Fair evaluation methods for image watermarking systems

M. Kutter[a] and F. A. P. Petitcolas[b]

[a]Signal Processing Laboratory, Swiss Federal Institute of Technology,

Ecublens, 1015 Lausanne, Switzerland

[b]The Computer Laboratory, University of Cambridge,

Pembroke Street, Cambridge CB2 3QG, United Kingdom

ABSTRACT

Since the early 90's a number of papers on "robust" digital watermarking systems have been presented but none of them uses the same robustness criteria. This is not practical at all for comparison and slows down progress in this area. To address this issue, we present an evaluation procedure of image watermarking systems. First we identify all necessary parameters for proper benchmarking and investigate how to quantitatively describe the image degradation introduced by the watermarking process. For this, we show the weaknesses of usual image quality measures in the context watermarking and propose a novel measure adapted to the human visual system. Then we show how to efficiently evaluate the watermark performance in such a way that fair comparisons between different methods are possible. The usefulness of three graphs: "attack vs. visual-quality," "bit-error vs. visual quality," and "bit-error vs. attack" are investigated. In addition the receiver operating characteristic (ROC) graphs are reviewed and proposed to describe statistical detection behavior of watermarking methods. Finally we review a number of attacks that any system should survive to be really useful and propose a benchmark and a set of different suitable images.

Keywords: digital watermarking, benchmark, evaluation, quality metric, robustness

Further author information:

M. K.: e-mail: Martin.Kutter@epfl.ch

F. A. P. P.: e-mail: w+fapp2@cl.cam.ac.uk

1. INTRODUCTION

At the beginning of 1990 the idea of digital watermarking, embedding imperceptible information into audiovisual data, has emerged. Since then worldwide research activities have been increasing dramatically and the industrial interest in digital watermarking methods keeps growing. The first academic conference on the subject was organised in 1996.[2] Digital watermarks have mainly three application fields: data monitoring, copyright protection, and data authentication. The first watermarking methods were proposed for digital images by Caronni[8,9] in 1993, although earlier publications already introduced the idea of tagging images to secretly hide information and ensure ownership rights.[43,42] Since then, the idea of digital watermarking has been extended to other data such as audio and video. For recent overviews of digital watermarking methods the reader is referred to Anderson,[2] Aucsmith,[3] and Swanson et al.[40]

Besides designing digital watermarking methods, an important and often neglected issue addresses proper evaluation and benchmarking. This not only requires evaluation of the robustness, but also includes subjective or quantitative evaluation of the distortion introduced through the watermarking process. Only few authors (e.g., Braudaway[7] or Kutter et al.[22]) report quantitative results on the image degradation due to the watermarking process. In general, there is a tradeoff between watermark robustness and watermark visibility. Hence, for fair benchmarking and performance evaluation one has to ensure that the methods under investigation are tested under comparable conditions.

In this paper we propose a way to evaluate and compare performances of "robust" invisible watermarking systems. In Section 2 we redefine the generic watermarking scheme and identify important parameters and variables. Distortion metrics and attacks on watermarks are described in Section 3 and Section 4, respectively. In Section 5 we propose different graphs useful for performance assessment. Our benchmark procedure and an image database are introduced in Section 6.

2. DIGITAL WATERMARKING: FRAMEWORK, DEFINITIONS AND PARAMETERS

In order to identify important watermarking parameters and variables, we first need to have a look at the generic watermarking embedding and recovery schemes. In the following we use the same notation for sets and their elements; the difference should be obvious to the reader.

Figure 1 illustrates the generic embedding process. Given an image I, a watermark W and a key K (usually the seed of a random number generator) the embedding process can be defined as a mapping of the form: $I \times K \times W \rightarrow \tilde{I}$ and is common to all watermarking methods.

The generic detection process is depicted in Figure 2. Its output is either the recovered watermark W or some kind of confidence measure indicating how likely it is for a given watermark at the input to be present in the image $\tilde{I}'$ under inspection.

There are several types of watermarking systems. They are defined by their inputs and outputs:

- **Private watermarking** systems require at least the original image. *Type I* systems, extract the watermark W from the possibly distorted image $\tilde{I}'$ and use the original image as a hint to find where the watermark could be in $\tilde{I}'$ ($\tilde{I}' \times I \times K \rightarrow W$). *Type II* systems (e.g.,[9,10,36]) also require a copy of the embedded watermark for extraction and just yield a 'yes' or 'no' answer to the question: does $\tilde{I}'$ contain the watermark W? ($\tilde{I}' \times I \times K \times W \rightarrow \{0,1\}$). It is expected that this kind of scheme will be more robust than the others since it conveys very little information and requires access to secret material.[11]

- **Semi-private watermarking** does not use the original image for detection ($\tilde{I}' \times K \times W \rightarrow \{0,1\}$) but answers the same question. The only use of private and semi-private watermarking seems to be evidence in court to prove ownership and copy-control in applications such as DVD where the reader needs to know whether it is allowed to play the content or not. A large number of the currently proposed schemes fall in this category.[5,20,26,27,46,49,56]

- **Public watermarking** (also referred to as *blind* watermarking) remains the most challenging problem since it requires neither the secret original I nor the embedded watermark W. Indeed such systems really extract n bits of information (the watermark) from the watermarked image: $\tilde{I}' \times K \rightarrow W$.[15,16,21,23,41,57] Public watermarks have much more applications than the others and we will focus our benchmark on these systems. Actually the embedding algorithms used in public systems can always be used into a private one improving robustness at the same time.

- There is also **asymmetric watermarking** (or *public key watermarking*) which has the property that any user can read the watermark, without being able to remove it.

After grouping the different systems, we can now identify important parameters and variables.

- **Amount of embedded information** – This is an important parameter since it directly influences the watermark robustness. The more information one wants to embed, the lower is the watermark robustness. The information to be hidden depends on the application. In order to avoid small scale proprietary solutions, it seems reasonable to assume that one wants to embed a number similar to the one used for ISBN* (roughly 10 digits) or better ISRC† (roughly 12 alphanumeric characters). On top of this, one should also add the year of copyright, the permissions granted on the work and rating for it.[31] This means that roughly 70 bits of information should be embedded in an image. This does not include extra bits added for error correction codes.

- **Watermark embedding strength** – There is a tradeoff between the watermark embedding strength (hence the watermark robustness) and quality. Increased robustness requires a stronger embedding, which in turn increases the visual degradation of the images.

- **Size and nature of the picture** – Although very small pictures do not have much commercial value, a watermarking software needs to be able to recover a watermark from them. This avoids a "Mosaic" attack[34] on them and allows tiling, used very often in Web applications. For printing applications high resolution images are required but one also wants to protect these images after they are resampled and used on the Web. Photographers and stock photo companies have great concerns about having their work stolen and most of them still rely on small images, visible watermarks and even "rollover java scripts"‡ to reduce infringement. Furthermore the nature of the image has also an important impact on the watermark robustness. Very often methods featuring a high robustness for scanned natural images have a surprisingly reduced robustness for synthetic images (e.g., computer generated images). A fair benchmark should use a wide range of picture sizes, from few hundred to several thousands pixels, and different kind of images.

- **Secret information (e.g., key)** – Although the amount of secret information has no direct impact on the visual fidelity of the image or the robustness of the watermark, it plays an important role in the security of

*International Standard Book Numbering

†International Standard Recording Code

‡These scripts are used to display images in such a way that they are replaced by another image (typically a copyright sign) when the user moves the cursor on it to save it. Contrary to popular belief, this does not provide any security.

the system. The key space, that is the range of all possible values of the secret information,[§] must be large enough to make exhaustive search attacks impossible. The reader should also keep in mind that many security systems fail to resist to very simple attacks because of bad software engineering.[1,34]

3. VISUAL QUALITY METRICS

As mentioned in the previous section, the watermark robustness dependents directly on the embedding strength, which in turn influences the visual degradation of the image. For fair benchmarking and performance evaluation, the visual degradation due to the embedding is an important and unfortunately often neglected issue. Since there is no universal metric, we review in this section the most popular pixel based distortion criteria and introduce one metric which makes use of the effect in the human visual system (HVS).

3.1. Pixel Based Metrics

Most distortion measures or quality metrics used in visual information processing belong to the group of *difference distortion measures.*[39] The first part of Table 7 lists the most popular difference distortion measures. Theses measure are all based on the difference between the original, undistorted and the modified, distorted signal. The second part of the same table shows distortion measures based on the correlation between the original and the distorted signal. For a comparative study of the measures the interested reader is referred to Eskicioglu and Fisher.[14]

Nowadays, the most popular distortion measures in the field of image and video coding and compression are the *Signal to Noise Ratio (SNR)*, and the *Peak Signal to Noise Ratio (PSNR)*. They are usually measured in *decibels* (dB): $SNR(dB) = 10\log_{10}(SNR)$.

Their popularity is very likely due to the simplicity of the metric. However, it is well known that these difference distortion metrics are not correlated with human vision. This might be a problem for their application in digital watermarking since sophisticated watermarking methods exploit in one way or the other the HVS. Using the above metric to quantify the distortion caused by a watermarking process might therefore result in misleading quantitative distortion measurements. Furthermore these metrics are usually applied to the luminance and chrominance channels of images. If the watermarking methods work in the same color-space, for example luminance modification, this does not pose a problem. On the contrary, if the methods use different color spaces, these metric are not suitable.

[§]Depending on the type of watermarking, the key space can be K, $W \times K$ or a subset of $I \times W \times K$.

3.2. Perceptual Quality Metrics

The weaknesses of the pixel-based distortion metrics have been known for a long time. In recent years more and more research concentrates on distortion metrics adapted to the human visual system by taking various effect into account.[47,52–54] In this paper, we make use of a distortion metric proposed by van den Branden Lamprecht and Farrell.[47] The perceptual quality measure exploits the contrast sensitivity and masking phenomena of the HVS and is based on a multi-channel model of the human spatial vision.

Computing the metric involves the following steps: coarse image segmentation, decomposition of the coding error and the original image into perceptual components using filter banks, computing the detection threshold for each pixel using the original image as masker, dividing the filtered error by the decision threshold, pooling over all color channels. The unity for the metric is given in *units above threshold* also referred to as *Just Noticeable Difference* (JND). The overall metric, *Masked Peak Signal to Noise Ratio (MPSNR)* is then given by:

$$MPSNR = 10 \log_{10} \frac{255^2}{E^2} \tag{1}$$

where E is the computed distortion. Since this quality metric does not have exactly the same meaning as the known dB's, it is referred to as *visual decibels* (vdB). A normalised quality rating is often more useful. We use the ITU-R Rec. 500 quality rating Q. The rating is computed as:

$$Q = \frac{5}{1 + N \times E} \tag{2}$$

where E is the measured distortion and N a normalisation constant. N is usually chosen such that a known reference distortion maps to the corresponding quality rating. Table 2 lists the ratings and the corresponding visual perception and quality.

The ITU rating has several advantages, such as not blowing up for not distorted images, over the $MPSNR$ quality metric and is hence more suitable for the watermarking purpose.

The software to compute the presented distortion metric is available for non commercial usage and can be downloaded at `<http://ltswww.epfl.ch/mpqm/>`.

4. POSSIBLE ATTACKS ON WATERMARKS

We propose here a list of attacks against which watermarking system could be judged. We do not make a difference between intentional and unintentional processing.

- **JPEG compression** – JPEG is currently one of the most widely used compression algorithms for images and any watermarking system should be resilient to some degree of compression.
- **Geometric transformations**
  - **Horizontal flip** – Many images can be flipped without loosing any value. Although resilience to flipping is usually straightforward to implement only very few systems do survive it.
  - **Rotation** – Small angle rotation, often in combination with cropping, does not usually change the commercial value of the image but can make the watermark un-detectable. Rotations are used to realign horizontal features of an image and it is certainly the first modification applied to an image after it has been scanned. For benchmarking we propose to crop the rotated image so that there is no need to add a fixed color border to it.
  - **Cropping** – In some cases, infringers are just interested by the "central" part of the copyrighted material, moreover more and more Web sites use image segmentation, which is the basis of the "Mosaic" attack.[32] This is of course an extreme case of cropping.
  - **Scaling** – As we noticed earlier, this happens when a printed image is scanned or when a high resolution digital image is used for electronic applications such as Web publishing. This should not be neglected as we move more and more toward Web publishing. Scaling can be divided into two groups, uniform and non-uniform scaling. Under uniform scaling we understand scaling which is the same in horizontal and vertical direction. Non-uniform scaling uses different scaling factors in horizontal and vertical direction (change of aspect ratio). Very often digital watermarking methods are resilient only to uniform scaling.
  - **Deletion of lines or columns** – This was our first attack on some copyright marking systems and is very efficient against any straightforward implementation of spread-spectrum techniques in the spatial domain. Removing k samples at regular intervals in a pseudo random sequence $(-1, 1)$ (hence shifting the next ones) typically divides by k the amplitude of the cross correlation peak with the original sequence.

- **Generalised geometrical transformations** – A generalised geometrical transformation is a combination of non-uniform scaling, rotation, and shearing.

- **Random geometric distortions (StirMark)** – These distortions were detailed in an earlier paper[33,34] and we suggested that image-watermarking tools, which do not survive them should be considered unacceptably easy to break.

- **Geometric distortions with JPEG** – Rotation, and scaling alone are not enough they should be also tested in combination with JPEG compression. Since most artists will first apply the geometric transformation and then save the image in a compressed format it makes sense to test robustness of watermarking system to geometric transformation followed by compression. However an exhaustive test should also include the contrary since it might be tried by willful infringers. It is difficult to chose a minimal "quality factor" for JPEG as artifact quickly appear. However experience from professionals shows that "quality factors" down to 70% are reasonable. Artists seem to use JPEG extensively as well as resizing.[17]

- **Enhancement techniques**

  - **Low pass filtering** – This includes linear and non-linear filters. Frequently used filters include median, Gaussian, and standard average filters.

  - **Sharpening** – Sharpening functions belong to the standard functionalities of photo edition software. These filters can be used as an effective attack on some watermarking schemes because they are very effective at detecting high frequency noise introduced by some digital watermarking software. More subtle attacks are based on the Laplacian operator[4]: in its simplest version the attacked image is $\tilde{I}' = I - \alpha\nabla^2(\nabla^2 I - I)$ where α is the strength of the attack.

  - **Histogram modification** – This includes histogram stretching or equalisation which are sometimes used to compensate poor lightening conditions.

  - **Gamma correction** – Very frequently used operation to enhance images or adapt images for display, for example after scanning.

  - **Color quantisation** – This is mostly applied when pictures are converted to the Graphics Interchange Format (GIF) extensively used for Web publishing. Color quantisation is very often accompanied by **dithering** which diffuses the error of the quantisation.
  - **Restoration** – These techniques are usually designed to reduce the effects of specific degradation processes but could also be used without priori knowledge of the noise introduced by the watermarking system.[33]

- **Noise addition** – Additive noise and uncorrelated multiplicative noise have been largely addressed in the communication theory and signal processing theory literature. Authors often claim that their copyright marking techniques survive this kind of noise but many forget to mention the maximum level of acceptable noise.
- **Printing-scanning** – This process introduces geometrical as well as noise-like distortions.
- **Statistical averaging and collusion** – Given two or more copies of the same image but with different marks, it should not be possible to remove the marks by averaging these images or by taking small parts of all images and reassembling them.
- **Over-marking** – In this case the attacker needs special access to the marking software. Current commercial implementations will refuse to add a watermark if another is already embedded. Consequently the attacks need to bypass the test implemented in the software.[6] However manufacturers have full access to the marking software and can perform this test without any difficulty.
- **Oracle attack** – When a public decoder is available, an attacker can remove a mark by applying small changes to the image until the decoder cannot find it anymore. A theoretical analysis of this attack and a possible countermeasure (randomising the detection process) have been presented recently.[24] One could also make the decoding process computationally expensive. However neither approach is really satisfactory in the absence of tamper-resistant hardware.

We consider this list to be a minimum for watermarking testing. Random non-linear imperceptible geometric distortions are still very challenging and solutions have not been discussed yet.

5. PERFORMANCE EVALUATION AND REPRESENTATION

In order to properly evaluate the performance of watermarking schemes and allow fair comparison between different schemes, the test setup conditions are of high importance. In this section the possible evaluation tools are outlined, together with the test setup and conditions. Table 3 lists useful graphs, together with the variable and fixed parameters for comparison.

For all evaluation strategies it is very important to perform the tests using different keys and a variety of images with changing image size and nature. The results should then be averaged and plotted. If performance evaluation on individual images is required, for example for direct performance comparison of two methods for one image, it is still very important that all tests are repeated several times, using different keys. In the following, *attack* refers to any attack of the previous section. The term *robustness* describes the watermark resistance to these attacks and can be measured by the *bit-error rate* which is defined as the ratio of wrong extracted bits to the total number of embedded bits. The *visual quality* is the result of a distortion metric such as $MPSNR$.

In order to illustrate the usefulness of the proposed graphs, we implemented a comparative scenario for two simple watermarking methods. Both methods are based on spread-spectrum modulation, but in different domains. One method uses the spatial domain while the other method uses a multi-resolution environment (three level wavelet transform with Daubechies 6 tap filters). The systems use a secret key which serves as seed for a pseudo random number generator used to generate the spread spectrum sequences. As robustness measure we use the bit-error rate, the metric for the visual quality is the rating Q introduced in Section 3.2, and the attack is JPEG compression. All tests were performed on the 512×512, 24-bit colour version of *lena*. Each test was repeated using each time a randomly chosen key. The watermark length is 100 bits.

5.1. Bit-Error vs. Attack Strength Graph

One of the most important graphs relates the watermark robustness to the attack. Usually this graph shows the bit-error rate as a function of the attack strength for a given visual quality. Several papers have used this graph, unfortunately without explicitly reporting the visual image quality. This evaluation allows direct comparison of the watermark robustness and shows the overall behaviour of the method towards attacks.

Figure 3 shows this graph for our example. Each test was repeated 10 times, using different keys, and the visual

quality rating was fixed to 4.5. It is clearly visible, that for a given visual quality the multi-resolution scheme has superior performance.

5.2. Bit-Error vs. Visual Quality Graph

The "bit-error vs. visual quality" graph shows the relationship between the bit-error and the visual image quality for a fixed attack. For a given attack, this graph can be used to determine the expected bit-error for a desired visual quality. This might be especially useful to determine the minimal visual quality for a desired bit-error rate under a given attack.

Figure 4 shows the graph for our example. The test was repeated 10 times for each image, using a different key. The attack was JPEG-compression at 75% quality. The individual results were then averaged and plotted. We can easily determine the maximal achievable visual quality such that, for example, the bit-error does not exceed a desired value. In addition, the same figure clearly illustrates that for a desired bit-error rate the multi-resolution watermarking scheme allows higher visual qualities.

5.3. Attack vs. Visual Quality Graph

The "attack vs. visual quality" graph illustrates the maximum allowable attack as a function of the visual quality for a given robustness. This graph allows immediate evaluation of the allowable watermark attack for given visual qualities. This is especially useful if the visual quality range is given and the corresponding maximal allowable distortion, i.e. watermark attack, needs to be evaluated. Furthermore this graph is very useful in comparing different watermarking methods since it facilitates immediate robustness comparisons for a given visual image quality at a fixed bit-error rate.

Figure 5 shows the graph for our example. The bit-error rate was fixed to 0.1 and every test was repeated 5 times using a different key. The graph clearly shows the superior performance of the multi-resolution approach. For a given visual quality, the spatial watermarking algorithm requires much higher compression qualities.

5.4. Receiver Operating Characteristic Graphs

Given any image a watermark detector has to fulfill two tasks: decide if the given image is watermarked and decode the encoded information.

The former can be seen as hypothesis testing in that the watermark decoder has to decide between the *alternative hypothesis* (the image is watermarked) and the *null hypothesis* (the image is not watermarked). In binary hypothesis testing two kinds or errors can occur: accepting the alternative hypothesis, when the null hypothesis is correct and accepting the null hypothesis when the alternative hypothesis is true. The first error is often called *Type I* error or *false positive* and the second error is usually called *Type II* error or *false negative*.

Receiver Operating Characteristic (ROC) graphs[58] are very useful in assessing the overall behavior and reliability of the watermarking scheme under inspection. Usually in hypothesis testing, a test statistic is compared against a threshold to decide for one or the other hypothesis. Comparing different watermarking schemes with a fixed threshold may result in misleading results. ROC graphs avoid this problem by comparing the test using varying decision thresholds. The ROC graph shows the relation between the *true positive fraction (TPF)* on the y-axis and the *false positive fraction (FPF)* on the x-axis.

The true positive-fraction is defined as:

$$TPF = \frac{TP}{TP + FN} \tag{3}$$

where TP is the number of true-positive test results, and FN is the number false negative tests. The false-positive fraction is defined as:

$$FPF = \frac{FP}{TN + FP} \tag{4}$$

where FP is the total number of false-positive test results, and TN is the number of true negative test results. In other words, the ROC graph shows TPF-FPF pairs resulting from a continuously varying threshold. An optimal detector has a curve that goes from the bottom left corner to the top left, and then to the top right corner. The diagonal from the bottom left corner to the top right corner describes a detector which randomly selects one or the other hypothesis with equal probability. Hence, the higher the detector accuracy, the more its curve approaches the top left corner. Often the integral under the curve is used as a detector performance measure.To generate these graphs, the same number of watermarked and non-watermarked images should be tested. If the overall performance of watermarking methods is to be evaluated, tests should included a variety of attacks with varying parameters.

Figure 6 shows the ROC graph for our example. Each test was repeated 10 times using a different key and the visual quality was set to 4.5. The attack was JPEG-compression and the quality factor was varied from 30% to 100% in steps of 5%. The two curves in the graph show, that the multi-resolution scheme features higher detection reliability. Furthermore it is interesting to note, that the spatial domain approach has a tendency to reject watermarked images.

6. A BENCHMARK

As we noticed in the introduction a number of broad claims have been made about the "robustness" of various digital watermarking method. Unfortunately the criteria as well as the pictures used to demonstrate these claims vary from one system to the other and recent attacks[24,25,33,34] show that the robustness criteria used so far are inadequate: JPEG compression, additive Gaussian noise, low pass filtering rescaling, and cropping have been addressed in most the literature[5,9–12,15,18–20,22,23,26,27,36,37,44–46,49,50,55,56] but specific distortions such as rotation have been rarely addressed.[21,29] In some cases the watermark is simply said to be "robust against common signal processing algorithms and geometric distortions when used on some standard images."

Most of the potential attacks detailed in Section 4 are actually implemented into the latest version of StirMark[35]: given a watermarked image, StirMark will apply these transformations with various parameters. Then the output images can be tested with watermark detection or extraction programs. The full process can be automated using a simple batch file and future version will propose a web interface were users can send their libraries.

6.1. Image Database

It is important to test an image watermarking software on many different images and for fair comparison the same set of sample images should always be used. Pictures can be interesting from the signal processing point of view: textured/smooth areas, size, synthetic, with straight edges, sharp, blur, brightness/contrast, etc. They should also cover a broad range of contents and types. It is impossible to get an exhaustive list of classes of pictures and stock photo companies have a lot of difficulties to set up a satisfactory index. However one can at least retain the main themes that are common among these libraries and that are used very often in the press in order to keep a wide range of kind of pictures: colors, textures, patterns, shapes, lightning.

Some image databases already exist for image processing research. The USC-SIPI Image Database is an example of such database[51] where one can find the "classics:" lena, baboon, peppers, etc. Using these databases for research

on digital watermarking and indeed copyright protection is somewhat hypocritical as "some of the images in the database were scanned from copyrighted material"[51] and the "origin of many is unknown."[51] Consequently we tried to find a wide range of other photographers and got the authorisation to use them freely for research on watermarking (including publication in proceedings or journals) as long as credit is given to the photographer.

6.2. Rating and Procedure

We suggest the following procedure for the robustness tests. For each image in the set:

- Embed a watermark with strongest strength which does not introduce annoying effects: the quality rating defined in Section 3.2 should be at least 4. In the case of private watermarking of type I, semi-public watermarking and public-watermarking system embed an 80-bit watermark.
- Apply the StirMark benchmark to produce a set of distorted version of the image. The set of applied distortions and the strength of these distortions highly depend on the application of the watermarking algorithm being evaluated. Currently StirMark implements only a fixed set of "attacks" but future versions will be customisable.
- Try to detect (say 80%) or recover (all) the watermark.

For this first version of our benchmark we did not put any weighting on the possible attacks. A watermark extracted or detected successfully – depending on the type of watermarking – gives one mark. For each type of attack these marks are averaged to give a mark out of 20 and the resulting marks are averaged to give the final mark.

Table 4 shows early results based on a subset of the transformation described previously and without using any quality measurement, just the naked eye and default software parameters.[¶] Although comparison should be done with great care, the table confirms what is currently achieved in term of robustness and what needs further research.

For the results summarised in table 5, we followed exactly[‖] the procedure detailed previously. The images used for the test were *lena*, *baboon*, *fishing boat*, *bear*, *skyline arch* and *watch*[**]. We will keep adding new results with other images but after four images we noticed that the average results were stable. Detailed results, including strength

[¶] For this evaluation, we used the images available at `<http://ltssg3.epfl.ch:1248/kutter/watermarking/database.html>`

[‖] Except for Signum SureSign for which we could not choose the strength of the embedding and used default parameters.

[**] These images are available at `<http://www.cl.cam.ac.uk/~fapp2/watermarking/benchmark/image_database.html>`

of the watermarks and $PSNR$ of the watermarked images, are at <http://www.cl.cam.ac.uk/~fapp2/watermarking/benchmark/>.

Two general remarks apply to these tests. First, we did not take into account the computation time which is also an important parameters, especially for the extraction process. Second, some of the tools we have tested have already been improved. For instance the method of the University of Geneva now also addresses shearing using log-log maps[30]; this was not the case for the version we tested.

7. CONCLUSION

In this paper we addressed the issue of how to perform fair benchmarking and performance evaluation of digital watermarking methods. In a first part we have shown that for a fair comparison between different methods, the visual degradation of the images has to be taken into account. We have reviewed a variety of commonly used distortion metrics. The drawback of these distortion metrics is that they are not correlated to the HVS. We therefore presented another distortion metric which is adapted to the HVS and hence more suitable for digital watermarking. In addition the metric allows comparison even if the distortion is in a different color channel. Distortion metrics such as the $PSNR$ are not suitable for this because they give a distortion value for all color channels, for example Y, U and V. Then we looked at how to evaluate the performance of different watermarking methods in a research environment. Further work could try to improve this measure by taking into account possible minor non linear geometric distortions.[††] We have proposed four different graphs which can be used to evaluate individual performance and allow fair comparison between different methods. We also proposed the use of ROC graphs, which are very useful in assessing the overall statistical detection behavior of watermarking methods. The usefulness of all graphs has been demonstrated by comparing two different watermarking methods.

As mentioned, the introduced performance evaluation is very useful in a scientific environment since one needs full access to the algorithms and their parameters. However in a commercial environment this is often not the case. We therefore propose a generic benchmark test which can be used to evaluate watermarking methods without going into technical details. The benchmark tests the robustness of the watermarking methods using a variety of attacks and distortions. The result is a single number between 0 and 20 which describes the overall performance of the

[††]Such measure could be based on the $GSSNR$ as it is block-based and these minor distortions would not affect the blocks very much.

methods. The higher the number, the better the performance. However it is important to note that even for this benchmarking the distortion has to be taken into account. This means that the methods should be parametrised such that the visual degradation is the same for all tested methods. As a basis for our benchmark and in order to limit the number of attacks to be tested, we enhanced SirMark[‡‡] to include a set of pre-defined typical attacks (rotation, scaling, colour quantisation, etc.) and better random geometric distortions.[33]

Furthermore it is important that all tests are run several times, using different keys and different images. We therefore propose a set of test-images to be used for the evaluation of watermarking methods. These image are freely usable for research purpose only as long as credit is given to the artist.

REFERENCES

1. Ross J. Anderson. Why cryptosystems fail. *Communications of the ACM*, 37(11):32–40, November 1994.
2. Ross J. Anderson, editor. *Information hiding: first international workshop*, volume 1174 of *Lecture Notes in Computer Science*, Isaac Newton Institute, Cambridge, England, May 1996. Springer-Verlag, Berlin, Germany, ISBN 3-540-61996-8.
3. David Aucsmith, editor. *Information Hiding: Second International Workshop*, volume 1525 of *Lecture Notes in Computer Science*, Portland, Oregon, USA, 1998. Springer-Verlag, Berlin, Germany, ISBN 3-540-65386-4.
4. R. Barnett and D. E. Pearson. Frequency mode LR attack operator for digitally watermarked images. *Electronics Letters*, 34(19):1837–1839, September 1998.
5. Mauro Barni, Franco Bartolini, Vito Cappellini, and Alessandro Piva. A DCT-domain system for robust image watermarking. *Signal Processing*, 66(3):357–372, May 1998. European Association for Signal Processing (EURASIP).
6. Anonymous (<zguan.bbs@bbs.ntu.edu.tw>). Learn cracking IV – another weakness of PictureMarc. <news:tw.bbs.comp.hacker> mirrored on <http://www.cl.cam.ac.uk/~fapp2/watermarking/image_watermarking/digimarc_crack.html>, August 1997. Includes instructions to override any Digimarc watermark using PictureMarc.
7. Gordon W. Braudaway. Results of attacks on a claimed robust digital image watermark. In van Renesse.[48] ISBN 0-8194-2556-7, ISSN 0277-786X.

[‡‡]StirMark was originally written by Markus G. Kuhn to test robustness against bilinear random geometric distortions.

8. G. Caronni. Ermitteln unauthorisierter verteiler von maschinenlesbaren daten. Technical report, ETH Zürich, Switzerland, August 1993.

9. Germano Caronni. Assuring ownership rights for digital images. In H.H. Brüggermann and W. Gerhardt-Häckl, editors, *Reliable IT Systems (VIS'95)*, pages 251–263. Vieweg Publishing Company, Germany, 1995.

10. Ingemar J. Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. A secure, robust watermark for multimedia. In Anderson,[2] pages 183–206. ISBN 3-540-61996-8.

11. Ingemar J. Cox and Matt L. Miller. A review of watermarking and the importance of perceptual modeling. In Rogowitz and Pappas.[38] ISBN 0-8194-2427-7, ISSN 0277-786X.

12. J. F. Delaigle, C. De Vleeschouwer, and B. Macq. Watermarking algorithm based on a human visual model. *Signal Processing*, 66(3):319–335, May 1998. European Association for Signal Processing (EURASIP).

13. Jana Dittmann, Petra Wohlmacher, Patrick Horster, and Ralf Steinmetz, editors. *Multimedia and Security – Workshop at ACM Multimedia'98*, volume 41 of *GMD Report*, Bristol, United Kingdom, September 1998. ACM, GMD – Forschungszentrum Informationstechnik GmbH.

14. Ahmet M. Eskicioglu and Paul S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communication*, 43(12):2959–2965, December 1995.

15. Frank Hartung and Bernd Girod. Watermarking of uncompressed and compressed video. *Signal Processing*, 66(3):283–301, May 1998. European Association for Signal Processing (EURASIP).

16. Alexander Herrigel, Joseph J. K. Ó Ruanaidh, Holger Petersen, Shelby Pereira, and Thierry Pun. Secure copyright protection techniques for digital images. In Aucsmith,[3] pages 169–190. ISBN 3-540-65386-4.

17. Marty Katz. Digital watermarks often fail on Web images. The New York Times, 11 November 1997.

18. E. Koch and J. Zhao. Towards robust and hidden image copyright labeling. In *Workshop on Nonlinear Signal and Image Processing*, pages 452–455, Neos Marmaras, Greece, June 1995. IEEE.

19. Deepa Kundur and Dimitrios Hatzinakos. A robust digital image watermarking method using wavelet-based fusion. In *International Conference on Image Processing*, pages 544–547, Santa Barbara, California, USA, October 1997. IEEE. ISBN 0-8186-8183-7.

20. Deepa Kundur and Dimitrios Hatzinakos. Digital watermarking using multiresolution wavelet decomposition. In *International Conference on Acoustic, Speech and Signal Processing (ICASP)*, volume 5, pages 2969–2972, Seattle, Washington, USA, May 1998. IEEE.

21. Martin Kutter. Watermarking resisting to translation, rotation, and scaling. In *Proceedings of SPIE International Symposium on Voice, Video, and Data Communications*, November 1998.

22. Martin Kutter, F. Jordan, and Frank Bossen. Digital watermarking of color images using amplitude modulation. *Journal of Electronic Imaging*, 7(2):326–332, April 1998.

23. Gerrit C. Langelaar, Jan C. A. van der Lubbe, and Reginald L. Lagendijk. Robust labeling methods for copy protection of images. In Ishwar K. Sethin and Ramesh C. Jain, editors, *Storage and Retrieval for Image and Video Database V*, volume 3022, pages 298–309, San Jose, California, USA, February 1997. The Society for Imaging Science and Technology (IS&T) and the International Society for Optical Engineering (SPIE), SPIE. ISBN 0-8194-2433-1, ISSN 0277-786X.

24. Jean-Paul M. G. Linnartz and Marten van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In Aucsmith,[3] pages 258–272. ISBN 3-540-65386-4.

25. Maurice Maes. Twin peaks: The histogram attack on fixed depth image watermarks. In Aucsmith,[3] pages 290–305. ISBN 3-540-65386-4.

26. Gianluca Nicchiotti and Ennio Ottaviano. Non-invertible statistical wavelet watermarking. In *9th European Signal Processing Conference (EUSIPCO'98)*, pages 2289–2292, Island of Rhodes, Greece, 8–11 September 1998. ISBN 960-7620-05-4.

27. N. Nikolaidis and I. Pitas. Robust image watermarking in the spatial domain. *Signal Processing*, 66(3):385–403, May 1998. European Association for Signal Processing (EURASIP).

28. Paulo Roberto Rosa Lopes Nunes, Abraham Alcaim, and Mára Regina Labuto Fragoso da Silva. Quality measures of compressed images for classification purposes. Technical Report CCR-146, IBM Brasil, Rio Scientific Center, P.O. Box 4624, 20.0001 Rio de Janeiro, Brazil, October 1992.

29. Joseph J. K. Ó Ruanaidh and Thierry Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Processing*, 66(3):303–317, May 1998. European Association for Signal Processing (EURASIP).

30. Shelby Pereira, Joseph J. K. O'Ruanaidh, Frédéric Deguillaume, Gabriella Csurka, and Thierry Pun. Template based recovery of fourier-based watermarks using log-polar and log-log maps. In *International Conference on Multimedia Systems (ICMS'99)*, Firenze, 7–1 June 1999. IEEE. To appear.

31. Adrian Perrig. A copyright protection environment for digital images. Diploma dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, February 1997.

32. Fabien A. P. Petitcolas. Weakness of existing watermarking schemes. `<http://www.cl.cam.ac.uk/~fapp2/watermarking/image_watermarking/>`, October 1997.

33. Fabien A. P. Petitcolas and Ross J. Anderson. Weaknesses of copyright marking systems. In Dittmann et al.,[13] pages 55–61.

34. Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on copyright marking systems. In Aucsmith,[3] pages 218–238. ISBN 3-540-65386-4.

35. Fabien A. P. Petitcolas and Markus G. Kuhn. StirMark 2. `<http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/>`, November 1997.

36. Christine I. Podilchuk and Wenjun Zeng. Digital image watermarking using visual models. In Rogowitz and Pappas,[38] pages 100–111. ISBN 0-8194-2427-7, ISSN 0277-786X.

37. Stéphane Roche and Jean-Luc Dugelay. Mécanismes de sécurit liés la transmission des images. In *COompression et REprsentation des Signaux Audiovisuels (CORESA'97)*, Issy-les-Moulineaux, France, March 1997.

38. Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors. *Human Vision and Electronic Imaging II*, volume 3016, San Jose, California, USA, February 1997. The Society for Imaging Science and Technology (IS&T) and the International Society for Optical Engineering (SPIE), SPIE, ISBN 0-8194-2427-7, ISSN 0277-786X.

39. Khalid Sayood. *Introduction to Data Compression*, chapter 7, page 142. Morgan Kaufmann Publishers, 1996.

40. Mitchell D. Swanson, Mei Kobayashi, and Ahmed H. Tewfik. Multimedia data-embedding and watermarking technologies. *Proceedings of the IEEE*, 86(6):1064–1087, June 1998.

41. Mitchell D. Swanson, Bin Zu, and Ahmed H. Tewfik. Robust data hiding for images. In *7th Digital Signal Processing Workshop (DSP 96)*, pages 37–40, Loen, Norway, September 1996. IEEE.

42. K. Tanaka, Y. Nakamura, and K. Matsui. Embedding secret information into a dithered multilevel image. In *Proceeding of the 1990 IEEE Military Communications Conference*, pages 216–220, September 1990.

43. K. Tanaka, Y. Nakamura, and K. Matsui. Embedding the attribute information into a dithered image. *Systems and Computers in Japan*, 21(7), 1990.

44. A. Z. Tirkel, C. F. Osborne, and T. E. Hall. Image and watermark registration. *Signal Processing*, 66(3):373–383, May 1998. European Association for Signal Processing (EURASIP).

45. A. Z. Tirkel, G. A. Rankin, R. M. van Schyndel, W. J. Ho, N. R. A. Mee, and C. F. Osborne. Electronic watermark. In *Digital Image Computing, Technology and Applications (DICTA'93)*, pages 666–673, Macquarie University, Sidney, 1993.

46. Dimitrios Tzovaras, Nikitas Karagiannis, and Michael G. Strintzis. Robust image watermarking in the subband or discrete cosine transform domain. In *9th European Signal Processing Conference (EUSIPCO'98)*, pages 2285–2288, Island of Rhodes, Greece, 8–11 September 1998. ISBN 960-7620-05-4.

47. Christian J. van den Branden Lambrecht and Joyce E. Farrell. Perceptual quality metric for digitally coded color images. In *Proceeding of EUSIPCO*, pages 1175–1178, Trieste, Italy, September 1996.

48. Rudolf L. van Renesse, editor. *Optical Security and Counterfeit Deterrence Techniques II*, volume 3314, San Jose, California, USA, January 1998. The Society for Imaging Science and Technology (IS&T) and the International Society for Optical Engineering (SPIE), SPIE, ISBN 0-8194-2556-7, ISSN 0277-786X.

49. R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne. A digital watermark. In *International Conference on Image Processing*, volume 2, pages 86–90, Austin, Texas, USA, 1994. IEEE.

50. G. Voyatzis, N. Nikolaidis, and I. Pitas. Digital watermarking: an overview. In *9th European Signal Processing Conference (EUSIPCO'98)*, pages 9–12, Island of Rhodes, Greece, 8–11 September 1998. ISBN 960-7620-05-4.

51. Allan G. Weber. The usc-sipi image databse: Version 5. `<http://sipi.usc.edu/services/database/Database.html>`, October 1997. Singal and Image Processing Institute at the University of Southern California.

52. S. J. P. Westen, R. L. Lagendijk, and J. Biemond. Perceptual image quality based on a multiple channel HVS model. In *Proceeding of ICASP*, volume 4, pages 2351–2354, 1995.

53. Stefan Winkler. A perceptual distortion metric for digital color images. In *Proc. ICIP*, volume 3, pages 399–403, Chicago, IL, October 1998.

54. Stefan Winkler. A perceptual distortion metric for digital color video. In *SPIE Proceedings of Human Vision and Electronic Imaging*, volume 3644, San Jose, CA, January 1999.

55. Raymond B. Wolfgang and Edward J. Delp. A watermark for digital images. In *International Conference on Images Processing*, pages 219–222, Lausanne, Switzerland, September 1996. IEEE.

56. Raymond B. Wolfgang and Edward J. Delp. A watermarking technique for digital imagery: further studies. In *International Conference on Imaging, Systems, and Technology*, pages 279–287, Las Vegas, Nevada, USA, 30 June–3 July 1997. IEEE.

57. J. Zhao and E. Koch. Embedding robust labels into images for copyright protection. In *International Congress on Intellectual Property Rights for Specialised Information, Knowledge and New Technologies*, Vienna, Austria, August 1995.

58. Mark H. Zweig and Gregory Campbell. Receiver–operating characteristics (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577, 1993.

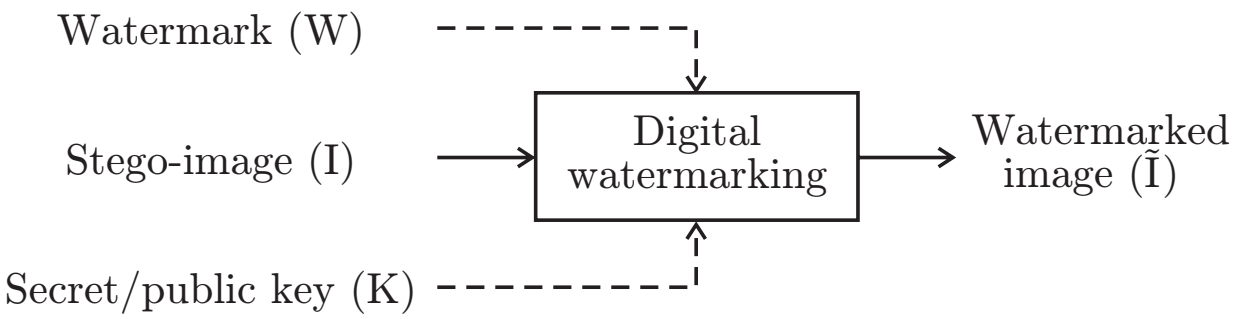

Figure 1. Generic digital watermark embedding scheme.

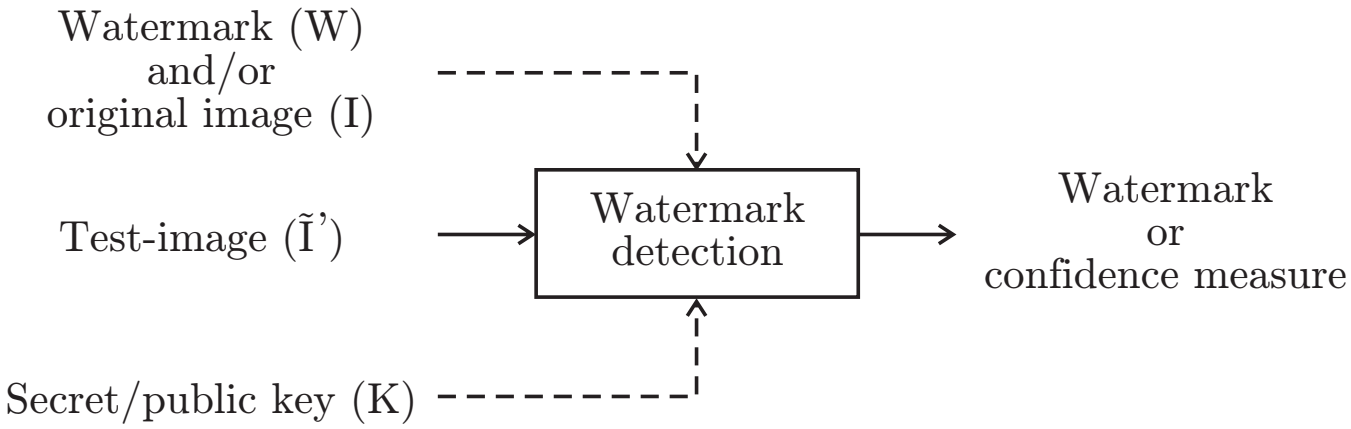

Figure 2. Generic digital watermark recovery scheme.

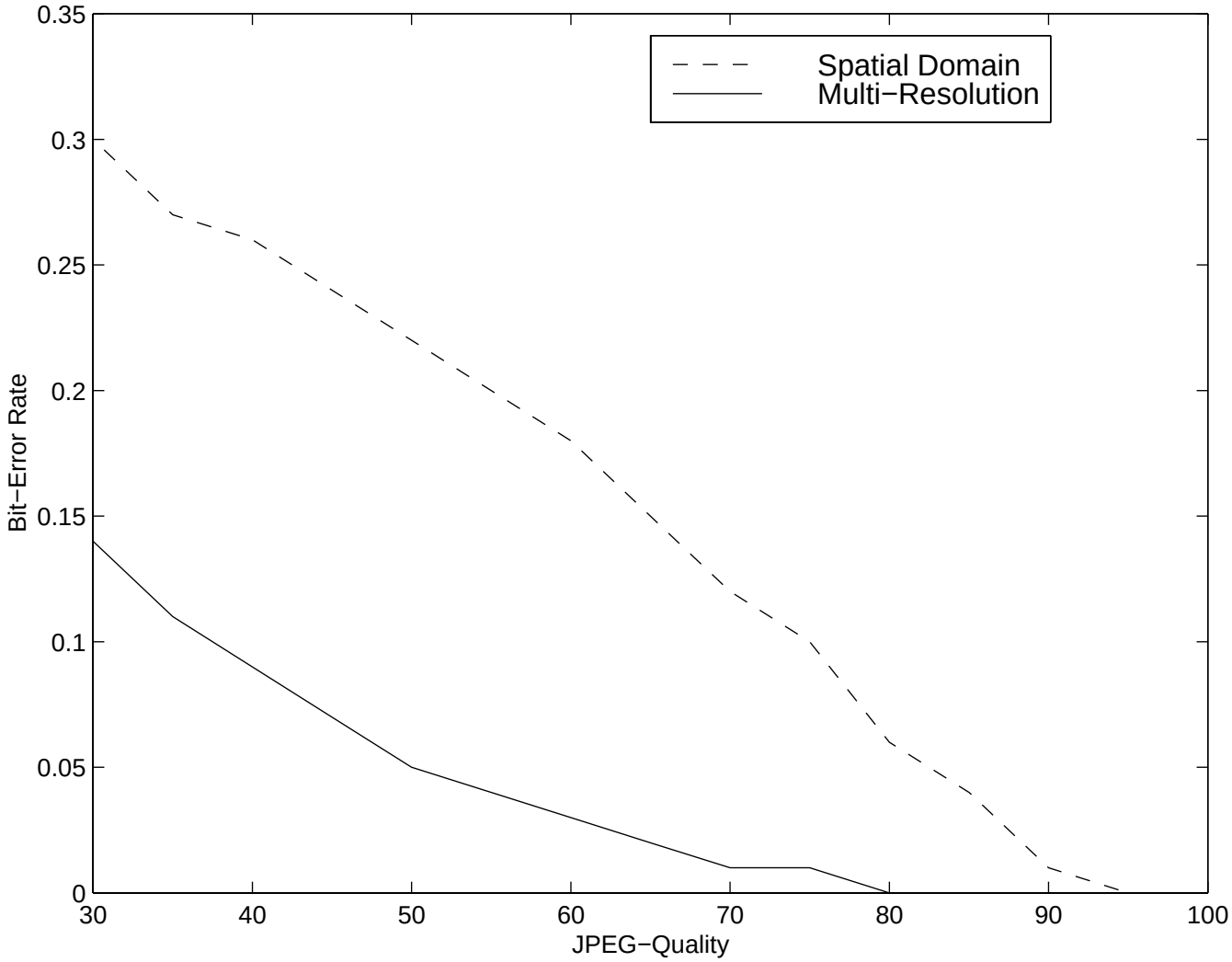

Figure 3. Bit-error vs. attack graph for spread-spectrum modulation in a spatial and multi-resolution environment. The visual quality was fixed to 4.5. It is clearly visible, that the multi-resolution approach has a higher robustness.

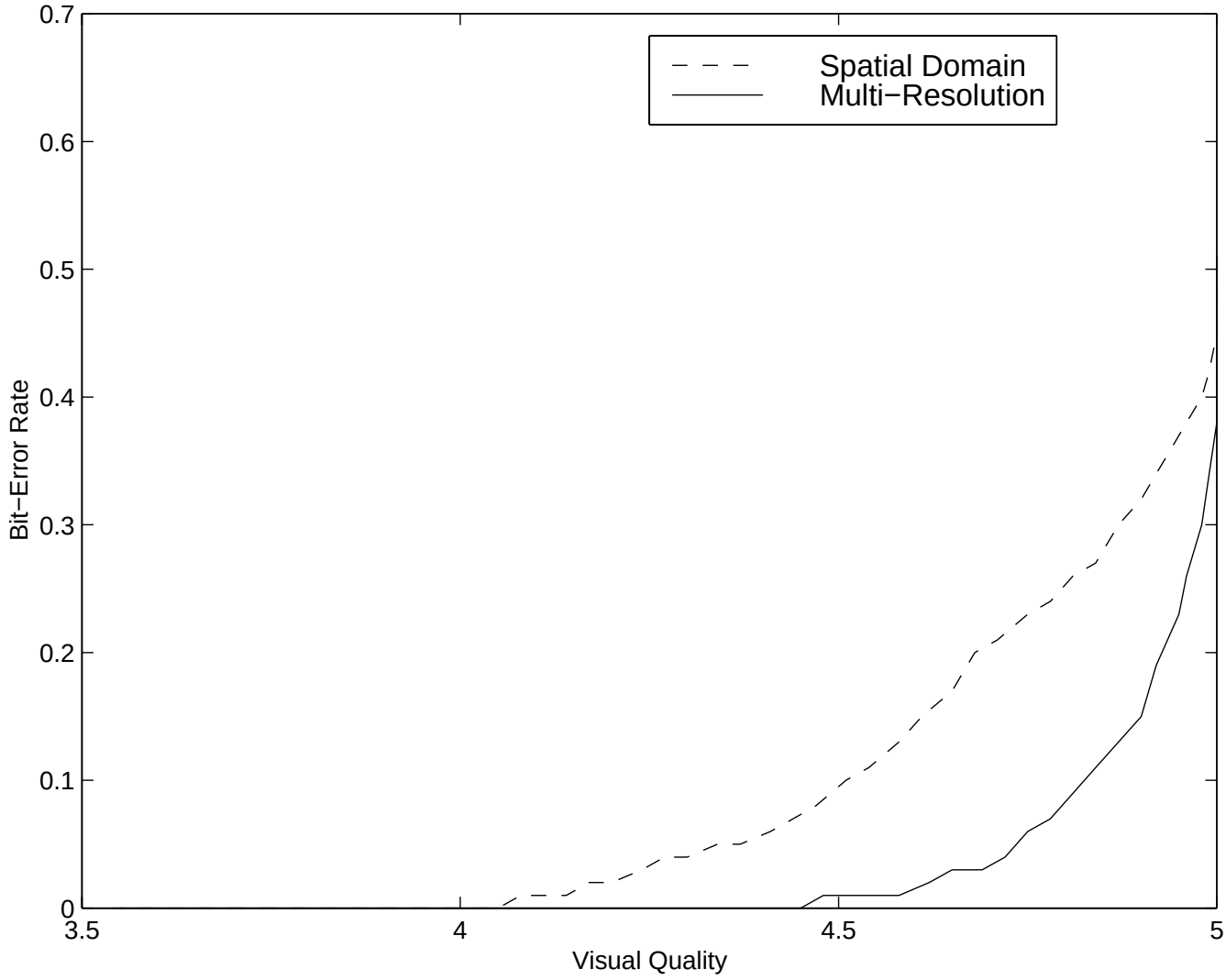

Figure 4. Bit-error vs. visual quality graph for spread-spectrum modulation in a spatial and multi-resolution environment. The attack was fixed to JPEG compression with 75% quality. Again the multi-resolution approach shows superior performance.

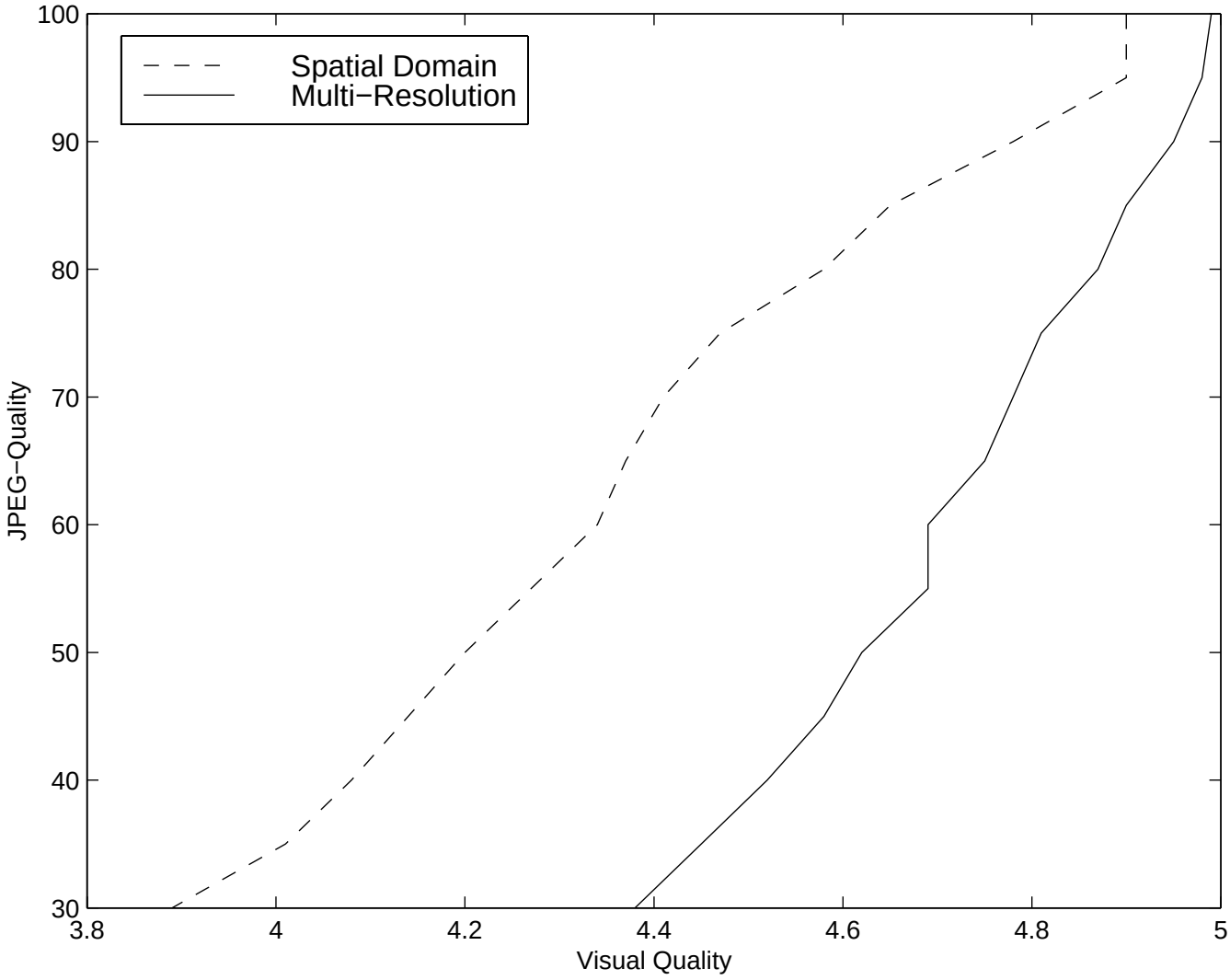

Figure 5. Attack vs. visual quality graph for spread spectrum modulation in a spatial and multi-resolution environment. The bit-error rate was set to 0.1. The curves clearly show that the multi-resolution approach accommodates larger compression ratios for a given visual quality.

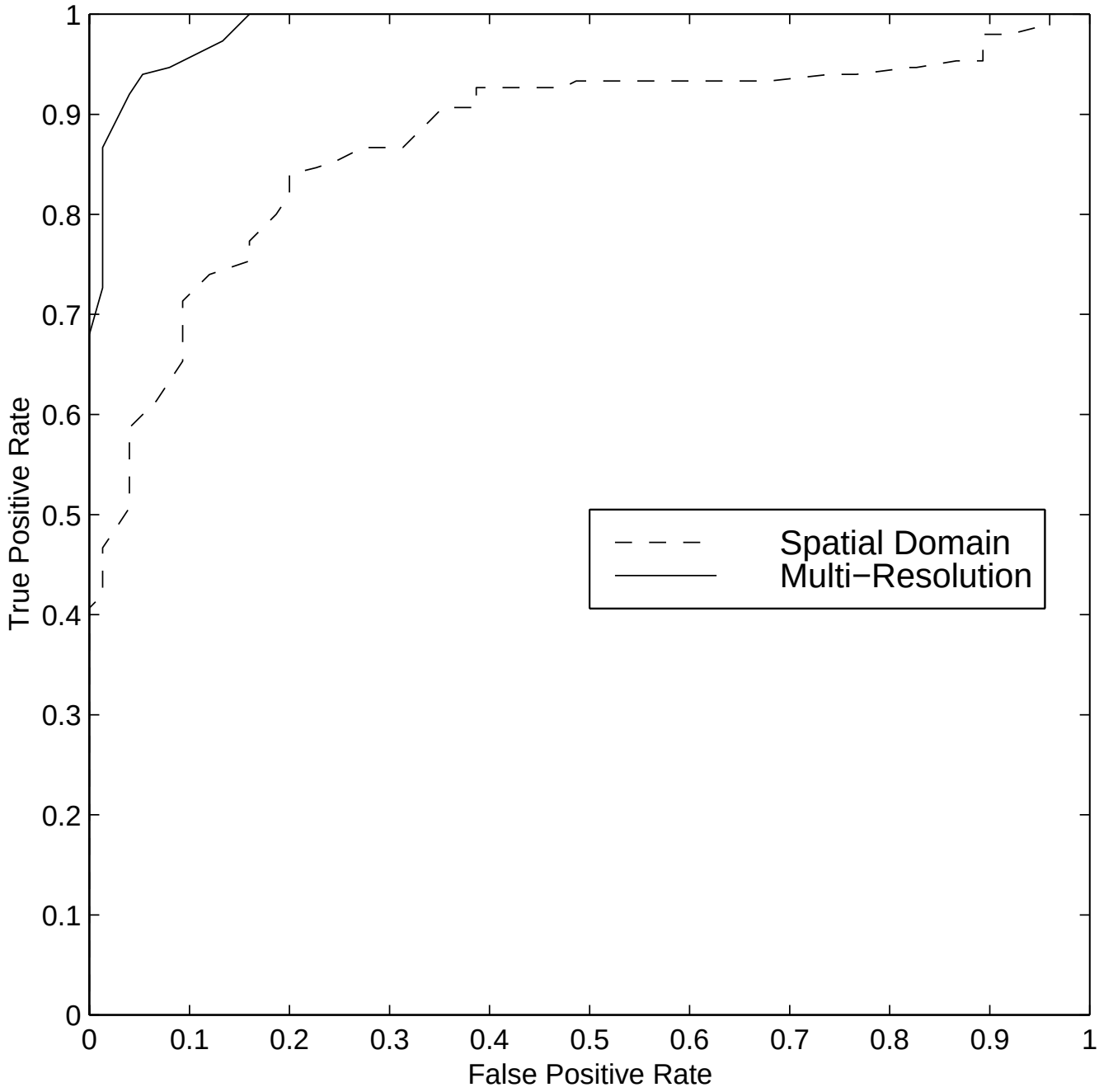

Figure 6. ROC graph for spread spectrum modulation in a spatial and multi-resolution environment. The curve corresponding to the multi-resolution approach is closer to the top left corner, which indicates it superior performance.

Difference Distortion Metrics

Maximum Difference	$MD = \max_{m,n} \lvert I_{m,n} - \tilde{I}_{m,n} \rvert$
Average Absolute Difference	$AD = \frac{1}{MN} \sum_{m,n} \lvert I_{m,n} - \tilde{I}_{m,n} \rvert$
Norm. Average Absolute Difference	$NAD = \sum_{m,n} \lvert I_{m,n} - \tilde{I}_{m,n} \rvert / \sum_{m,n} \lvert I_{m,n} \rvert$
Mean Square Error	$MSE = \frac{1}{MN} \sum_{m,n} \left(I_{m,n} - \tilde{I}_{m,n} \right)^2$
Normalised Mean Square Error	$NMSE = \sum_{m,n} \left(I_{m,n} - \tilde{I}_{m,n} \right)^2 / \sum_{m,n} I_{m,n}^2$
L^p-Norm	$L^p = \left(\frac{1}{MN} \sum_{m,n} \lvert I_{m,n} - \tilde{I}_{m,n} \rvert^p \right)^{1/p}$
Laplacian Mean Square Error	$LMSE = \sum_{m,n} \left(\nabla^2 I_{m,n} - \nabla^2 \tilde{I}_{m,n} \right)^2 / \sum_{m,n} (\nabla^2 I_{m,n})^2$
Signal to Noise Ratio	$SNR = \sum_{m,n} I_{m,n}^2 / \sum_{m,n} \left(I_{m,n} - \tilde{I}_{m,n} \right)^2$
Peak Signal to Noise Ratio	$PSNR = MN \max_{m,n} I_{m,n}^2 / \sum_{m,n} \left(I_{m,n} - \tilde{I}_{m,n} \right)^2$
Image Fidelity	$IF = 1 - \sum_{m,n} \left(I_{m,n} - \tilde{I}_{m,n} \right)^2 / \sum_{m,n} I_{m,n}^2$

Correlation Distortion Metrics

Normalised Cross-Correlation	$NC = \sum_{m,n} I_{m,n} \tilde{I}_{m,n} / \sum_{m,n} I_{m,n}^2$
Correlation Quality	$CQ = \sum_{m,n} I_{m,n} \tilde{I}_{m,n} / \sum_{m,n} I_{m,n}$

Others

Structural Content	$SC = \sum_{m,n} I_{m,n}^2 / \sum_{m,n} \tilde{I}_{m,n}^2$
Global Sigma Signal to Noise Ratio	$GSSNR = \sum_b \sigma_b^2 / \sum_b \left(\sigma_b - \tilde{\sigma}_b \right)^2$ where $\sigma_b = \sqrt{\frac{1}{P} \sum_{\text{block}b} I_{m,n}^2 - \left(\frac{1}{P} \sum_{\text{block}b} I_{m,n} \right)^2}$
Sigma Signal to Noise Ratio	$SSNR = \frac{1}{P} \sum_b SSNR_b$ where $SSNR_b = 10 \log_{10} \frac{\sigma_b^2}{(\sigma_b - \tilde{\sigma}_b)^2}$
Sigma to Error Ratio	$SER_b = \frac{\sigma_b^2}{\frac{1}{P} \sum_{\text{block}b} (I_{m,n} - \tilde{I}_{m,n})^2}$
Histogram Similarity	$HS = \sum_{c=0}^{255} \lvert f_I(c) - f_{\tilde{I}}(c) \rvert$ where $f_I(c)$ is the relative frequency of level c in a 256 levels image.

Note: $\nabla^2 I_{m,n} = I_{m+1,n} + I_{m-1,n} + I_{m,n+1} + I_{m,n-1} - 4I_{m,n}$

Table 1. Commonly used pixel based visual distortion metrics. $I_{m,n}$ represents a pixel, whose coordinates are (m,n), in the original, undistorted image, and $\tilde{I}_{m,n}$ represents a pixel, whose coordinates (m,n), in the watermarked image. $GSSNR$, $SSNR$ and SER require the division of the original and watermarked images into B blocks of P pixels (e.g., 4×4 pixels). More details are given in Nunes.[28]

Rating	*Impairment*	*Quality*
5	Imperceptible	Excellent
4	Perceptible, not annoying	Good
3	Slightly annoying	Fair
2	Annoying	Poor
1	Very annoying	Bad

Table 2. ITU-R Rec. 500 Quality ratings on a scale from 1 to 5.

	Parameter			
Graph Type	*Visual Quality*	*Robustness*	*Attack*	*Bits*
Robustness vs. attack	fixed	variable	variable	fixed
Robustness vs. visual quality	variable	variable	fixed	fixed
Attack vs. visual quality	variable	fixed	variable	fixed
ROC	fixed	fixed	fixed / variable	fixed

Table 3. Different graphs and corresponding variables and constants.

	Digimarc	**SureSign**	**EikonaMark**	**Giovanni**	**SysCoP**
	1.51	**3.0 Demo**	**3.01**	**1.1.0.2**	**1.0R1**
Filtering (3×3 median, Gaussian)	100	100	100	60	80
Scaling (0.5, 0.75, 0.9, 1.1, 1.5, 2)	70	100	0	63	0
Cropping (1, 2, 5, 10, 15, 20, 25, 50 %)	100	100	0	15	0
Rotation (-2, -1, -0.5, 0.5, 1, 2)	82	58	0	10	0
JPEG (90, 85, 80, 75, 60, 50, 25, 15, 10, 5)	56	72	90	12	58
GIF Conversion	100	100	100	60	80
Horizontal flip	100	100	0	0	0
StirMark 1.0	80	80	0	0	0
StirMark 2.2	0	0	0	0	0

Table 4. Early robustness tests (August 1998) for various digital watermarking products. Values are percentage of survival to attack. For each product 5 test images (baboon, benz, girl, glasses, and lena) have been used and for each image 42 transformations have been applied using StirMark 2. Each image has been watermarked using the best parameters that do not give obvious and annoying distortions. Although comparison should be done with great care (not all systems have the same applications, some systems are public other semi-private, etc.), the table confirms what is currently achieved and what needs further research.

	Digimarc	Unige	SureSign
Signal enhancement			
Gaussian	100	100	100
Median	100	100	100
Sharpening	100	100	100
FMLR	100	67	100
Compression			
JPEG	65	52	87
GIF/Colour quantisation	100	100	100
Scaling			
Without JPEG 90	81	81	97
With JPEG 90	72	81	83
Cropping			
Without JPEG 90	100	81	94
With JPEG 90	98	72	91
Shearing			
X	50	13	42
Y	50	4	42
Rotation			
Auto-crop	95	74	37
Auto-scale	97	77	51
Other geometric trans.			
Col. & line removal	100	69	89
Horizontal flip	100	100	100
Random geometric dist.	17	0	0

Table 5. Summary of the results for the benchmark presented in this paper. We tested the following piece of software: Digimarc's Batch Embedding Tool 1.00.13, Digimarc's ReadMarc 1.5.8, the watermarking tool of the University of Geneva (version 15 January 1999) and Signum Technologies' SureSign Server 1.94. The partition in the table means that the conditions of the experiments were slightly different for SureSign as explained in the body of this paper.