Watermarking Attack (BOWS contest)

Bennour J.[a], Dugelay J-L.[a] and Matta F.[a]

[a]Institut Eurecom
2229 route des Cretes, B.P.193
06904 Sophia-Antipolis, France
Email: bennour@eurecom.fr, dugelay@eurecom.fr, matta@eurecom.fr

ABSTRACT

In this paper, we describe some attack strategies we have applied to three different grayscale watermarked images in the particular context of BOWS (Break Our Watermarking System) Contest[1]; we also propose a possible use of BOWS as a teaching tool for master students.

Keywords: Watermarking, attack, BOWS.

1. INTRODUCTION

Many people and laboratories are working on solutions to be able to protect multimedia documents. A wide variety of watermarking algorithms have been proposed for this purpose. But today, the watermarking community needs some advanced attack techniques in order to compare the performances of the different proposed watermarking technologies. In this quite particular context we can cite the BOWS (Break Our Watermarking System) contest .
Three different grayscale watermarked images (see figure 1) are available for download on the "BOWS web site" and participants to BOWS are asked to erase the watermark from all the three images by using any action they want while granting the best PSNR between the watermarked image and the attacked one. For the first phase of the BOWS contest, the watermarking algorithm is unknown to the attackers.
In this paper, we describe the main attack techniques we have applied to the three watermarked images. We also propose a possible use of the BOWS contest as a teaching tool for Master students, and we conclude with some suggestions for a future edition.

2. TECHNIQUES TO BROKE THE WATERMARK

In this section, we describe the techniques used to erase the watermark while keeping a PSNR more than 30 dB. The first technique we have used exploits self-similarities of the image. This system was successful for the first image with no problem whereas for the two other images the watermark has been removed but with some difficulties to preserve the minimum required PSNR. Some simple image processing tools based on filtering, averaging, etc. were then used to increase the PSNR.

Figure 1. Watermarked images

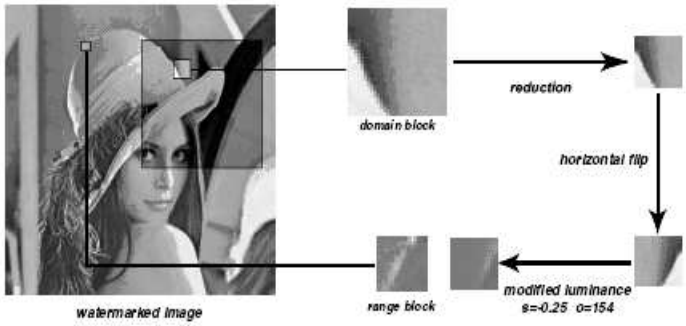

Figure 2. Self similarities process

2.1. Self similarities attack

We summarize here the principle and the main steps of the self similarity attack. More details can be found in.[2]

This attack is based on self-similarities of the image. Usually correlation between neighbor pixels is taken into account. With self similarities, it is the correlation between different parts of the image which is of interest. The basic idea of the attack consists in substituting some parts of the picture with some other parts of itself which are, or look, similar. The process is presented in figure 2. The objective is to approximate, to stir the watermark signal while keeping clear the cover signal.

In its basic version, the original image is scanned block by block. Those blocks are labeled range blocks (block R_i) and have a given dimension $n * n$. Each block R_i is then associated with another block D_i which look similar according to a Root Mean Square (RMS) metric.

The block D_i is labeled domain block and is searched in a codebook counting Q blocks Q_j. Those blocks may be blocks from the same image or from an external unwatermarked database. In practice, for a given range block R_i, a window is randomly selected in the image. The blocks belonging to this window provide the codebook. Each block Q_j is scaled if needed in order to match the dimensions of the range block R_i. A set of T_k geometrically transformed blocks $T_k(Q_j)$ is then built. For each transformed block $T_k(Q_j)$, the photometric scaling s and offset o is computed by minimizing the error between the transformed block $g = T_k(Q_j)$ and the range block $f = R_i$ by the Least Mean Square method.

Eventually, the transformed block $s.T_k(Q_j) + o$ which has the lowest RMS distance with the range block R_i is found and the corresponding block Q_j will be the domain block D_i associated with the range block R_i. Since the two blocks R_i and D_i look similar, we can substitute R_i with the transformed version of D_i. As a result, the image will be slightly modified but the watermark signal will be randomly spread through the image and the detector will be unable to retrieve it.

In[3] [4] improved versions of the self-similarity attack are presented. The basic idea is that several blocks can be combined to compute the replacement block instead of a single one. We can use PCA (Principal Component Analysis) on blocks to build the efficient codebook.

2.2. Other simple techniques

We describe here some simple techniques used to improve the PSNR of the images attacked using the self-similarity process described in section 2.1.

- Assuming we have an estimate of the watermark but we do not know with which strength we have to retrieve it we can use the following procedure in order to increase the PSNR. We denote I_o the original watermarked image, I_x the image with the watermark removed.

1. We compute the estimated watermark W_x by subtracting I_o from I_x

$$W_x = I_o - I_x. \quad (1)$$

2. We compute

$$I_x^{'} = I_o - \alpha.W_x. \quad (2)$$

α is tuned manually to the point where $I_x^{'}$ contains no watermark and has the highest possible PSNR. If $\alpha > 1$ the attack will be more important but the PSNR will decrease and if $\alpha < 1$ the attack will be less important (i.e. the watermark could be still detected) but the resulting PSNR will be higher (i.e. image of better quality). Eventually we can use a low pass filtering version of W.

- We improve some results by averaging several (successfully) attacked images. We obtain better results if the used methods to attack the image are different.
- We also improve our results by taking care to preserve the average gray values associated with subregions or blocks of images.

It is worth mentioning that we attacked the three images without any knowledge about the watermarking algorithm used (first phase of the BOWS Contest). We started to work on the BOWS contest during the last week of the competition. The obtained results are as follows:

Average	Image 1	Image 2	Image 3
33.65 dB	35.92 dB	32.84 dB	32.85 dB

3. BOWS AS A TEACHING TOOL

For the second phase of the BOWS contest during which the algorithm used to watermark the three images (documented in[5]) was made public, we have proposed to the Master students of the Institute Eurecom[67] to participate to the BOWS Contest.

A homework adapted from BOWS was then proposed (a preliminary draft of this homework is reported in appendix A): the aim of this homework is to break the watermark of the three images while preserving the highest possible quality of the content. This exercise allows master students to discover the difficulty of breaking a watermark and to develop a critical point of view about watermarking algorithms. In addition, it is a way to use some basic notions of image processing such as filtering, DCT transform and so on.

39 students have participated to this homework and were asked to provide their results as well as a report explaining how they broke the watermark within a delay of two weeks. The homework was appreciated and 21 students among 39 were able to break the watermark while keeping a PSNR more than 30 dB.

The majority of students, have applied some filters over the first image to break the watermark (Gaussian and Average filters). For the two other images, they have modified the DCT coefficients used in the insertion process.

The general conclusion provided by almost all the students is that the first image was relatively easy to attack because of its frequency content. The second and third images are more textured areas, it is indeed non trivial to attack. The knowledge of the watermarking technique being employed was useful.

We drew attention that some students have used the oracle attack to achieve a better PSNR for an attacked image e.g. they have computed some functions to randomly switch pixels from the attacked image by original ones. Repetitive blind tests were possible as the protocol did not include any restriction.

4. CONCLUDING REMARKS AND OPEN QUESTIONS

In this short note, we have described the techniques used to attack three watermarked images in the particular context of BOWS contest. We have first used an efficient technique based on self similarities in images and we have second improved the obtained results by using some basic image processing tools. BOWS has been also used as a teaching tool for master students. It was an attractive training for students to better understand the difficulty of breaking a watermark and to develop a critical point of view about watermarking algorithms. Nevertheless, there are some open questions in case of a possible future BOWS-II.

From an industrial point of view, we can discuss on one side how to encourage companies to propose their watermarking algorithms in order to face several and potential products and on the other side how much the proposed conditions are realistic: number of images to attack, PSNR between the original and the attacked image versus watermarked and attacked one, watermarking algorithm not publicly available during the first phase of the BOWS contest, etc.

From a scientific point of view, we can discuss about the best way to encourage participants to design new attacks with little tuning and working on different images and/or different watermarking technologies.

In addition to all this points, we have to keep in mind that policy of a possible future edition must nevertheless remain readable to attract as much participants as possible.

APPENDIX A. HOMEWORK BASED ON BOWS (FIRST DRAFT VERSION PROPOSED TO MASTER STUDENTS AT EURECOM DATED ON 27 OF MARCH 2006)

A.1. General advices

A.1.1. Before...

Before starting the homework, read the paper "*Applying Informed Coding and Embedding to Design a Robust, High capacity Watermark*", M. L. Miller, G. J. Doerr and I. J. Cox, IEEE Trans. on IP.13(6): 792-807, 2004. This paper describes the watermarking algorithm used to embed the watermark into the three images to attack. The companion C++ source code is available at http://www.adastral.ucl.ac.uk/gwendoer.

A.1.2. After...

- Save your best three images .raw and write a report including explanations on how you did and your comments. Good explanations mean that your results are (more or less) easy to reproduce thanks to them.
- Send an email to us in which you indicate the best PSNR you obtained and clearly state the complete path to get resulting images (do not forget to give the correct rights to the directories and files.) as well as your report by Monday 10 April.
- Even if you can collude during the period (i.e. exchange and share with some other students some ideas and/or some partial results or tools), you should nevertheless write at the end a personal report and send us individual results.
- Like Lab. sessions and Quiz, this homework will be evaluated and will be part of your final grade.

A.1.3. Image format...

Three different grayscale watermarked images in raw format and size 512x512 are available for download on the "BOWS" web-site (*http://lci.det.unifi.it/BOWS/*).
If your software cannot work with .raw formats, you can convert pictures to .pgm (or any other lossless format).

A.2. Goals

The aim of this homework (adapted from BOWS) is to investigate how and when an image watermarking system can be broken though preserving the highest possible quality of the content. The aim of this homework is to better understand which are the disparate attacks and comprehend the degree of difficulty of breaking the watermark.

A.3. Rules

Three different grayscale watermarked images in raw format and size 512x512 are available for download on the "BOWS" web-site. By RAW image format we mean that the pixel values are stored in the file (8 bits per pixels) without any header, with the image scanned from left to right and from top to bottom. Hence, reading the bytes in the file and interpreting them as unsigned char variables you immediately have the values of the pixels. Sometimes this .raw format is referred to as .dat or raster scan format.

You are asked to erase the watermark from all the three images by using any action you want while granting the best PSNR (Peak-Signal-to-Noise-Ratio) between the watermarked image and the attacked one (we have to send by e-mail your .raw images with the best PSNR). To verify your action, you can upload the attacked images on the "BOWS" web-site through an ad-hoc interface and ask to run the detection process; finally you obtain as answer the result of the detection and the PSNR achieved. If the attack has been successful, the image thumbnail will show the word "removed".

You must provide or/and store:

1. Your three images with the watermark removed and the best PSNR.
2. Your explanations on how you obtained the three images.

A.4. Useful tools and preliminary hints

- Assuming you have an estimate of the watermark but you do not know with which strength you have to retrieve it you can use the following procedure and tune α in order to obtain the best result.

  $W = I_{orig} - I_{att}$
  $I_{att}^{'} = I_{orig} - \alpha * W.$

  - If $\alpha = 1$ then $I_{att}^{'} = I_{att}$.
  - If $\alpha > 1$ your attack will be more important but your PSNR will decrease
  - If $\alpha < 1$ your attack will be less important (i.e. the watermark could be still detected) but the resulting PSNR will be higher (i.e. image of better quality).

  Eventually we can use a low pass filtering version of W.

- We can improve some results by averaging several (successfully) attacked images. You may obtain better results if the used methods to attack the image are different (explain why?).
- We can improve your result by taking care to preserve the average gray values (or DC component) associated with subregions or blocks, or more generally any other information not used by the signer, for example some AC coefficients (in case where the algorithm is working in the DCT domains on blocks).
- Colluders: It is allowed to collaborate with some other attackers. Collusion means fair (win-win) exchange of results or to share some observations or partial understandings/knowledge. If you use this option, do not forget to mention it in your report (which information and from who).
- ....

A.5. Some Matlab Codes

- *Compute DCT block by block...*
  imgDct = blkproc(img,[8 8],'dct2(x)');

- *Alpha the watermark...*
  I = imread('imm_x.bmp');
  I2 = imread('$imm_{x_{att}}$.bmp');
  W = imsubtract(int16(I),int16(I2));
  W2 = immultiply(W, alpha);
  H = fspecial('average');
  I3 = imsubtract(int16(I), int16(W2));
  imwrite(uint8(I3),'$imm_{x_{att_{plus}}}$.bmp','BMP');

- *Perserve average value of blocks...*
  function I= $mean_{adj}$(IO,IA,bsize)
  x=1;
  y=1;
  for j=1:(size(IO,2)/bsize-1)
  for i=1:(size(IO,1)/bsize-1)
  MO=mean(mean(IO(x:x+bsize-1,y:y+bsize-1)));
  MA=mean(mean(IA(x:x+bsize-1,y:y+bsize-1)));
  DM=MO-MA;
  IA(x:x+bsize-1,y:y+bsize-1)=imadd(IA(x:x+bsize-1,y:y+bsize-1),DM);
  x=x+bsize;
  end;
  x=1;
  y=y+bsize;
  end;
  imwrite(IA,'$A2x_4$.bmp','BMP');

REFERENCES

1. "Bows, http://lci.det.unifi.it/bows/."
2. C. Rey, G. Doërr, G. Csurka, and J.-L. Dugelay, "Toward generic image dewatermarking?," in *ICIP 2002, IEEE International Conference on Image Processing, September 22-25, 2002 - New York, USA - Volume 2 pp 633-636*, Sep 2002.
3. G. Doërr and J.-L. Dugelay, "How to combat block replacement attacks?," in *IH'05, 7th Information Hiding Workshop, June 6-8, 2005, Barcelona, Spain - Also published in LNCS Volume 3727*, Jun 2005.
4. G. J. Doërr, J.-L. Dugelay, and D. Kirovski, "On the need for signal-coherent watermarks," *IEEE Transactions on Multimedia, Volume 8 N5, October 2006* , 2006.
5. M. L. Miller, G. Doërr, and I. J. Cox, "Applying informed coding and embedding to design a robust high-capacity watermark," *IEEE Transactions on image processing, Volume 13, N6 - June 2004* , 2004.
6. "Eurecom institut, http://www.eurecom.fr."
7. "Course ImSecu Eurecom, http://www.eurecom.fr/util/coursdetail.en.htm?id=7."