Blur/Deblur attack against document protection systems based on digital watermarking

Sviatoslav Voloshynovskiy[1] , Alexander Herrigel[2] and Thierry Pun[1]

Computer Science Department,
University of Geneva,
24 rue General Dufour,
CH 1211, Geneva 4, Switzerland,

Digital Copyright Technologies,
Rte de la Chocolatiere 21,
CH-1026 Echandens, Switzerland,
{svolos,Thierry.Pun}@cui.unige.ch,alexander.herrigel@dct-group.com
WWW home pages: http://vision.unige.ch, http://www.dct-group.com

Abstract. A growing concern emerges regarding the possibility of counterfeiting currencies using digital imaging technologies. In order to help developing resistance against this new type of fraud, this paper presents possible attacking scenarios against supposedly sophisticated document protection systems based on digital watermarking. These new attacks, which would allow even an average counterfeiter to reproduce banknotes or passports created using systems with built-in watermark detector.

1 Introduction

Image/video acquisition and reproduction systems afford virtually unprecedented opportunities for the forgery and counterfeiting of documents, ID cards, passports, banknotes and other valuable documents. The low price and simultaneously high quality of modern scanners, printers, image editors, as well as the computational power of current computers, offer a powerful basis for average or even inexperienced counterfeiters to easily create high-quality reproductions of the above documents. Moreover, the fast distribution of information, technologies and software via Internet presents real unconstrained opportunities for counterfeiters to access and distribute such knowledge.

To fight such counterfeiting, a number of world leading companies and universities propose to use digital watermarking as a possible solution [1, 4, 12, 11]. The latest proposal [10] aims at creating a complete security architecture to prevent: input/output of valuable documents in/out of computers, as well as editing and further distribution or usage of the faked documents. The main idea of this proposal is to integrate the watermark detection in every piece of multimedia hardware. Any attempt to use a scanner, photo- or web cameras to digitize a valuable document, or a printer to reproduce copies is to be immediately indicated on the display with the operating system refusing to continue the process.

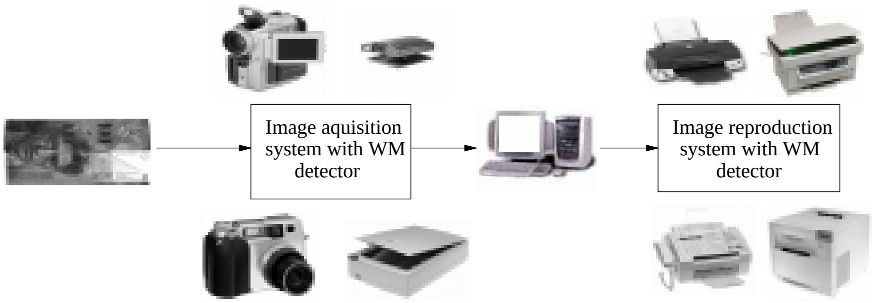


Image aquisition system with WM detector
Image reproduction system with WM detector

Fig. 1. Generalized block-diagram of a document protection system based on digital watermarking technology.

A further possibility is to send a message informing about such attempt to a specialized organization keeping track of forgeries. ID or personalized document information, or information about hardware seller or consumer can be used as a watermark. The generalized block-diagram of this approach can be depicted as in Figure 1.

This scenario is very attractive, but presents many security threats at different levels of the system architecture. We will briefly consider the main ones:

1. The system architecture does not support any future enhancement of the digital watermarking technology. This practically means that the watermarking system, being once accepted, should be compatible in the future with all possible modifications. This is not very likely in such a dynamically developing field as digital watermarking is.

2. To be exploited worldwide, the system assumes key management based on a public scheme that advocates the usage of the publicly known key for watermark embedding. This does not achieve an adequate level of security. Potentially, everybody or at least all digital imaging systems manufacturers will be able to utilize the watermarking key to counterfeit some dedicated documents or currencies.

3. The proposed system would become a de-facto worldwide standard for the worldwide digital industry. It is necessary to note, that no certification of watermarking technologies exist at the moment. Projects aiming at such certification, such as the European project Certimark, are just starting their activity [9].

4. The practical introduction of the system makes sense only under the condition of a global worldwide agreement between all manufacturers and countries to utilize the same technology. Otherwise, the consumers will have always a choice between systems with or without some functional constraints (the same story is happenning with the DVD industry). Additionally, without such an counterfeiters will always be able to order imaging systems from countries or companies which does not joint this agreement.

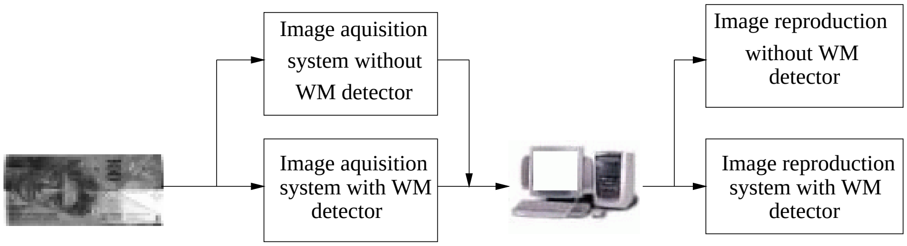


Image aquisition system without WM detector
Image aquisition system with WM detector
Image reproduction without WM detector
Image reproduction system with WM detector

Fig. 2. The generalized block-diagram of replacement attack.

5. Counterfeiters with a sufficient amount of technological and financial resources will always be able reproduce the simplest versions of the imaging devices themselves.

Therefore, in practice, an additional scenario that was not foreseen in proposal is possible. The counterfeiters can utilize the so-called "replacement attack". The main idea of the replacement attack, according to Figure 2, consists in the usage of devices without watermark detectors or of documents without watermarks. Obviously, someone can buy imaging devices today and keep them in the future. The same is true for banknotes. One can keep old banknotes without watermarks and reproduce them later even on an equipment that contains a built-in watermark detection.

The problem is even more complicated by the fact that a large number of image processing attacks against digital watermarking exist already which can be easily utilized by the attackers and by the counterfeiters [13].

We may further consider an even easier attacking scenario that does not require any auxiliary equipment like in the replacement attacks. The analysis we propose in the text that follows consists of a set of possible attacks that exploit the weak points of modern digital watermarking and the specifics of the proposed system architecture. The rest of the paper is structured as follows. In section 2 we describe the general concept of blur/deblur attacks whose goal are to prevent watermark detection. Section 3 discusses the restoration algorithm which would permit an attacker t enhance images after the blur attack. In Section 4, we present possible attacking scenarios against digitized documents containig watermarks. Section 5 introduces split/merge attack. Finally, section 6 presents the results of attacks and section 7 contains our conclusions.

2 Blur/Deblur attacks

In this section we describe the general concept of blur/deblur attacks. The block-diagram of this attacking scenario is depicted in Figure 3. The main idea of the proposed attack consists in the simultaneous exploitation of the weaknesses

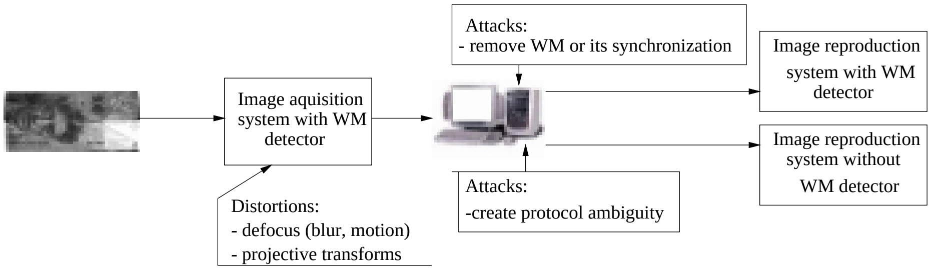

Attacks:
- remove WM or its synchronization
Image reproduction system with WM detector
Image aquisition system with WM detector
Image reproduction system without WM detector
Attacks:
-create protocol ambiguity
Distortions:
- defocus (blur, motion)
- projective transforms

Fig. 3. The blur/deblur attacking scenario.

of digital watermarking technologies and of the deficiencies of the considered architecture of the document protection system.

Consider the following scenario in details, starting from the image acquisition system. The basic assumption used for the design of the proposed secure architecture of the document protection system [10] is that a counterfeiter cannot easily digitize or take a picture of a protected document, due to the use of imaging devices with a built-in watermark detector. Therefore, the first line of attack consists in preventing successful detection of the watermark by such devices. To reach this goal, the counterfeiter can utilize some prior knowledge about the weaknesses of current watermarking technologies, available for example from the StirMark benchmarking home page of Fabien Petitcolas [8], from the European project Certimark [9], or more generally from any publication that deals with watermarking attacks or benchmarking. It is commonly known in the watermarking community that for example random bending distortions or image smoothing are still considerable weak points for the majority of watermarking algorithms. Therefore, the counterfeiter can exploit these attacks to disable watermark detection.

As the practical examples of these attacks, one can use: defocusing or blurring changing the focus distance in the photo-, video- or web cameras; putting the documents on some distance from the scanner or copy machine working surface to create some defocusing; slightly mutually move document and imaging device during scanning or taking picture; putting documents in such plane with respect to the imaging device to create generally projective geometrical transform. The list of possible pre-distortions can be considerably extended depending on the particularities of the imaging technology. We will refer to any possible distortion in this scenario as a blur. This operation can be generally modeled as:

$$y = Hx + n \tag{1}$$

$$x = s + w \tag{2}$$

where y is the blurred image at the output of an imaging device, x is a watermarked image or document, H is the blurring operator and n is the noise of the imaging system. The original image is s and a watermark created by some additive linear watermarking technology is denoted as w. The geometrical distortions can be modeled as global affine or projective transforms.

The only problem that could appear due to such blur attack lies in the degradation of the quality of the obtained image which is an important issue for the counterfeiter whose goal is to further exploit the faked document. Therefore, the quality of y should be as high as possible. To reach this objective, the attacker can utilize techniques that allow the inversion of the imaging equation (1); this is described in the next section.

3 Restoration of blurred image

The inversion of (1) is an ill-posed problem that requires to use either dedicated deterministic regularization or stochastic approach [2]. We will use here a stochastic approach based on a maximum a posteriori probability (MAP) estimator:

$$\widehat{x} = \arg\max_{\tilde{x}\in\Re^N} \{ln\ p_n\left(y \mid \widetilde{x}\right) + ln\ p_X\left(\widetilde{x}\right)\} \tag{3}$$

where $p_n\left(.\right)$ is the p.d.f. of the noise and $p_X\left(.\right)$ is the prior distribution of image. More generally estimators like penalized likelihood can be considered as well as minimum description length (MDL) estimator [6] could be also used. One can use sophisticated prior models of image with good edge-preserving properties like Markov Random Fields (MRF), Huber, Generalized Gaussian, Talvar or line model [3]. Since we are following the attacking scenario of the average attacker the resulting restoration algorithm should be either very simple to implement for example in Matlab, or its solution should easily be found on Internet. This motivates us to choose a simple non-stationary Gaussian model for the image $x \sim N\left(\overline{x}, R_x\right)$ with local mean $\overline{x}$ and covariance matrix R_x, and a Gaussian model for the noise $n \sim N\left(0, R_n\right)$. Assuming image and noise are conditionally i.i.d. one can determine:

$$\widehat{x} = (H^T {R_n}^{-1} H + C^T {R_x}^{-1} C)^{-1} H^T {R_n}^{-1} y \tag{4}$$

where T denotes transpose, and C represents a high-pass filtering (decomposition operator) and which can be also rewritten as $Cx = (I - A)x = x - Ax = x - \bar{x}$, where I is the unitary matrix, A is a low-pass filter used to compute the non-stationary local mean $\bar{x}$. The obtained solution corresponds to Wiener filter. The above maximization problem can be efficiently solved using the method of successive approximation [2], which yields the following iteration:

$$\widehat{x}^{k+1} = \widehat{x}^k + \beta[H^T y - (H^T H + \lambda C^T C)\widehat{x}^k] \tag{5}$$

where $\widehat{x}^k$ is the image estimate at iteration k and β is the relaxation parameter. To simplify the programming one can use a stationary assumption about the image prior that results in the Tikhonov regularization with constant regularization parameter $\lambda = \frac{\sigma_n^2}{\sigma_x^2}$, where $R_n = \sigma_n^2 I$ and $R_x = \sigma_x^2 I$. The iterative methods make possible to incorporate also a number of deterministic constraints into the solution in very simple manner. Therefore, for comparatively low cost of programming, the attacker can obtain quite powerful restoration technique.

4 Attacking scenarios

Once the image is restored from the blur the counterfeiter can use different attacking scenarios to reach the final goal, i.e. to create a faked document. We assume the simplest linear additive watermarking scheme (2). The theoretical analysis of the possible attacks against this scheme is reported by Voloshynovskiy *et al* [13]. Therefore, we will concentrate only on the most appropriate group of estimation-based attacks depending on the image reproduction system available for the counterfeiter (Figure 3).

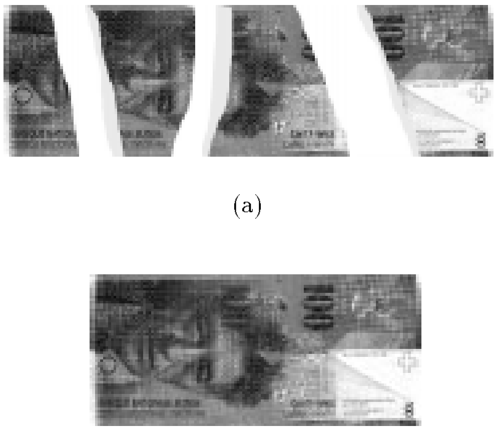
The first possible scenario assumes that the counterfeiter has only access to a reproduction system with built-in watermark detector. Therefore, the main goal of the attack should consist in the removal of the watermark or of the damaging of its synchronization without introduction of visible artifacts. The possible candidates for these attacks are: removal (denoising/lossy compression, denoising and perceptual remodulation), synchronization removal (random bending attack, template removal, projective transforms) [13].

If the counterfeiter has also access to a printing system without watermark detector, the spectrum of possible attacks can be considerably extended. The printing can be performed without any image modification. Secondly, to decrease public confidence, or even to damage the economy of other countries or to decrease their international reputation, the counterfeiters might be interested in creating public distrust in the currency or in other valuable documents. As the possible attacking scenario that perfectly fits this goal, the copy attack can be used [5]. Moreover, the counterfeiters can try to increase their personal interest based on the weaknesses of the watermarking protection system. For example, the watermark corresponding to larger banknote nominal could be embedded in smaller ones, if the bank machines are using watermarking for checking the denomination of the banknotes.

5 Split/merge attack

In this section we propose another new attack that we call the split/merge attack. A split/merge attacking game can be considered in the framework of the above attacks. However, since it can be used independently we consider it in more details. The split/merge attack is in its spirit similar to the mosaic attack proposed for Internet cracking of digital watermarking technologies used for copyright protection [7]. The same basic idea can be used for the counterfeiting of valuable documents on two levels. First, at image digitization the attacker shows/visualizes only a part of the document that is not small enough for the watermark detector in the imaging device to fail. The rest of the document is shadowed or cut on some orthogonal cells or pieces. At the second stage, another part of document is shown up to the imaging device in an amount that prevents successful watermark detection. This operation is repeated until the whole document is digitized.

Secondly, the printing of a watermarked document can be accomplished even using reproduction devices equipped with a watermark detector. The printing process is straightforward: the counterfeiter prints the whole document piece by piece on the same paper. This process could be also applied for the printing of the documents after copy attack on the equipment containing a watermark detector.



(a)

(b)

Fig. 4. An example of image that (a) can be presented to an imaging system according to the split/merge attack and (b) the final image printed in 4 stages using all croped pieces (b).

6 Test Results

We use the described attacks to test the possibility of counterfeiting an entire document protection architecture with some "virtual watermarking technologies". This means that these technologies are to our knowledge not directly exploited in some currency protection device, but since they are publicly available and represent the state-of-the art in copyright protection applications, they allow to evaluate the future tactics of counterfeiters. Moreover, it is not very likely that some more robust technologies will appear quickly than those that are already patented and implemented in the commercial watermarking tools. As an example, we used Digimarc, SysCop and IkonaMark methods to test the proposed attacks.

The results for the proposed blur/deblur attack are shown in Figure 5. The test image of a 100 Swiss frank banknote Figure 5(a) was watermarked using the PhotoShop version of Digimark, Figure 5(b). The image was printed and scanned; the watermark was successfully detected. The blurring was then applied, resulting in the image shown in Figure 5(c). The attempt to detect the watermark failed. The resulting image after restoration is shown in Figure 5(d). The image is of sufficient quality to be used for further counterfeiting and the watermark is successfully detected. The attacks described in Section 4 can be applied depending on the final goal of counterfeiter.

We performed the simulation of the watermarking system proposed in the patent [10] to show the efficiency of the proposed attack even against systems that are not publicly available as software. The above pattern in Figure 6 is referred to the fine art watermark modulation that uses changing of the line width and density. The Figure 6 shows the results of the applied blur/deblur attack for this type of watermarking systems. The performed modeling clearly indicates that the proposed attack is efficient against simulated system proposed in [10].

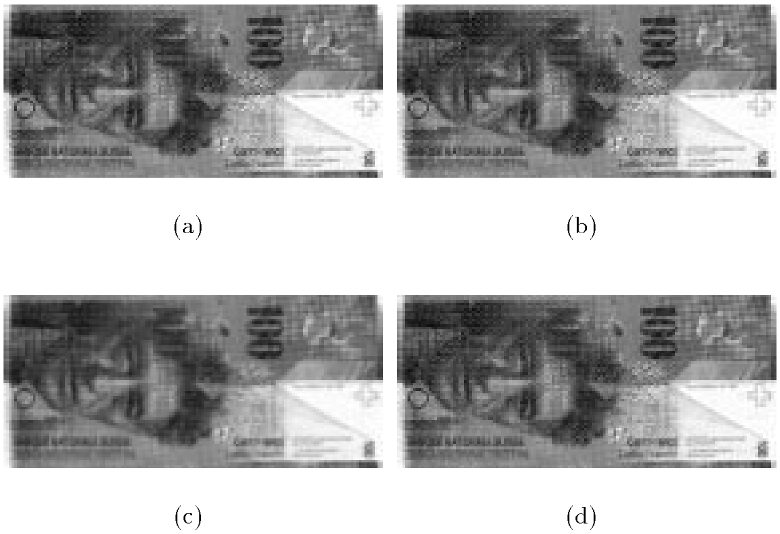
The split/merge attack was also successfully tested against the Digimark and SysCop algorithms, by splitting the image into 6 parts. Both systems were unable to detect the watermark from the small pieces.

7 Conclusions

In this article we have considered possible attacks against the recently proposed concept of exploiting digital watermarking as a tool against counterfeiting and forge of valuable documents. The critical analysis performed clearly shows that even an average counterfeiter could easily overcome all security measures and forger the documents supposed to be protected by this system. Moreover, the protocol attacks scenarios discussed here show that even introducing new extremely robust watermarking algorithms seems to be questionable in view of improving the security level of the considered architecture. Deeper investigations should be carefully performed before considering this system as the working prototype on the world-wide level.

ACKNOWLEDGMENTS

We thank Shelby Pereira and Frederic Deguillaume for their valuable insights. This work is partly supported by European project CERTIMARK.



(a) (b)

(c) (d)

Fig. 5. Results of testing: (a) original image of Swiss banknote, (b) banknote with Digimark watermark embedded from PhotoShop with the maximum durability 4, (c) image after defocusing (watermark is not detected), (d) restored image (watermark is successfully detected).

References

1. S. Carr B. Perry and P. Patterson. Digital watermarks as a security feature for identity documents. In R. van Renesse and W. Vliegenthart, editors, *SPIE's 12th Annual Symposium, Electronic Imaging 2000: Optical Seciruty and Counterfeit Deterrence Techniques III*, volume 3973 of *SPIE Proceedings*, pages 80–87, San Jose, California USA, 27–28 January 2000.
2. A. Katsaggelos ed. *Digital image restoration.* Springer Verlag, 1991.
3. D. Geman and S. Geman. Stochastic relaxation, gibbs distributions and the bayesian restorations of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(6):367–383, 1984.
4. A. Herrigel, S. Voloshynovskiy, and Z. Hrytskiv. Optical/digital identification/verification system based on digital watermarking technology. In *SPIE International Workshop on Optoelectronic and Hybrid Optical/Digital Systems for Image/Signal Processing ODS'99*, SPIE Proceedings, Lviv, Ukraine, 20–24 sep 1999.
5. M. Kutter, S. Voloshynovskiy, and A. Herrigel. Watermark copy attack. In Ping Wah Wong and Edward J. Delp, editors, *IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II*, volume 3971 of *SPIE Proceedings*, San Jose, California USA, 23–28 jan 2000.
6. J. Liu and P. Moulin. Complexity-regularized image denoising. In *Proc. of 4th IEEE International Conference on Image Processing ICIP97*, pages 370–373, Santa-Barbara, CA, 1997.

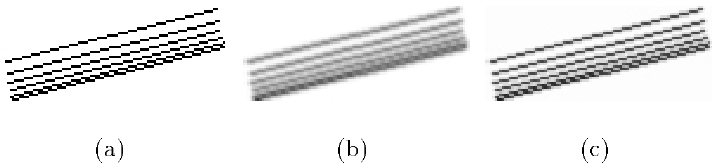


Fig. 6. Results of fine art modulation watermarking: (a) original image fragment of simulated watermark from Swiss banknote, (b) blurred pattern, (c) pattern after restoration.

7. F. A. P. Petitcolas and R. J. Anderson. Attacks on copyright marking systems. In *2nd International Information Hiding Workshop*, pages 219–239, Portland, Oregon, USA, April 1998.
8. Fabien Petitcolas. http://www.cl.cam.ac.uk/ fapp2/watermarking/.
9. European project Certimark. http://www.certimark.org/.
10. G. Rhoads. Digital watermarking and banknotes. *European patent application # 0961239A2*, 1999.
11. A. Jaffe S. Church, R. Fuller and L. Pagano. Counterfeit detterrence and digital imaging technology. In R. van Renesse and W. Vliegenthart, editors, *SPIE's 12th Annual Symposium, Electronic Imaging 2000: Optical Seciruty and Counterfeit Deterrence Techniques III*, volume 3973 of *SPIE Proceedings*, pages 37–46, San Jose, California USA, 27–28 January 2000.
12. S. Spannenburg. Developments in digital document security. In R. van Renesse and W. Vliegenthart, editors, *SPIE's 12th Annual Symposium, Electronic Imaging 2000: Optical Seciruty and Counterfeit Deterrence Techniques III*, volume 3973 of *SPIE Proceedings*, pages 88–98, San Jose, California USA, 27–28 January 2000.
13. S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun. Attack modelling: Towards a second generation watermarking benchmark. *Signal Processing*, accepted, Jauary 2001.