A SIMPLE BENCHMARKING SYSTEM FOR PERFORMANCE EVALUATION OF STILL IMAGE WATERMARKING METHODS

V. Solachidis, S. Tsekeridou, A.Tefas and I.Pitas

Department of Informatics, Aristotle University of Thessaloniki,
Box 451, 54006 Thessaloniki, Greece,
Tel:+3031996361, Fax:+3031998419, e-mail: pitas@zeus.csd.auth.gr

ABSTRACT

A benchmarking system for performance evaluation and comparison of still image watermarking methods is presented in this paper. Robustness against a variety of attacks is evaluated. Performance assessment is based on ROC curve estimation and comparison. As an example, two distinct watermarking methods are tested by the benchmarking system and conclusions about their comparative performance are drawn.

1. INTRODUCTION

The development of multimedia services introduced new requirements for multimedia security and copyright protection techniques. Copying and reproduction of digital images and their distribution through the World Wide Web made watermark protection an essential requirement in image distribution. Thus, watermarking research has been developed rapidly in the last years.

Many methods have been presented in the literature and several watermarking software packages have been developed [1], thus introducing the need for evaluation of the relative performance of these methods against various types of attacks and distortions. A first attempt for the definition of benchmarking tool is reported in [2].

A simple benchmarking system for still image watermarking methods is presented in this paper. Two distinct methods are tested by the benchmarking system to evaluate their comparative performance. The watermarked images are distorted by a variety of attacks implemented in the benchmarking system. Thus, performance against attacks is also evaluated based on ROC curves.

In Section 2 the benchmarking system is presented. The watermarking method performance requirements are depicted in Section 3. Experimental results after testing the performance of two state of the art methods are shown in Section 4. Finally, conclusions are drawn in Section 5.

2. BENCHMARK SYSTEM

The benchmarking system involves the definition of the image database, the quality/distortion levels and the performance assessment criterion.

This work has been supported by the European Project IST-1999-10987 CERTIMARK

2.1. Greyscale image selection

Five 8-bit greyscale images of size 960×960 pixels, depicting portable Byzantine icons, have been selected for the comparison: PANAGIA1, PANAGIA2, PANAGIA3 (depicting the Virgin Mary), GENESIS, ANGELS. These images are typical images existing in the two image databases of Byzantine icons installed in Northern Greece.

2.2. Distortion levels

Image distortions introduced by the watermarking methods should be invisible, thus preserving the quality of the original image. As the power of the watermark signal increases, it is evident that such distortions become more visible. In our experimental setup, the signal to noise ratio (SNR) has been used as a quantitative measure for the distortions introduced by the watermarking systems. SNR values do not always correlate well with the visual perception of image distortions. However, since image quality metrics based on models of visual perception are difficult to be implemented and no globally agreeable metric exists, we have decided to still use SNR. Three SNR values, namely 40dB, 35dB and 30dB, have been chosen. For an SNR of 40dB, the watermark is clearly invisible. When SNR is 35dB, the watermark is almost invisible. Finally, when SNR falls to 30dB, the watermark starts becoming visible. A watermarked image having one of the three specified SNR values is obtained by properly adjusting the power of the watermark. Since some of the methods are not so flexible in terms of changing the watermark power, the above SNR values were approximated within a range of 1dB. It is also noted that the visibility of the watermark is also dependent on the domain where embedding is performed. For example, when embedding is performed to the detail components of the wavelet transform, its invisibility is greatly enhanced even for small SNR values.

2.3. Performance assessment

A correlator-based watermark detector is considered. The value of the correlator is normalized by the watermark power. Thus, the output of the detector module is given by:

$$\rho = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} x_w(i,j)\cdot w(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} w(i,j)^2} \tag{1}$$

Table 1: EER for each attack and SNR for both methods

Attack	*SNR 30*		*SNR 35*		*SNR 40*	
	Method 1	*Method 2*	*Method 1*	*Method 2*	*Method 1*	*Method 2*
No distribution	$2.37\ 10^{-283}$	$2.62\ 10^{-27}$	$1.81\ 10^{-124}$	$9.93\ 10^{-18}$	$1.25\ 10^{-54}$	$6.41\ 10^{-5}$
Jpeg compression 1/10	$1.14\ 10^{-74}$	$6.8\ 10^{-27}$	$2.1\ 10^{-29}$	$1.03\ 10^{-17}$	$2.26\ 10^{-15}$	$6.48\ 10^{-5}$
Jpeg compression 1/20	$7.24\ 10^{-23}$	$2.26\ 10^{-26}$	$7.65\ 10^{-13}$	$3.99\ 10^{-16}$	$5.87\ 10^{-6}$	$6.68\ 10^{-4}$
Jpeg compression 1/30	$9.97\ 10^{-10}$	$4.24\ 10^{-17}$	$5.29\ 10^{-7}$	$8.31\ 10^{-9}$	$3.35\ 10^{-3}$	$1.51\ 10^{-2}$
Jpeg compression 1/40	$3.02\ 10^{-4}$	$7.25\ 10^{-3}$	$3.81\ 10^{-3}$	$6.03\ 10^{-2}$	$5.21\ 10^{-2}$	$1.9\ 10^{-1}$
Jpeg compression 1/50	$1.37\ 10^{-2}$	$8.89\ 10^{-3}$	$4.97\ 10^{-2}$	$9.64\ 10^{-2}$	$2.66\ 10^{-1}$	$4.11\ 10^{-1}$
Median filtering (3×3)	$1.77\ 10^{-27}$	$1.93\ 10^{-23}$	$1.32\ 10^{-13}$	$5.6\ 10^{-15}$	$6.98\ 10^{-7}$	$2.91\ 10^{-4}$
Median filtering (5×5)	$9.57\ 10^{-3}$	$3.04\ 10^{-16}$	$3.05\ 10^{-2}$	$7.98\ 10^{-10}$	$1.36\ 10^{-1}$	$4.35\ 10^{-3}$
Moving average filtering (3×3)	$1.29\ 10^{-25}$	$3.7\ 10^{-25}$	$2.48\ 10^{-13}$	$1.07\ 10^{-15}$	$1.76\ 10^{-5}$	$2.1\ 10^{-4}$
Moving average filtering (5×5)	$2.03\ 10^{-1}$	$2.49\ 10^{-19}$	$1.19\ 10^{-1}$	$3.16\ 10^{-11}$	$4.1\ 10^{-1}$	$2.35\ 10^{-3}$
Gaussian noise $(\sigma = 10)$	$8.9\ 10^{-267}$	$9.54\ 10^{-26}$	$4.62\ 10^{-118}$	$4.83\ 10^{-17}$	$6.1\ 10^{-50}$	$1\ 10^{-4}$
Gaussian noise $(\sigma = 20)$	$1.08\ 10^{-227}$	$3.43\ 10^{-22}$	$1.61\ 10^{-91}$	$1.32\ 10^{-14}$	$4.64\ 10^{-41}$	$4.48\ 10^{-4}$
Histogram equalization	$1.43\ 10^{-171}$	$2.49\ 10^{-26}$	$3.85\ 10^{-128}$	$1.68\ 10^{-17}$	$3.44\ 10^{-74}$	$6.28\ 10^{-5}$
Quantization (7 bits)	$1.87\ 10^{-283}$	$2.35\ 10^{-27}$	$2.01\ 10^{-124}$	$1\ 10^{-17}$	$1.29\ 10^{-54}$	$6.4\ 10^{-5}$
Quantization (6 bits)	$1.7\ 10^{-283}$	$3.17\ 10^{-27}$	$3.35\ 10^{-122}$	$1.06\ 10^{-17}$	$1.47\ 10^{-54}$	$6.43\ 10^{-5}$
Scaling $\times 2$	0	$1.44\ 10^{-28}$	$6.5\ 10^{-205}$	$1.28\ 10^{-18}$	$2.46\ 10^{-71}$	$4.28\ 10^{-5}$
Scaling $\times 0.5$	$4.61\ 10^{-25}$	$1.87\ 10^{-22}$	$2.59\ 10^{-11}$	$6.54\ 10^{-15}$	$6.97\ 10^{-5}$	$3.48\ 10^{-4}$
Rotation $2°$	$3.97\ 10^{-2}$	$8.86\ 10^{-4}$	$1.11\ 10^{-1}$	$1.5\ 10^{-2}$	$1.85\ 10^{-1}$	$2.18\ 10^{-1}$
Cropping 75%	$2.69\ 10^{-275}$	$4.73\ 10^{-7}$	$4.54\ 10^{-126}$	$4.17\ 10^{-5}$	$6.02\ 10^{-50}$	$3.26\ 10^{-2}$

In (1),M, N are the image dimensions, w denotes the watermark signal, whereas x_w denotes the watermarked image appropriately transformed to the embedding domain. In order to decide for the presence of a certain watermark in an image, ρ is compared against a threshold T. If $\rho > T$, then the detector concludes that the watermark is present. T may be chosen so that the probability of false alarm P_{fa}, i.e. the probability of detecting a watermark in an image that is not watermarked or is watermarked with a different watermark, is smaller than a predefined value. For this P_{fa} and T, a respective probability of rejection P_{rej}, i.e. the probability to fail to detect a watermark in an image when it actually exists, is additionally defined. Such probabilities are measured under the assumption that the correlation values attain a gaussian distribution. These values of P_{fa} and P_{rej} for a given threshold T, represent a point of the receiver operating curve (ROC) of the watermarking system. To be able to better evaluate the watermarking systems' performance, ROC curves instead of ROC points are estimated, by varying the value of T and measuring P_{fa} and P_{rej} for each value. During this process, each image is watermarked in turn with one of 100 different watermarks (500 watermarks in total, for all five images) and correlation values are evaluated for two cases: i) when another watermark, not belonging to the above set, is used for detection leading to a gaussian-like distribution having zero mean value (left pdf), and ii) when the right watermark inside the above set is used for detection, leading to another gaussian-like distribution with mean value, ideally 1, but usually lying in the range $(0, 1]$ (right pdf). Correlation values for both cases are evaluated under a variety of image attacks, linear or non-linear, that will be summarized in the next subsection. Under non-linear attacks, the right pdf for each image exhibits shifts in its mean value, since the role the image plays in the correlation values of (1) cannot be entirely cancelled out. This leads to an overall distribution that is not a single gaussian pdf but rather a mixture of equiprobable gaussian pdfs:

$$f(y) = \sum_{k=1}^{K} \alpha f_k(y) \qquad (2)$$

where $f(y)$ is the overall pdf, $f_k(y)$ are the single gaussian pdfs, α is a weight equal to $\frac{1}{K}$ and K is the total number of single gaussian-like pdfs for each image. K is equal to 5 in our case, since 5 different images are used. Thus, by considering (2), one may conclude that the overall error probability may be evaluated as the equally weighted sum of the error probabilities evaluated by each single gaussian pdf. P_{fa} values are estimated using the approximated left pdf, while P_{rej} values are evaluated using the approximated right pdf.

3. PERFORMANCE REQUIREMENTS-BENCHMARKS

For each of the five images, for each of the three SNR values and for each watermarking method, the following performance criteria are considered:

1. *Detectability after JPEG compression.* Minimum requirement: robustness to 1:10 compression ratio. The following compression ratios are tested: 1:10 up to 1:50 in increments of ten. Watermarked images are compressed using the JPEG compression/decompression routines provided by the Independent JPEG Group. Since these routines employ the notion of quality factor, a lookup table that provides for each

value of compression ratio the corresponding quality factor has been experimentally evaluated for each of the five images (Table 2).

Table 2: Correspondence between compression ratio and JPEG quality factor for each image under study.

	Image	Compression Ratio 1:10	1:20	1:30	1:40	1:50
Quality	angels	71	28	14	9	6
	genesis	76	33	17	10	7
	panagia1	47	16	9	6	-
	panagia2	35	12	7	-	-
	panagia3	46	16	9	6	-

2. *Detectability after lowpass filtering.* Minimum requirement: robustness to 3×3 moving average and median filtering. Tests are conducted with 5×5 window sizes, as well.
3. *Detectability after noise addition.* Minimum requirement: robustness to additive zero-mean Gaussian noise with standard deviation $\sigma = 10$. Tests are conducted with $\sigma = 20$, as well.
4. *Detectability after histogram equalization.*
5. *Detectability after gray level reduction using a uniform quantizer.* Minimum requirement: robustness after grayscale reduction to 128 gray levels. Reduction down to 64 gray levels has is considered.
6. *Detectability after scaling using linear interpolation.* Minimum requirements: robustness to zooming with a factor of 2, decimation with a factor of 0.5 (in each dimension). Zooming with a factor of 1.5 and decimation with a factor of 0.75 are also considered during the evaluation of the left pdf, in order to somehow account for the possibility of the correlator attaining a correlation peak at an incorrect scaling factor.
7. *Detectability after cropping.* Minimum requirement: robustness to 75% cropping of the original image. Tests are also conducted with 85% cropping during the evaluation of the left pdf.
8. *Detectability after rotation.* Minimum requirement: robustness to a rotation of 2^o. Tests are conducted with a rotation angle of 10^o for the evaluation of the left pdf. Rotation is performed by mapping pixels of the destination image to the source image. Nearest neighbor interpolation is used on the source image.
9. *Detectability after multiple watermarking.* Minimum requirement: detectability among 5 distinct watermarks.

It is noted that the left pdf, needed for P_{fa} error evaluation, is approximated by combining the correlation values obtained for all cases of attacks aiming at better evaluation assessment.

4. EXPERIMENTAL RESULTS

Performance results for two selected watermarking method s [3],[4] are reported here. The first one, Method 1 [3], embeds selfsimilar watermarks in the wavelet domain. The second one, Method 2, [4], embeds circularly symmetric watermarks in the Fourier domain. ROC curves for all SNRs and attacks can not be presented due to lack of space. Thus EERs for each attack and SNR for both methods are displayed in Table 1. Comparison is performed by evaluating the respective ROC curves under all image attacks mentioned in the previous subsection:

1. *No distortion.* The first method exhibits better performance than the second, but both attain very small P_{fa} and P_{rej} values, less than 10^{-4} for all SNR cases. The superiority of the first method is attributed to the fact that the watermark there affects more pixels of the watermarked image (medium and high frequencies in the wavelet domain) than in the second method (embedded only in the medium frequencies in the Fourier domain)(Figure 1a).
2. *Moving average filtering.* Both methods prove almost similarly robust against 3×3 moving average filtering. False alarm and rejection probabilities are both below the limit of 10^{-4}. In the case of 5×5 moving average filtering however only the second watermarking method proves sufficiently robust when $SNR = 30$ db and 35 db. The deterioration in performance for the first method is due to the watermark embedding in high frequencies as well, which are greatly affected by intense smoothing caused by filtering with large window sizes (Figure 1c).
3. *Median filtering.* Similar conclusions are drawn for the performance of both methods as in the case of moving average filtering (Figure 1b).
4. *JPEG compression.* When SNR is 30dB or 35dB, both methods exhibit remarkable robustness for compression ratios up to 1/30. When SNR is equal to 40dB, robustness is noted for compression ratios up to 1/10 (Figure 1d).
5. *Noise addition.* The first method proves rather robust against zero-mean noise addition for all SNR cases and both σ values (10 and 20). The second method has also satisfactory performance but in the case that $SNR = 40$dB and $\sigma = 20$ its performance deteriorates.
6. *Histogram equalization.* Both methods have good performance against histogram equalization for all SNR cases.
7. *Gray level reduction.* Such image modification proves to only slightly affect the performance of both methods exhibited under no distortion for all SNR cases. What is worth mentioning is that their performance is almost identical irrespectively of the final number of gray levels.
8. *Scaling.* Both methods prove rather robust against uniform linear scaling, even in the case when the scaling factor is equal to 0.5 translating to significant image smoothing and less correlation reliability as well. In the case of $SNR = 30$ dB and for a scaling factor of 2, the first method attains very small error probabilities that cannot be computed resulting in no ROC curve.

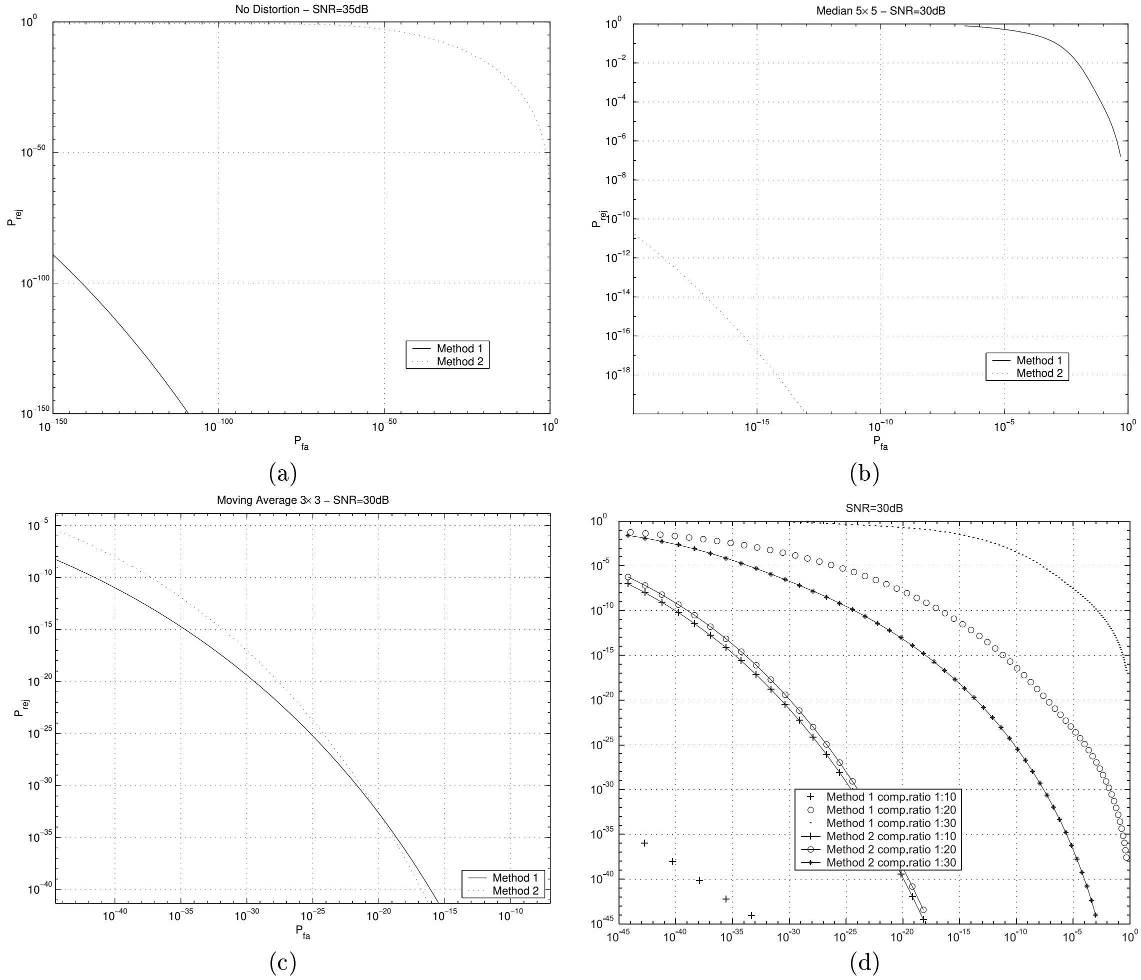

Figure 1: ROC curves for both methods where the images are (a) not distorted (b) filtered by median 5×5 (c)filtered by moving average filter 3×3, (d)compressed by JPEG compression (ratios 1/10, 1/20, 1/30)

9. *Cropping.* The first method is superior to the second method in the case of cropping. This may attributed to the fact that the latter affects a much smaller number of image pixels and since cropping reduces even more this number, the correlation values finally evaluated tend to be quite low.
10. *Rotation.* Both methods prove to behave badly in the rotation case even of 2^o, with the only exception of the first method in the case of SNR being 30dB. It is noted however that, due to the circular symmetry of the watermarks embedded, no rotation has been attempted during the estimation of the correlation values.
11. *Multiple Watermarking.* In this case, the first method proves once more superior in performance than the second one for all SNR cases.

5. CONCLUSIONS

A benchmarking system for performance evaluation and comparison of still image watermarking methods is presented. Watermarking methods are compared through ROC curve evaluation and comparison under a variety of image attacks. As an example, two distinct watermarking methods are considered for comparison.

REFERENCES

[1] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1079–1107, July 1999.

[2] F.A.P. Petitcolas, R.J. Anderson, and M.G. Kuhn, "Attacks on copyright marking systems," in *Second International Workshop on Information Hiding*, Portland, USA, April 15-17 1998, pp. 219–239.

[3] S. Tsekeridou and I. Pitas, "Embedding self-similar watermarks in the wavelet domain," in *Proc. of ICASSP'00*, Istanbul, Turkey, 5-9 June 2000, to appear.

[4] V. Solachidis and I. Pitas, "Circularly symmetric watermark embedding in 2-d dft domain," in *Proc. of ICASPP'99*, Phoenix, Arizona, USA, 15-19 March 1999.