The Watermark Template Attack

Alexander Herrigel[#], Sviatoslav Voloshynovskiy[*#], and Yuriy Rytsar[#]

[#] DCT - Digital Copyright Technologies
Research & Technology
Stauffacher-Strasse 149
CH-8004 Zurich
Switzerland

[*]CUI - University of Geneva
Department of Computer Science
24 rue Général Dufor,
CH-1211 Geneva 4
Switzerland

ABSTRACT

This paper presents a new attack, called the watermark template attack, for watermarked images. In contrast to the Stirmark benchmark [1-2], this attack does not severely reduce the quality of the image. This attack maintains, therefore, the commercial value of the watermarked image. In contrast to previous approaches [3-4], it is not the aim of the attack to change the statistics of embedded watermarks fooling the detection process but to utilize specific concepts that have been recently developed for more robust watermarking schemes. The attack estimates the corresponding template points in the FFT domain and then removes them using local interpolation. We demonstrate the effectiveness of the attack showing different test cases that have been watermarked with commercial available watermark products. The approach presented is not limited to the FFT domain. Other transformation domains may be also exploited by very similar variants of the described attack.

Keywords: Digital watermark, template, attack, robustness.

INTRODUCTION

Copyright infringements of digital images can be detected by digital watermarks, embedded by special software programs. Visible and invisible watermarks may be applied for copyright protection. Visible watermarks direct the observer to the fact that the image is copyright protected. The invisible watermark utilizes the inability of the human vision system to perceive small differences in optical data. These slight differences are exploited by special software programs for embedding copyright information directly into the images. Invisible and robust watermarks have the advantage that they can not be identified and destroyed easily if the watermark process is robust against specific image transformations such as lossy compression, change of contrast, vector quantization, rotation, scaling, translation, cropping, change of aspect ratio, color editing, and morphing. A lot of new approaches [14, 15] have been proposed by academia and industry to derive and to develop watermarking systems that are robust and satisfy the requirement to recocer the watermark after any affine transformations.Along this development, different schemes have been proposed [6-8] that are based on so-called templates. These templates are a synchronization pattern that enables the estimation of the applied affine transformation. Based on this estimation, the inverse transformation can be applied and then the watermark estimated and recovered. The following two figures illustrate such templates applied for different types of watermarking technologies [6,7].

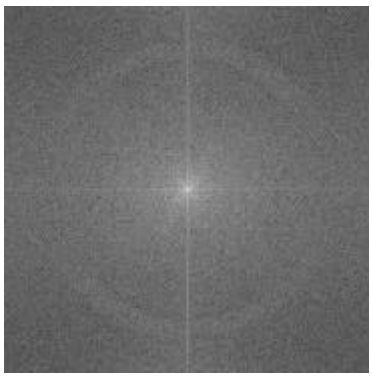


Figure 1: A FFT based watermarking technique with a synchronization template.

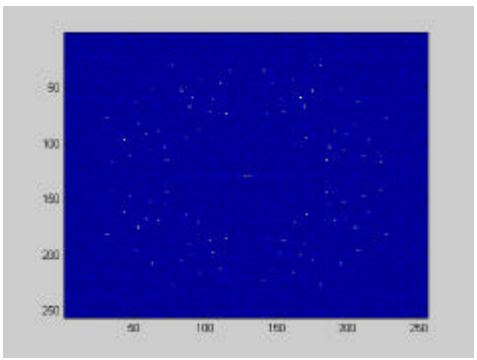


Figure 2: A FFT based template for a spatial domain based watermarking technique.

The results achieved by utilizing synchronization patterns with respect to performance and the robustness against affine transformation are convincing. Three schemes [6-8], developed between 1996 and 1999, had a leading score rating in the Stirmark benchmark. Although a lot of progress has been made for more robust schemes utilizing these templates, a system security aspect has been totally ignored. If these templates have characteristic features that are independent from the specific image, then these charcatseristics can be exploited to derive a new attack that has severe consequences to the performance and robustness of the different watermark technologies. This attack, called the watermark template attack, identifies the specific features of a template and partly removes the specific template to fool the synchronization process applied during the detection process. The attack presented in this paper requires no key knowledge.

TEMPLATE USAGE

The template contains no information but is merely a tool or synchronization pattern applied to recover possible affine transformations of the watermarked image. The watermark detection process is based on a two phases. First we determine the affine transformation (if any) undergone by the watermarked image, then we invert the transformation and decode the watermark.

Embedding the Template: The template applied during the marking process is very specific to the watermarking technology. Depending on features of the specific watermark technology, there are different strategies for the template generation. We don't need to know how the specific template in a domain is constructed from the perspective of a counterfeiter, since the template applied will always generate some peaks in the traget domain used for the template. Some approaches [7] use a template of approximately 25 points. The points of the template are uniformly distributed in the DFT domain with the low frequencies being excluded. The points are chosen pseudo-randomly based on a secret key. The low frequencies are excluded since they contain the bulk of the spectral power and represent noise during the decoding process. The strength of the template is determined adaptively as well. A good visible quality and performance can be achieved if we embed insertpoints at a strength equal to the local average value of DFT points plus one standard deviation. Points in the high frequencies are inserted less strongly since in these regions the average value of the high frequencies is usually lower than the average value of the low frequencies.

Detecting the Template: Some approaches [7] transform the template matching problem into a point-matching problem. After this problem has been solved, the top candidates for the template points are identified. If an affine transformation has been applied, the identified template poiints will differ from the orginal ones. This change is exploited to estimate the applied affine transformation. The corresponding inverse affine transformation is then applied for a better synchronization of the watermark.

WATERMARKING ATTACKS

Following the attack classification [5, 9] , we distinguish between four different classes of attacks: Removal attacks, geometrical attacks, cryptographic attacks, and protocol attacks.

Removal attacks remove a watermark from the watermarked data. These approaches consider the inserted watermark as noise with a given statistics and estimate the original, non watermarked data from the watermarked data. Based on a maximum a posteriori watermark estimation a remodulation of the watermark is applied such that a modified distribution is generated that has the least favorable noise distribution for the watermark detector.
Geometrical attacks do not to remove the embedded watermark, but distort by alterations the watermarked data. The attacks are usually applied such that the watermark detector looses synchronization with the embedded information. The watermark is still in the data but the verification process is fooled since it can't resynchronize the data given with the embedded watermark.

Cryptographic attacks do not exploit the very specific embedding or detection process. Based on exhaustive key search approaches, they try to guess the key applied for the embedding. If such a key has been found, the watermark can be overwritten. Other approaches try to fool the generation of the random vectors to predict the positions of the modufied pixels.

The protocol attacks generate protocol ambigiuties in the watermark process. Typically examples are the copy attack [10] and the approach to find a inverse mapping from the watermarked data [13]. The purpose of this attack is to generate ambigious original data that does not correspond to the same orginal data the copyright owner had to generate the watermarked data.

The template attack partly destroyes the snychonization pattern of the watermarked data. If an affine transformation has been applied on the protected data, this transformation can't be recovered. The template attack, belongs, therefore, to the class of the geometrical attacks.

THE WATERMARK TEMPLATE ATTACK

The scheme of the attack can be easily derived. Our main goal is to destroy without any key knowledge the synchronization pattern of the watermark such that we can fool later the detection process after an affine transformation of the image has been applied. The attack has the following phases:

Phase 1	• Read the watermarked image and apply a median filter (Wiener filter is also adequate) to estimate the original image. • Subtract the estimated orginal image from the watermarked image and store the result.
Phase 2:	• Compute the discrete Fourier transform. • Identify the maximum peaks. • Modify the amplitude of the maximum peaks by replacing the specific amplitude with the average amplitude value of the neigbours within a considered window (window size is a system parameter). • Compute the inverse discrete Fourier transform after the amplitude of the peaks have been modified and write the result in a new file. This file is the attacked image.

This attack utilizes latest results in the watermarking domain, since there are only a very few watermark schemes that have a robust detection for a 64 bit watermark against any affine transformations such as change of aspect ratio, scaling, or rotation. Different tests have shown that the attack may even break the detection process of some watermarking technologies if no affine transformation was applied. The following figures illustrate a simple example (technology from a US manufacturer, integrated in Photoshop Version 5.5, application domain: currency protection). Figure 3 shows the original image, Figure 4 shows the watermarked image (highest durability with verification), Figure 5 shows the template of the watermarked image, Figure 6 shows the watermarked image after the template attack, Figure 7 shows the template of the attacked image, Figure 8 shows the attacked image after the rotation and Figure 9 shows the failed detection message. The quality of the image after the attack can be improved applying some denoising or restoration techniques. For the target application domain, this is not necessary, since currency notes are exchanged in daily life with usage dependent quality factors. Currency protection by digital watermarks is a new application field that is investigated by different groups from academia and industry.

Figure 3: The orginal image.

Figure 4: The watermarked image (US currency).

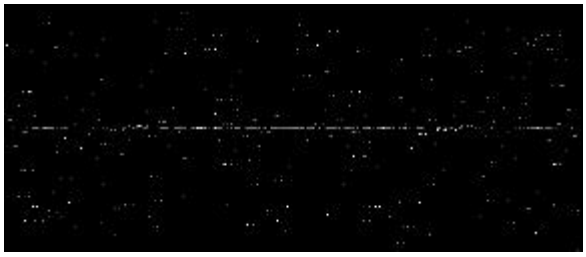


Figure 5: The template of the watermarked image (US currency).

Figure 6: The watermarked image after the template attack.

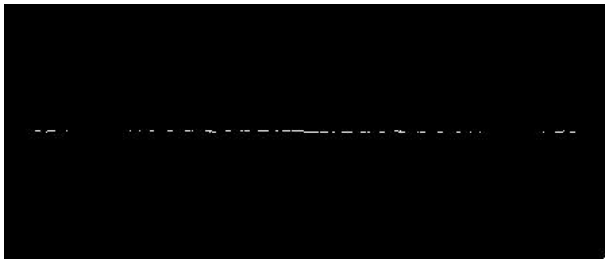


Figure 7: The template of the watermarked image after the template attack.

Figure 8: The rotated image after the attack.

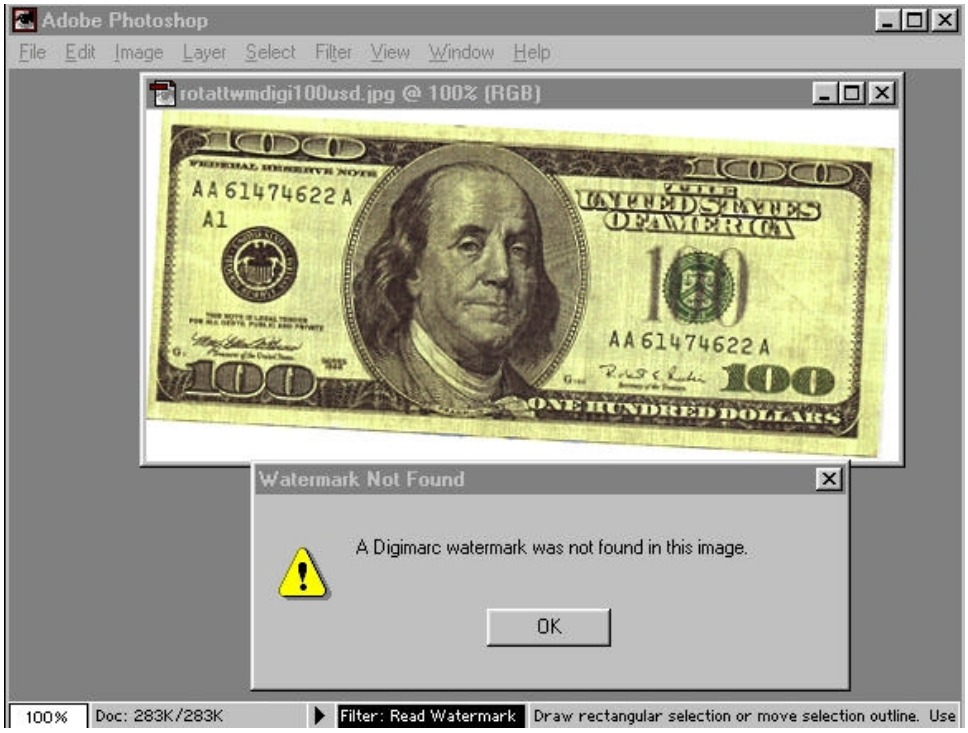


Figure 9: Screen capture of the failed detection.

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Template Removal Implementation
Attack test program for watermark technologies
Gray & color images.
Version 4.0, September 2000
Copyright (C) 2000, DCT, Zurich, Switzerland
All rights reserved
Legal relevant notice:
Authorized program execution only
for testing purposes.
Any other usage is not authorized
and prohibited.
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
*** Watermarked: (445 x 188), 83660 pixels ***
Template Removal.....50%....(515).....100%
Output File Name -----> removed_wmdigi100usd.bmp
Output File Name -----> test11.jpg
```

Figure 10: Screen capture of the attack program.

RESULTS

We have run numerous test cases to check how often the above described attack can be successfully applied. Based on the provided data from the Stirmark benchmark and the commercial market[1], we have tested the performance of the attack for a representative set of different image classes. After the tests we have noticed that there was not a single example the attack was not successful. The tests have been performed with different watermarking systems from academia and industry. Sometimes, the result of the attack was so severe that it was not necessary to apply an affine transformation in a post-processing step. For example, the watermark could not be detected for many images applying the commercial solution of an US manufacturer in Photoshop 5.0. Since the Stirmark benchmark images are quite known, we have listed in the appendix of the paper only the test results of a subset from our test cases. During the tests we have also realized that some manufactures have slightly changed the structure of the template from one product version to the other. This chnage of the template construction had no impact on our tests since the attack does not depend on the specific construction procedure of the template in the FFT domain.

CONCLUSIONS

The watermark template attack is a new attack against template based watermarking systems. The attack exploits parts of the specific algorithm concept of the watermarking technology and damages the applied template, i.e. the synchronization pattern, such that the detector is not able to identify the type of affine transformation applied. Since the template based approaches try to reduce the number of embedded information that constitute the template to avoid potential visible artifacts, the attack described in this paper is very efficient. This attack has a severe influence on the good rating in the Stirmark benchmark for the template based watermarking systems. If a technology does not resist to the template attack, a user in his role of the copyright owner has to face very similar threats and risks as with a different watermark technology that is less robust against affine transformations. We have shown by different test cases of different image classes that advanced and commercially available watermarking systems are not robust against this new attack.

We are currently working on self-referencing and repetitive watermarking schemes [10-12] that are resistant against the template attack and allow to recover the watermark from any affine transforms.

REFERENCES

1. Markus G. Kuhn and Fabien A. P. Petitcolas. StirMark. <http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/>, November 1997.
2. 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.
3. Sviatoslav Voloshynovskiy, Shelby Pereira, Alexander Herrigel, Nazanin Baumgärtner and Thierry 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, 23-28January 2000. (Paper EI 3971-34)
4. Sviatoslav Voloshynovskiy, Alexander Herrigel, Frédéric Jordan, Nazanin Baumgärtner and Thierry Pun, A noise removal attack for watermarked images, In J. Dittmann, K. Nahrstedt and P. Wohlmacher eds., *Multimedia and Security Workshop*, Orlando, Florida, USA, 30-31 October 1999. (at the 7th ACM Multimedia Conference (Multimedia 99))
5. Svyatoslav Voloshynovskiy, Shelby Pereira, Victor Iquise and Thierry Pun, Attack modelling: Towards a second generation benchmark, *Signal Processing, Special Issue: Information Theoretic Issues in Digital Watermarking*, 2001. V. Cappellini, M. Barni, F. Bartolini, Eds.
6. Digimarc Corporation, US patent 5,822,436, Photographic Products And Methods Employing Embedded Information.
7. Shelby Pereira, Joseph J. K. O Ruanaidh, Frederic Deguillaume, and Thierry Pun.Template Based Recovery of Fourier-Based Watermarks Using Log-polar and Log-log Maps, IEEE Int. Conf. on Multimedia Computing and Systems, ICMCS‘99 Florence, Italy, June 1999.

[1] Since the images applied in the Stirmark benchmark do not always share the same properties as commercial images, we have used in our tests also some professional images from Fratelli Alinary, which have been given to us for testing and benchmarking purposes. We would like to thank Fratelli Alinari, especially Mr. Andrea de Polo, for providing these test images. Fratelli Alinari is the copyright holder of these images. These images may not be used for any business purpose without the written permission of Fratelli Alinari, http://www.alinari.com.

8. Alessandro Piva, Department of Electronic Engineering, University of Florence, Italy: Improving DFT Watermarking robustness through optimum detection and synchronisation, Multimedia and Security Workshop at ACM Multimedia'99, GMD, Report 85
9. Martin Kutter, Sviatoslav Voloshynovskiy and Alexander 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. (Paper EI 3971-35)
10. M. Kutter, Watermarking resisting to translation, rotation and scaling, *Proc. of SPIE: Multimedia systems and applications*,Volume 3528, pp.423-431, Boston, USA, November, 1998.
11. Sviatoslav Voloshynovskiy, Frederic Deguillaume and Thierry Pun, Content adaptive watermarking based on a stochastic multiresolution image modeling, In *Tenth European Signal Processing Conference (EUSIPCO'2000)*, Tampere, Finland, September 5-8 2000.
12. PCT patent application, PCT/IB00/01089, Digital Copyright Technologies, Method for Adaptive Digital Watermarking Robust Against Geometric Transforms.
13. S. Craver, N. Memon, B. Yeo, and M. Young, Can invisible marks resolve rightful ownerships ? IS&T/SPIE Electronic Imaging'97: Storage and Retrieval of Image and Video Databases, 1997, pp. 310-321.
14. Special issue of Proc. IEEE, vol. 87, July, 1999.
15. Special issue of IEEE Signal Processing Magazine, September, 2000.

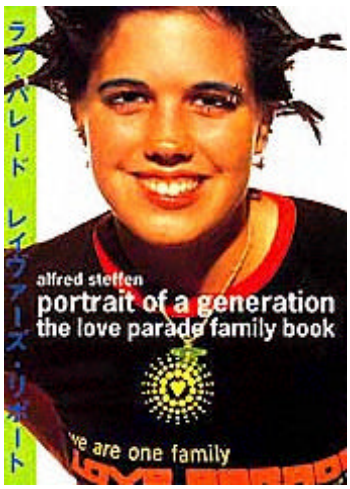
APPENDIX

Watermarked image

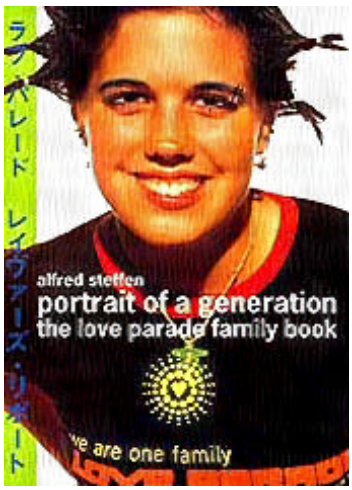
Watermarked image after the template removal

Watermarked image

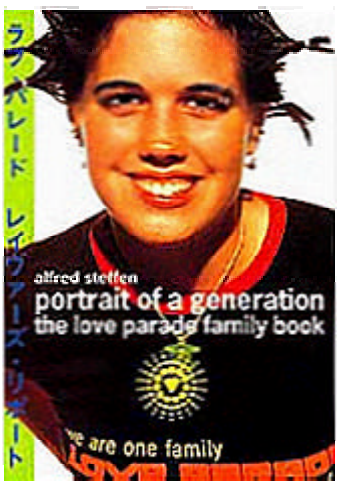
Watermarked image after the template removal



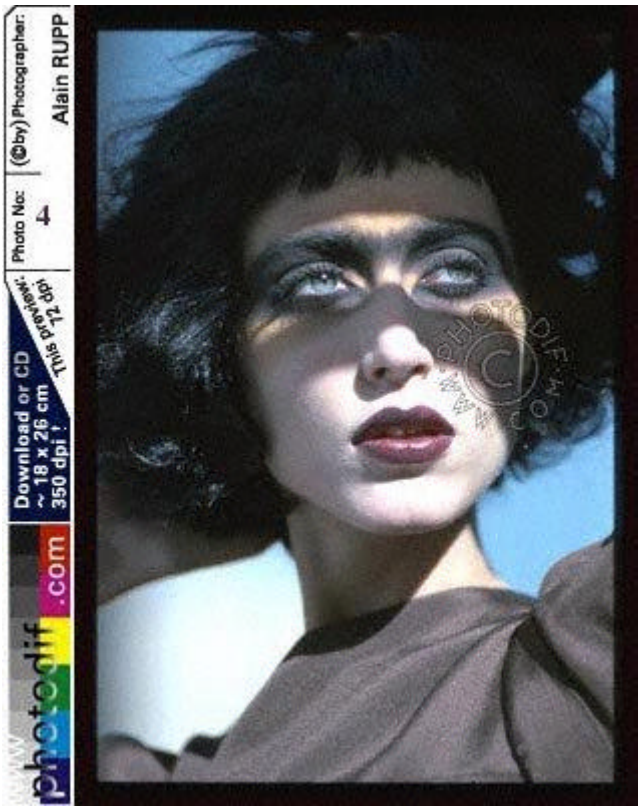
Watermarked image



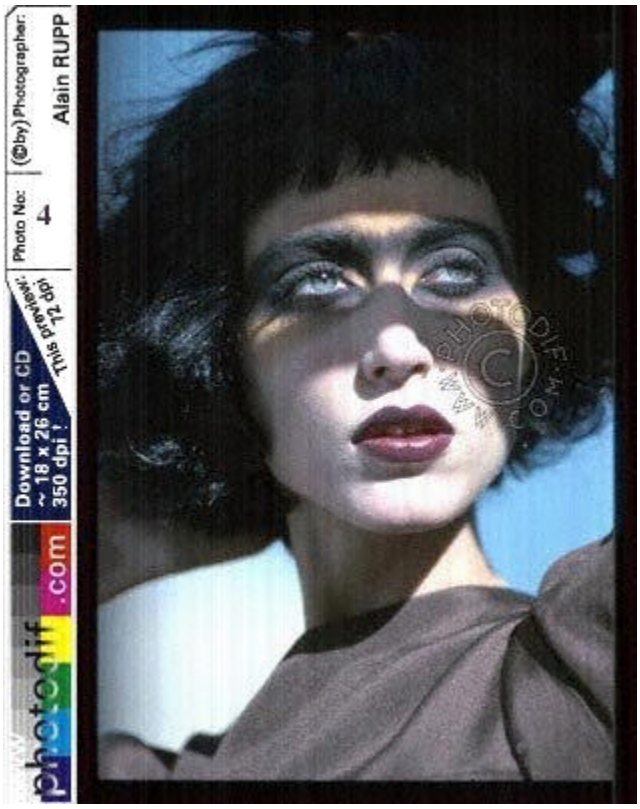
Watermarked image after the template removal



Attacked image after sharpen operation



Watermarked image



Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal

Watermarked image

Watermarked image after the template removal