^2 + \sigma_x^2}y. \tag{12}$$

This solution is similar to the closed form Wiener filter solution.[9] The same RLS solution could also be rewritten in the form of Lee filter[9]:

$$\hat{x} = \bar{x} + \frac{\sigma_x^2}{w\sigma_n^2 + \sigma_x^2}(y - \bar{x}). \tag{13}$$

The principal difference with classical Wiener or Lee filters is the presence of the weighting function w. This weighting function depends on the underlying assumptions about the statistics of the cover image. In the rest of this paper, we will only consider the Lee version of the solution which coincides with the classical case of Gaussian prior (shape parameter $\gamma = 2$) of the cover image ($w = 1$). It is important to note that the shrinkage solution of image denoising problem previously used only in the wavelet domain can easily be obtained from Equation 8 in the next closed form:

$$\hat{x} = \bar{x} + \max(0, |y - \bar{x}| - T)sign(y - \bar{x}), \tag{14}$$

where $T = \frac{\sigma_n^2}{\sigma_x}\sqrt{2}$ is the threshold for practically important case of Laplacian image prior. This coincides with the soft-thresholding solution of the image denoising problem.[11]

5. COPY INSERTION

The predicted watermark from the previous section could directly be added to the target image. However, this solution is not optimal and would result in a variety of artifact in the target image. Therefore, before adding the predicted watermark to the target image we need to process it. As mentioned above, the goal of the processing is to adapt the copy of the watermark to the target image in order to keep it imperceptible while maximizing its energy. This process is actually the same as in a standard watermark embedding scheme. There are many ways to adapt the watermark to the target image, such as methods exploiting the contrast sensitivity and masking phenomena of the HVS.[7,12] In this work we decided to used the *noise visibility function (NVF)* proposed by Voloshynovskiy et al.[14] The noise visibility function is defined as:

$$NVF = \frac{\omega}{\omega + \sigma_x^2\theta}, \tag{15}$$

where σ_x^2 the cover image variance, and ω is a local weighting function defined by:

$$\omega = \gamma[\eta(\gamma)]^\gamma \frac{1}{|r|^{2-\gamma}}, \tag{16}$$

where $r = \frac{x-\bar{x}}{\sigma_x}$. θ is a tuning parameter defined by:

$$\theta = \frac{100}{\sigma_{max}^2} \tag{17}$$

There NVF can be used for two different models. In the first model we assume a non-stationary Gaussian cover image. In this case the NVF in Equation 15 reduces to:

$$NVF(i,j) = \frac{1}{1 + \sigma_x^2(i,j)\theta}, \tag{18}$$

where $\sigma_x^2(i,j)$ denotes the local variance of the image in a window centered at the coordinates (i,j). In the second model we assume a stationary generalized Gaussian cover image. In this case, the NVF defined in Equation 15 reduces to:

$$NVF(i,j) = \frac{\omega(i,j)}{\omega(i,j) + \sigma_x^2}, \tag{19}$$

where $\omega(i,j)$ is the local weighting factor, and σ_x^2 is constant. For the watermark copy attack we will work with the non-stationary Gaussian model, and therefore use Equation 18.

The NVF characterizes the local texture of the image and varies between 0 and 1, where it is 1 for flat areas and 0 for highly textured areas. It can therefore be used to describe the local texture masking phenomena for the watermark, that is in ares with high activity the watermark strength may be increased due to the masking effect. In addition to the masking effect, we will also take the contrast sensitivity into account and combine it with the noise

visibility function. The contrast sensitivity is described by the Weber-Fechner law, which says that the detection threshold of noise is approximately proportional to the local luminance. The final weight is then given as:

$$W = ((1 - NVF)\alpha + NVF(1 - \alpha))\, lum, \tag{20}$$

where $\alpha = (0, \ldots, 1)$ describes the relation between the watermark strength in textured areas and flat areas, and lum the local luminance. If we set $\alpha = 1$, the watermark will be concentrated on textured areas, that is edges and corners. If the set $\alpha = 0$, the watermark will mainly be put into flat areas.

The fake watermarked image is then generated by scaling the weighting function W, multiplying it by the sign of the predicted watermark, and then adding the result to the target image:

$$\tilde{t} = t + \beta W \text{sign}(\hat{w}), \tag{21}$$

where β is the overall watermark strength.

6. RESULTS

To test the proposed watermark copy attack we applied it to two publicly available watermarking tools for still images. We refer to these tools as software A and B. The watermarking technologies in the attacked tools have the following properties:

- **Software A**: spatial domain spread spectrum based watermarking with a frequency template for geometrical reference, 64 bits payload.
- **Software B**: spatial domain spread spectrum based watermarking scheme with weighting mask based on the human visual system, 64 bits payload.

The attack was tested on the two gray scale images *lena* and *cameraman*, both of size 256×256. For the watermark prediction we used a standard adaptive Wiener filter with a window size of 5×5. The tests were as follows. First, we signed the *lena* image using one of the mentioned watermarking tools. Then we copied the watermark from the watermarked *lena* image to the *cameraman* image and tried to extract the watermark form the counterfeit image. The top two images in Figure 6 show the watermarked images corresponding to the two watermarking tools A) and B). The second row of the same figure shows the predicted watermark from the stego images in the top row. It is interesting to note the different characteristics for the predicted watermarks. The last row of Figure 6 shows the target images after inserting a scaled version of the sign of the predicted watermark.

In both cases we were able to retrieve the watermark form the counterfeit images. We made additional test with different images and succeeded each time to copy the watermark from the original stego image to the target image.

7. DISCUSSION

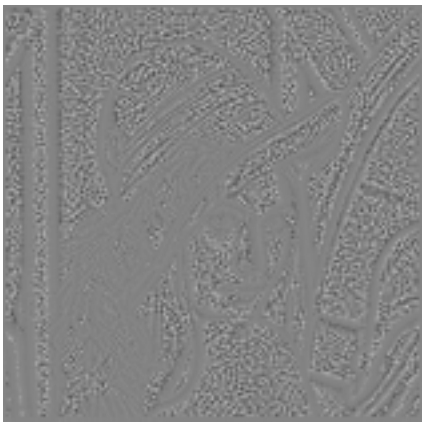
Independent of the application of the digital watermarking technology, the proposed attack puts in question the link between the cover data and the embedded information. Let us consider one specific application of digital watermarking and look at the weakness caused by the proposed attack. We consider the case where the watermark has an identification or authentication purpose in a passport picture. The idea of this application is to embed information about the owner of the passport, such as the name or social security number, into the image. The goal of the watermark is to uniquely link the image to the legal owner of the passport. If someone changes the image in a passport, the new image would not contain the required watermark and hence allow for the detection of the falsified document. However, if we can show that the watermarking technology used to protect the passport images does not resist the watermark copy attack, then the watermark is useless because anyone could easily replace the original image by a new image and copy the original watermark to the new image.

Similar scenarios are possible if the watermark is used for copyright protection or data monitoring. If a watermark does not resist the watermark copy attack, then a user can newer be sure if a detected watermark really belongs to the data under inspection. As a result, in many applications the use of a watermark would be useless.

(a) *Watermarked with Software A*

(b) *Watermarked with Software B*

(c) *Predicted watermark Software A*



(d) *Predicted watermark Software B*

(e) *Counterfeit image Software A*

(f) *Counterfeit image Software A*

Figure 5. Watermark copy attack applied to two watermarking tools.

8. CONCLUSIONS

The watermark copy attack is a new attack belonging to the group of protocol attacks. The concept of the attack consists in copying a watermark from a stego image to a target image without using any specific information about the watermarking technology. The attack consist of three main steps. In the first step we compute a prediction of the watermark in the stego data. Instead of directly predicting the embedded watermark, we propose to compute the prediction through a denoising process. In other word, the watermark prediction is computed by taking the difference between the stego image and a denoised version of the stego image. To perform the denoising we looked at ML-estimates and MAP-estimates of the cover image and propose closed form and iterative solutions for special cases of noise and cover image statistics. In the second step the predicted watermark is processed. The goal of the processing step is to maximize the energy of the watermark under the constraint of imperceptibility. To adapt the watermark to the target image we propose to use a noise visibility function. In the last step the processed prediction of the watermark is added target image.

The effectiveness of the attack was tested by applying it to two publicly available watermarking tools. We have shown that for both tools it was possible to copy the watermark from the stego image to the target image. This new attack has several important implications depending on the application of digital watermarking. If a technology does not resist to the copy attack, a user may not be sure if a detected watermark really belongs to the data under inspection. This is a big problem for may applications. New watermarking technologies should therefore take the watermark copy attack into account during the technology design process. We are current working on optimized versions of the attack for a variety of media. Furthermore, we are investigating solutions to make watermarking technologies resilient to the watermark copy attack.

REFERENCES

1. I.J. Cox and J.-P. Linnartz. Some general methods for tampering with watermarks. *IEEE Journal on Selected Areas of Communications (JSAC)*, 1997.
2. I.J. Cox, J.-P. Linnartz, and T. Shamoon. Public watermarks and resistance to tampering. In *Proceedings of the International Conference on Image Processing (ICIP)*, 1997.
3. Frank Hartung and Martin Kutter. Multimedia watermarking techniques. *Proceedings IEEE: Special Issue on Identification and Protection of Multimedia Information*, 87(7):1079–1107, July 1999.
4. Carl W. Helstrom. *Probability and Stochastic Processes for Engineers.* Macmillan, 1991.
5. Markus G. Kuhn and Fabien A. P. Petitcolas. StirMark. `<http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/>`, November 1997.
6. M. Kutter and F. Petitcolas. A fair benchmark for image watermarking systems. In Ping Wah Wong and Edward J. Delp, editors, *Proceedings of the SPIE, Security and Watermarking of Multimedia Contents*, volume 3657, pages 226–239, San Jose, CA, USA, January 1999. IS&T, The Society for Imaging Science and Technology and SPIE, The International Society for Optical Engineering, SPIE.
7. Martin Kutter. *Digital Image Watermarking: Hiding Information in Images.* PhD thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1999.
8. G. Langelaar, R. Lagendijk, and J. Biemond. Removing spatial spread spectrum watermarks by non-linear filtering. In *Proceedings EUSIPCO98*, volume 4, pages 2281–2284, 1998.
9. Jae S. Lim. *Two-Dimensional Signal and Image Processing.* Prentice Hall, 1990.
10. Adrian Perrig. A copyright protection environment for digital images. Diploma dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, February 1997.
11. P.Moulin and J. Liu. Analysis of multiresolution image denoising schemes using generalized-gaussian priors. In *Proc. IEEE Sig. Proc. Symp. on Time-Frequency and Time-Scale Analysis*, Pittsburgh, PA, USA, October 1998.
12. Christine I. Podilchuk and Wenjun Zeng. Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas of Communications (JSAC)*, 16(4):525–539, May 1998.
13. Unknown. UnZign watermark removal software. http://altern.org/watermark/, July 1997.
14. Sviatoslav Voloshynovskiy, Alexander Herrigel, Nazanin Baumgrtner, and Thierry Pun. A stochastic approach to content adaptive digital image watermarking. In *Proceeding of International Workshop on Information hiding*, Dresden, Germany, September 1999.

15. Sviatoslav Voloshynovskiy, Shelby Pereira, Alexander Herrigel, Nazanin Baumgrtner, and Thierry Pun. Generalized watermarking attack based on watermark estimation and perceptual remodulation. In *Proceedings of SPIE: Security and Watermarking of Multimedia Content II*, San Jose, CA, USA, January 2000. SPIE.
16. S. Craver N. Memon B.-L. Yeo and M. Yeung. Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks, and implications. *IEEE Journal on Selected Areas in Communications (Special issue on Copyright and Privacy Protection)*, 16(4):573–586, May 1998.