Attacks on Digital Image Watermarks in the Discrete Wavelet Transform Domain

Andreja Samčović

***Abstract* - In the last few years, a large number of schemes have been proposed for hiding copyright marks and other information in digital images. Watermaking is a potential method for protection of ownership rights on digital images. This paper presents a number of attacks that enable the information hidden by them to be removed or otherwise rendered unusable.**

I. INTRODUCTION

Digital information is now readily available due to advances in the compression, storage and communication technologies. The amount of digital information that can be found in the Internet and the popularity of the Internet corroborate this observation. Unfortunately, the protection of this information, especially in circumstancies where the owners hope to generate revenue through controlled dissemination, is yet to be standardized. There are at least two consequences of the status quo, First, digital information that is already available is being illegally re-distributed and thereby robbing the legal owners of deserved revenue. Second, this situation discourages content generators from sharing their work with the wider community.

Digital media could be copied easy without loss of quality. Digital watermarking is an appropriate tool for protection of ownership. Digital image watermarking is a communication method in which additional information called watermark is embedded directly and imperceptibly into a digital image, also called original data or host data, to form watermarked data. Loosely analogous to watermarks in article documents, the embedded information is bound to the watermarked data wherever if goes. The embedded information should still be decodable from the watermarked data, even if the watermarked data is processed, copied, or redistributed. Potential applications of digital watermarking include copyright protection, distribution tracing, authentication and authorized access control [1].

Some types of image processing methods can be applied with the explicit goal of hindering watermark reception. In watermarking technology, an attack is any processing that may impair detection of the watermark. Attacks on digital watermarking schemes have two effects: either they reduce the effective channel capacity of fully disable the detection of the embedded watermark. Because it is not possible to enumerate all possible attacks, it is very difficult or even impossible to assess if given system is robust in the general case. What we should realize here is the fact that robustness requirements are application dependent. As a logical consequence, for a given application we first need to define the desired level of robustness and security and then to test against corresponding types of attacks. The danger is, of course, that some attacks may be forgotten or new attacks will emerge in the future. There is no way to avoid this kind of threat and it exists for all types of security related applications. What we should remember from the above comments is that attacks and their efficiencies are application dependent and that is only possible to guarantee robustness to attacks that are known at the time of application development [2].

Better understanding of the mechanisms of possible attacks will lead to the development of more efficient and robust watermarking techniques [3].

Andreja Samčović is with University of Belgrade, Faculty of Transport and Traffic Engineering, Vojvode Stepe 305, 11040 Belgrade, Serbia, E-mail: andrej@sf.bg.ac.rs

II. CLASSIFICATION OF WATERMARKING ATTACKS

Five different groups of attacks can be identified: removal attacks, geometrical attacks, cryptographic attacks, protocol attacks and other attacks, which is illustrated in Fig. 1.

2.1 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. This category of attacks includes denoising, quantization, remodulation, and averaging. Not all of these methods always come close to complete watermark removal, but they may damage the watermark information significantly.

The main idea of removal attacks consists of assuming that the watermark is additive noise relative to the original image. The removal attacks are those which further add noise to the watermarked image. This noise may have any of a number of different statistical distributions such as Gaussian or Laplacian. The removal attacks exploit the linear additive model in order to derive optimal estimators used for denoising and consequently removing of the watermark. In other cases both the removal attacks and the interference attacks can be combined such as in the denoising with perceptual remodulation attacks [4].

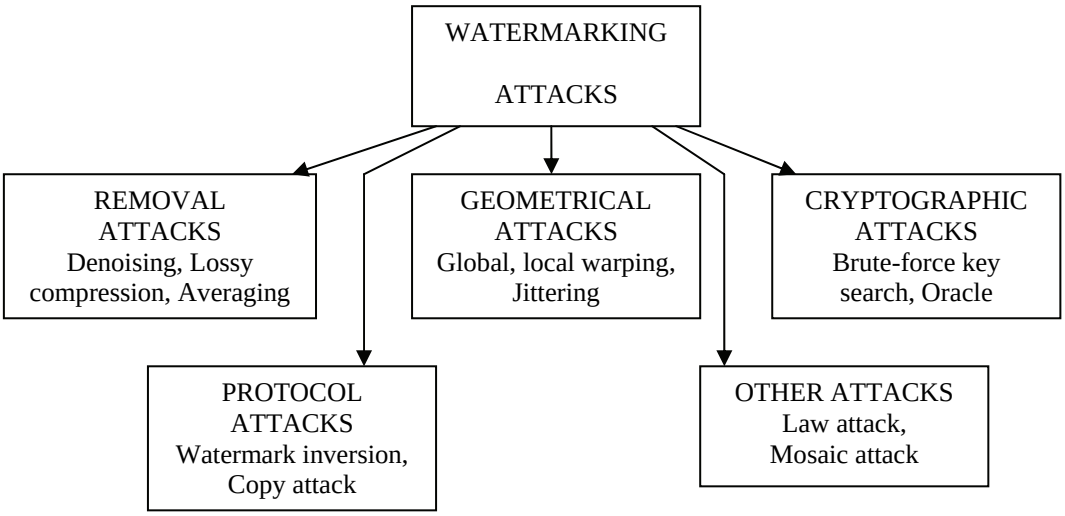


Fig.1. Classification of watermarking attacks

2.2 Geometrical Attacks

In contrast to the removal attacks, geometrical attacks intend to remove the embedded watermark itself, but to distort it through spatial alterations of the stego-data. The attacks are usually such that the watermark detector loses synchronization with the embedded information.

The most well known integrated software versions of these attacks are *Unzign* and *StirMark.* Unzign introduces local pixel jittering and is very efficient in attacking spatial domain watermarking schemes. StirMark introduces both global geometrical and local distortions. The global distortions are rotation, scaling, change of aspect ratio, translation and shearing that belong to the class of general affined transformations. The line / column removal and cropping / translation are also integrated in StirMark. Most recent watermarking methods survive after these attacks due to the usage of special synchronization technique. If robustness to global affined transformations is a solved problem, the local random alterations integrated in StirMark still remain an open problem almost for all techniques.

The so-called random bending attack exploits the fact that the Human Visual System (HVS) is not sensitive against shifts and local affined modifications. Therefore, pixels are locally shifted, scaled and rotated without significant visual distortions. The synchronization removal attacks belong also to this category. The synchronization consists of inserting peaks in the Discrete Fourier Domain (DFT). This method is called „template".

2.3 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 brute-force search for the embedded secret information. Practically, application of these attacks is restricted due to their high computational complexity.

Cryptographic attacks cover, for example, direct attacks to find the secret key or attacks called collusion attacks. Cryptographic attacks are very similar to the attacks used in cryptography. There are the brute-force attacks, which aim at finding secret information 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 in this category is so-called Oracle attack which can be used to create a non-watermarked image when a watermark detector device is available.

2.4 Protocol Attacks

Protocol attacks neither aim at destroying the embedded information nor at disabling the detection of the embedded information (deactivation of the watermark). Rather, they take advantage of semantic deficits of the watermark's implementation. The protocol attacks aim at attracting the concept of the watermarking application. The first protocol attack was proposed by Craver et al. [5]. They introduced the framework of invertible watermark and showed that for copyright protection applications watermarks need to be non-invertible. The idea of inversion consists of the fact that an attacker who has a copy of the stego-data can claim that the data contains also the attacker's watermark by subtracting his own watermark. This can create a situation of ambiguity with respect to the real ownership of the data. The requirement of non-invertability on the watermarking technology implies that it should not be possible to extract a watermark from non-watermarked image. As a solution to this problem, the authors proposed to make watermarks signal-dependent by using a one-way function.

Consequently, a watermark must not be invertible or to be copied. A copy attack, for example, would aim at copying a watermark from one image into another without knowledge of the secret key used for the watermark

embedding to create ambiguity with respect to the real ownership of data. It also belongs to the group of the protocol attacks. 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.

If the watermarking system or protocol makes not only the watermarked image, but also additional devices publicly available, the presence of such devices can be exploited. When exploiting the presence of a watermark detector, a test-image should be created near the detection boundary and then successively change single pixels until the detector response indicates that a particular pixel value has significant influence on the watermark. This way, a set of influential pixels can be determined which has the largest influence on the detector while introducing low disturbance into the image when manipulated. This process has linear complexity. With the presence of a watermark inserter, the difference image between the watermarked and the original image can be easily computed and analyzed. A public watermark inserter is provided by the *Digital Versatile Disc* (DVD) system for copy generation management.

2.5 Other Attacks

The Mosaic attack consists of chopping an image up into a number of smaller subimages, which are embedded in a suitable sequence in a web page. Common web browsers render juxtaposed subimages stuck together, so they appear identical to the original image. This attack appears to be quite general. All marking schemes require the marked image to have some minimal size. Thus, by splitting an image into sufficiently small pieces, the mark detector will be confused.

III. Discrete Wavelet Transform in Watermarking

The wavelet transform (WT) has been extensively studied in last decade. Many applications of the wavelet transform, such as compression, signal analysis and signal processing have been found. There are many good tutorial books and papers on this topic. Here, we just introduce the necessary concepts of the Discrete Wavelet Transform (DWT) for the purpose of this paper.

The basic idea of the DWT for a one-dimensional signal is the following. A signal is split into two parts, usually high and low frequencies. The edge components of the signal are largely confined in the high frequency part. The low frequency part is split again into two parts of high and low frequency. This process is continued until the signal has been entirely decomposed or stopped before by the application at hand. For compression and watermarking application, generally no more than five decomposition steps are computed. Furthermore, from the DWT coefficients, the original signal can be reconstructed.

The wavelet transform decomposes an image into three spatial directions, i.e. the horizontal **HL**, the vertical **LH** and the diagonal **HH**. At each level of decomposition, the magnitude of the DWT coefficients is larger in the lowest subbands ("approximation" **LL** subband), and smaller for other subbands ("detail" subbands: **HL**, **LH** and **HH**). The most significant coefficients in a subband are those with large magnitudes. The high resolution subbands help in locating the edge and texture patterns for an arbitrary image.

IV. Simulation Results

For the purpose of robustness testing the following set of ten standard test-images with the size of 512 x 512 pixels are used: *Barbara, Boat, Cameraman, Couple, Einstein, Elaine, F16, Goldhill, House* and *Lena*. The watermark is firstly converted into ASCII code and than encoded with the error correction code (ECC) in order to improve the robustness. Here, the robustness of the algorithm will be tested for the watermark sequence encoded with three different ECCs and for the watermark sequence that is directly embedded without using ECC. The following ECCs are used in order to determine which ECC performs the best from the robustness point of view:

(15,7) Bose-Chaudhuri-Hocquenghem (BCH) code,
(7,4) Hamming code, and
(15,7) Reed-Solomon (RS) code

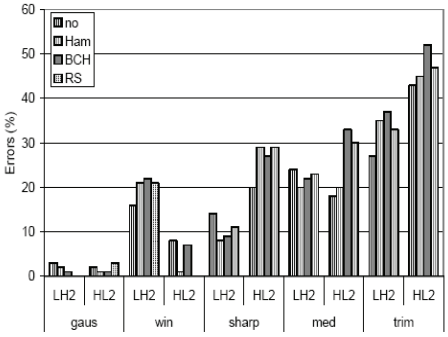
The same watermark is embedded in all detail subbands of the two-level DWT according to the embedding procedure. In order to fit our sequence to the codeword of the ECC for Hamming code, the 8-bit representation of the particular character will be used. For other ECC as well as for the directly embedded watermark sequence, the 7-bit representation will be used. The characteristic of the embedded watermark will be given in the Table I:

TABLE I
CHARACTERISTICS OF THE EMBEDDED WATERMARK

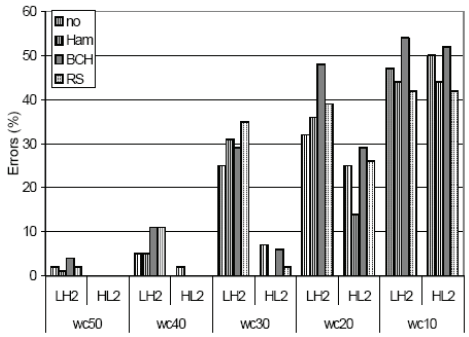
	MESSAGE LENGTH (bits)	ENCODED MESSAGE LENGTH (bits)	ADDITIONAL INFORMATION (bits per character)
NO ECC	147	147	7
BCH	147	315	7
HAMMING	168	294	8
RS	147	360	7

The Table I shows that with the Reed-Solomon coding more then twice of bits have to be embedded into the DWT subband compared to the approach without ECC. This fact must be taken into account when designing the watermark scheme due to the possible problem with the capacity of the cover image.

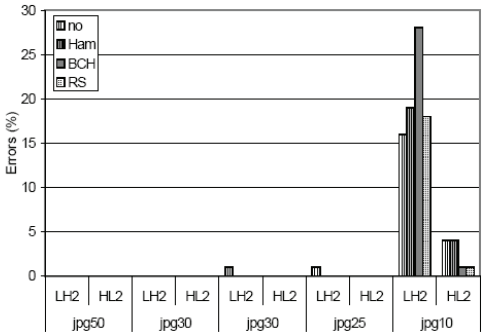
In the testing, several non-geometrical processing operations are applied watermarked test-images: median filtering with 3 x 3 window size (med), Gaussian filtering with 5 x 5 window (gaus), Wiener filtering with 5 x 5 window (wien), trimmed mean filtering with 7 x 7 window (trim), sharpening with 3 x 3 high-pass filter (sh), JPEG compression with different quality factors from 50 to 10 (jpg50, jpg40, jpg30, jpg25, jpg15, jpg10), as well as JPEG compression with different bit rates from 0,5 to 0,1 bits per pixel (bpp) (wc50, wc40, wc30, wc20 and wc10).



(a)



(b)



(c)

Fig. 2. Simulation results for: (a) different filtering attacks, (b) JPEG 2000, (c) JPEG compression attacks

The watermark is extracted separately from every subband in order to compare the robustness of the watermark embedded in that subband. The results for the *Lena* image are given in Fig. 2. The similar results are obtained for other test-images. All graphs in Fig. 2 present different attacks on the x-axes. The results are calculated as the total number of not correctly extracted watermark bits (errors) divided by the total number of watermark bits, expressed in percentage and presented on the y-axes of all three graphs. The best results are obtained for the watermark embedded in the subbands $\mathbf{HL_2}$ and $\mathbf{LH_2}$ and only results for these subbands are presented. The results for other tested subbands were not good and they were not being further considered. This was expected due to the fact that the common signal processing operations like filtering and compression will be most effective in the high frequencies (level 1 of the DWT decomposition).

From Fig. 2 it can be concluded that for the most attacks Reed-Solomon code gives less errors than other ECCs. It can also be concluded that the results strongly depend on the subband in which the watermark sequence was embedded. In some cases like trimmed mean filtering better results are obtained without using ECC.

V. Conclusion

Although the above classification makes it possible to have a clear separation between the different classes of attacks, it is necessary to note that very often a malicious attacker applies not only a single attack at the moment, but rather a combination of two or more attacks. The better understanding of possible attacks will lead to the development of more efficient and robust watermarking techniques.

References

[1] G.Voyatzis, I.Pitas, "The use of watermarks in the protection of digital multimedia products", *Proceedings of the IEEE*, Vol.87, No.6, pp 1197-1207, July 1999.

[2] B.Macq, J.Dittmann, E.Delp, "Benchmarking of image watermarking algorithms for digital rights management", *Proceedings of the IEEE*, Vol.92, No.6, pp 971-984, June 2004.

[3] Z.Bojković, J.Turán, A.Samčović, L.Ovsenik, "Coding, streaming and watermarking – some principles in multimedia signal processing", *Acta Electrotechnica et Informatica*, Vol.4, No.3, pp 13-20, 2004.

[4] F.Petitcolas, R.Anderson, M.Kuhn, "Attacks on copyright marking systems", 2nd Workshop on Information hiding, in *Vol. 1525 of Lecture Notes in Computer Science*, Portland, Oregon, USA, pp 218-238, 14-17. April 1998.

[5] S. Craver, S. Katzenbeisser, „Copyright Protection Protocols Based on Asymmetric Watermarking: The Ticket Concept", in *Communications and Multimedia Security Issues of the New Century*, Kluwer Academic Publishers, pp 159-170, 2001.