Digital watermarking of still images in the presence of de-synchronization attacks: theoretical analysis and development of practical algorithms

WP2: Development of practical algorithms

M. Barni
University of Siena

N. Merhav
Israel Institute of Technology

Sponsored by

MiUR
Under contract no RBIN04AC9W

WP2.2 Second report on the development of practical algorithms
10 September 2008
Preface

This report is a living document that is updated at the end of each reporting period to describe the activity carried out within the practically-oriented work-package of the FIRB project no RBIN04AC9W. The final version, that will be ready at the end of the third year, will thus contain a comprehensive and self-contained description of the main theoretical results obtained during the project. A similar report will account for the theoretical part of the project (WP1).

The present version of the report report (WP2.2), details the theoretical results obtained during the first 2 years of the project. As such it updates the report WP2.1 that accounted for the activity of the first year.

The structure of the report closely follows the structure of WP2 (devoted to the development of practical algorithms) of the project. Each chapter of the report corresponds to a task of WP2. This report is written so to be as self-contained as possible, however for a full understanding of the addressed issues it is suggested that it is read after reading the report about the theoretical investigation. For each chapter (section) the project year during which the activity described within the chapter (section) has been obtained is reported, so to allow an easy understanding of the temporal development of the project.

More specifically, the report is split into 4 chapters. Chapter 1 is dedicated to the validation of the DA models (namely LPCD and MRF) developed in WP1 against psychovisual experiments.

Chapter 2 is dedicated to the validation of the optimum embedding and detection strategies developed in WP1 on synthetic data.

In Chapter 3 the theoretical findings are tested against real data, e.g. natural images. In particular, the activity carried out so far, focused on the measurement of the actual de-synchronization capabilities of the geometric attacks studied within the project when they are applied to natural images watermarked with some of the most popular watermarking schemes proposed so far.

Finally, Chapter 4 considers the development of new watermarking systems based on the theory developed within the project. This activity will constitute the core of the third year of the project, however some work has already been done aiming at developing a perpetually sound quality metric to be used to control the perceptually admissibility of the displacement field used to register the watermark and the to-be-inspected images.
T2.1 - Validation of DA characterization against perceptual analysis
Project years: Y1,Y2

Abstract. The perceptual impact of the geometric distortions introduced by the multiresolution version of the C-LPCD model is evaluated by means of psychovisual experiments. Specifically, the maximum admissible distortion that can be applied before the distortion becomes visible is measured leading to the definition of the perceptually admissible subset of C-LPCD distortions. A similar analysis is performed with regard to the MRF model.

1 Introduction

Despite the existence of several studies on human perception of image quality, research has not dealt with geometric distortions. As we know this is a fundamental problem in digital watermarking when the presence of de-synchronization attacks must be taken into account. Specifically, given a distortion model, it would be useful to know the amount of geometric distortions that can be applied before the distorted image loses its commercial value or its meaning in order to develop an ad-hoc decoding algorithm and eventually obtain watermark synchronization through exhaustive search [1].

In addition to watermarking and de-synchronization attacks, other possible applications of this analysis are, for example, registration of biomedical images, that usually requires local and nonlinear transformations [2, 3], or collusion-secure fingerprinting techniques by random pre-warping [4]. In these last methods the host signal is randomly warped prior to watermarking in such a way to prevent traitors from obtaining a high-quality copy through collusion. The random pre-warping must be strong enough to avoid that registration techniques can undo the warping and, in the meantime, it must guarantee the invisibility of the distortion. For this reason gathering information about the subset of perceptually admissible geometric distortions is a vital requirement.

In the following section, the de-synchronization parameters allowing to maximize the geometric distortion while keeping the quality of the distorted images at an acceptable level are derived for both the LPCD (2) and the MRF (3) models.

2 The LPCD class of geometric attacks (year: Y1)

Though the class of LPCD DA is described in section T1.2 of the WP1 report, we briefly summarize it here.

By focusing on the 1D-case, let $y = \{y(1), y(2) \ldots y(n)\}$ be a generic signal and let $z = \{z(1), z(2)\ldots z(n)\}$ be the distorted version of $y$. Let $z(i)$ be a generic element of $z$, the LPCD model states that $z(i) = y(i + \Delta_i)$ where $\Delta_i$ is a sequences of i.i.d random variables uniformly distributed in a predefined interval $I = [-\Delta_{\text{max}}, \Delta_{\text{max}}]$. For simplicity we assume that $\Delta_i$ can take only integer values in $I$. To extend the model to the 2-D case, if $Z(i, j)$ is a generic pixel of the distorted image $Z$, we have $Z(i, j) = Y(i + \Delta_h(i, j), j + \Delta_v(i, j))$, where $Y$ is the original image and $\Delta_h(i, j)$ and $\Delta_v(i, j)$ are sequences of i.i.d integer random variables uniformly distributed in the interval $[-\Delta_{\text{max}}, \Delta_{\text{max}}]$.  

An important limit of the LPCD model is the lack of memory. The lack of memory is likely to be a problem from a perceptual point of view: with no constraints on the smoothness of the displacement field there is no guarantee that the set of LPCD distortions is perceptually admissible even by constraining $\Delta_{\text{max}}$ to be very small. This statement is confirmed by Figure (1).
Fig. 1. Local Permutation with Cancelation and Duplication (LP-CD) with $\Delta_{\text{max}} = 2$: (a) Lena; (b) Duomo.

2.1 Constrained LPCD (C-LPCD)

A way to overcome the limit of the lack of memory of the LP-CD model, in order to obtain better results from a perceptual point of view, is to require that the sample order, in the 1D case, is preserved (thus introducing memory in the system). In practice, the displacement of each element $i$ of the distorted sequence $z$ is conditioned on the displacement of the element $i - 1$ of the same sequence. In formulas, $z(i) = y(i + \Delta_i)$ where $\Delta_i$ is a sequence of i.i.d integer random variables uniformly distributed in the interval $I = [\max (-\Delta_{\text{max}}, \Delta_{i-1} - 1), \Delta_{\text{max}}]$. The 2D extension is obtained by applying the 1D C-LPCD transformation first by rows, then by columns. Figure (2) shows the Duomo and the Lena images distorted with the C-LPCD model (considering $\Delta_{\text{max}} = 2$).

Fig. 2. Constrained LPCD (C-LPCD) with $\Delta_{\text{max}} = 2$: (a) Lena; (b) Duomo.
After a visual inspection conducted on the images we can deduce that constrained LPCD is more perceptually pleasant than the previous models. Actually, as it is described in section T1.2 of the WP1 report, by casting the C-LPCD model into the framework of Markov Chain theory we observe that the model described above has the undesirable tendency to prefer small displacements, thus weakening the strength of the attack. Then we considered an improved version of the C-LPCD model whereby small and large displacements are given approximately the same occurrence probability. The results discussed in the sequel always refers to the improved version of C-LPCD. For a detailed description of the improved C-LPCD model we refer to section T1.2 of the WP1 report.

2.2 Multiresolution Constrained LPCD (MC-LPCD)

To further improve the C-LPCD model, and to make the distortions less perceptible, we considered a last model, a Multiresolution version of C-LPCD (MC-LPCD). In this case the Constrained LPCD model is applied at different resolutions to obtain the global displacement field. A low resolution displacement field is first generated, then a full size displacement field is built by means of bilinear interpolation. The full resolution field is finally applied to the original image to produce the distorted image.

More specifically MC-LPCD consists of two steps. In the first step C-LPCD is applied at a low level of resolution to obtain the displacement fields $\delta_h(i,j)$ and $\delta_v(i,j)$ with size $S^2 \times S^2$ where $L$ is the level of resolution and $S$ is the size of the original image. The displacement fields $\delta_h$ and $\delta_v$ are resized to the original image size by means of bilinear interpolation to obtain the full resolution displacement fields $\Delta_h$ and $\Delta_v$. Note that in this phase the values of the displacement fields are not enlarged in order to make the distortions less perceptible and take advantage of the multiresolution model. For this reason new non-integer displacement values are introduced. The full resolution displacement fields $\Delta_h$ and $\Delta_v$ are then applied to $\mathbf{Y}$ through a bilinear interpolation of the gray levels to find the final distorted image $\mathbf{Z}$ (introducing in this way new gray levels in the images).

Figure (3) shows two sample images distorted with MC-LPCD applied at different resolutions levels: in (a) the level of resolution is 3, (b) is the case of $L = 4$, and in figure (c) the level of resolution is 6 (in all the cases $\Delta_{\max} = 2$).

By observing the last figures it is clear that this model provides a good way to incorporate perceptual considerations within the model. In particular the image quality increases, from a perceptual point of view, if the MC-LPCD model is applied to a lower level of resolution but, in the meantime, the number of possible distortions decreases.

2.3 The psycho-visual experiments

Given the good performance of the MC-LPCD model, we carried out a set of objective and subjective tests to evaluate the model parameters that ensure the invisibility of the applied distortion. To do so, let us observe that from a perceptual point of view MC-LPCD has a different behavior for different values of $N$ and for different levels of resolution $L$. The goal of the tests was to establish the sensitivity of the visual system to the geometric distortions introduced by the model as a function of the control parameters $N$ and $L$. In this way we were able to identify the range of variation of the control parameters that do not affect image quality.

For objective testing, we used the PSNR measurement and state-of-the-art metrics such as the Universal Quality Index [5], the SSIM-index [6] and the RST based metric developed by I.Setyawan et al.[7]. It is worth mentioning that despite the great research effort, there is still a lack of objective visual quality metrics suitable for geometric distortions.

In the subjective test we applied the two alternatives forced choice (2AFC) paradigm: the users were asked to compare two images at time, the original image and the distorted image, and to indicate which one was the original.
Fig. 3. MC-LPCD applied at different resolution levels: (a) $L = 3$, (b) $L = 4$, (c) $L = 6$. 
The source image database used in both the tests included sixteen gray scale images, 512 × 512 pixel in size, and was derived from a set of source images that reflects adequate diversity in image contents. The images, in fact, included pictures of faces, houses, natural scenes and images without any specific object of interest. Some images have high activity, while some do not have much structures and are mostly smooth [8].

We chose to distort the source images through the MC-LPCD model using different distortion types obtained by changing the dimension of $N$ ($N = 5$, $N = 7$, and $N = 9$) and the level of resolution ($L = 6$, $L = 5$, $L = 4$, $L = 3$, $L = 2$). Specifically we produced thirteen distorted versions for each image (all the possible combinations of $N$ and $L$ except $N = 9$ $L = 2$ that always generates a visible distortions and $N = 9$ $L = 6$ because the level of resolution is smaller than $N$), for a total of 208 images.

The objective test Many image quality assessment algorithms have been shown to behave consistently when applied to certain kinds of distortions (e.g., JPEG compression), but the effectiveness of these metrics degrades when they are applied to a set of images distorted geometrically. However, just for completeness, we present the results we obtained by applying some of these metrics to the images used in the subjective test. For a given set of parameters, the comparison of the subjective test described in the following subsection with the objective quality measures will allow to establish the difficulty of objective metrics in predicting the perceived amount of geometrical distortions present in an image.

The results of the objective test are shown in figure (4). We do not show the results obtained with the Universal Quality Index because SSIM index it is an improvement of it.

![Figure 4](image-url)

**Fig. 4.** Plot of objective metrics versus the level of resolution used in the MC-LPCD model: (a) Peak Signal to Noise Ratio; (b) Structural Similarity based Image index.

We also applied the RST based metric developed by Setyawan [7] but the results we obtained are not meaningful. In our case, in fact, this metric is not able to predict the visual quality of the images because it does not find a local RST or affine transform (neither in a small interval) approximating the geometric distortion introduced by the MC-LPCD model.

The two alternative forced-choice test The subjective test we used is the two alternative forced-choice test (2AFC)[9]. Two stimuli are presented at each trial. One of these stimuli is the original image; the other one is a distorted version of it, and the observer is asked to select the original image.
Procedures for such experiment have been designed by following the ITU-T Recommendation P.910 [10], which suggests standard viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods.

The experiments were conducted by using the VP800 video card of the Cambridge Research Systems together with a high resolution digital monitor Mitsubishi DiamondPro 2070 with the external adaptor ViSaGe 71.02.00D2 [11]. To have a correct color representation a luminosity calibration was previously done through a videocamera ColorCAL [11].

The tests involved a panel of fifteen subjects, all naives with respect to image quality assessment methods and image impairments. Each subject was individually briefed about the goal of the experiment, and given a demonstration of the test. Subjects were shown images in a random order, the randomization was different for each subject. The test was performed in a dark room in free viewing conditions.

In order to analyze the results obtained with the subjective test an hypothesis test was conducted. This test tells us whether the subjective test, based on the number of sample points used, allows to make a statistically sound conclusion about the behavior of the MC-LPCD model. The test statistic on which the hypothesis test is based is a probability test: we tested the hypothesis $H_0$ that the probability $p = P(A)$ of an event $A$ equals a given constant $p_0$, using as data the number $k$ of successes of $A$ in $n$ trials. Specifically we tested the hypothesis that the probability $p$ that the users choose the original image in the 2AFC test (event $A$) is equal to $p_0 = \frac{1}{2}$ (it means that the original image and the distorted image are perceptually indistinguishable) using as observable data the number of times that the users decided for the original images in the 240 total comparisons ($16$ images $\times 15$ users).

The alternative hypothesis is that $P(A) < \frac{1}{2}$ and we use $\alpha = 0.05$ as significativity level [12].

<table>
<thead>
<tr>
<th>$L=6$</th>
<th>$L=5$</th>
<th>$L=4$</th>
<th>$L=3$</th>
<th>$L=2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>faces</td>
<td>N=7</td>
<td>N=5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>houses</td>
<td>N=7</td>
<td>N=5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>nature</td>
<td>N=9</td>
<td>N=7</td>
<td>N=5</td>
<td>-</td>
</tr>
<tr>
<td>others</td>
<td>N=7</td>
<td>N=5</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1. Maximum admissible distortion that can be applied to the images before the distortion becomes visible using the MC-LPCD model.

With the hypothesis test we derived the maximum admissible distortion that can be applied before the distortion becomes visible using the MC-LPCD model. Specifically, the results of the hypothesis test are shown in table (1). For each level of resolution and for each typology of images, we found the maximum value of $N$ that can be used while keeping the distortion invisible (the hypothesis $p_0 = \frac{1}{2}$ is accepted). The empty boxes correspond to cases in which it was not possible to find an adequate value of $N$, corresponding to that level of resolution, resulting in an invisible distortion. By looking at the results we note that, as expected, in images that do not have much structures, like natural images, the introduced distortions are less visible. Furthermore the results in table (1) show the difficulty of the objective metrics in predicting the degradation introduced by the model in the images: for example the plot $N = 5$ in figure (4.a) shows a loss of quality going from $L = 6$ to $L = 5$ while the subjective test assures the invisibility of the distortion in both cases. In the same way the loss of about 5 dB in the PSRN plot between the configurations $N = 5$, $L = 6$ and $N = 7$, $L = 6$ is not meaningful according table (1).

Remarks One of the main result of the experiments we carried out to validate the MC-LPCD model was the definition of an experimental methodology to be used throughout the rest of the project. Specifically, the next steps will be the evaluation of the visibility of the geometric artifacts introduced by the second DA class we developed (namely the MRF-DA class of attacks - see section T1.2 of
WP.1), and the development of a metric capable of measuring in an automatic way the annoyance of a geometric distortion.

The research carried out so far resulted in the following publications.


3 The MRF class of geometric attacks (year: Y2)

The exact definition of the MRF class of DA is described in section T1.2 of the WP1 report, however for sake of completeness we briefly summarize again it here.

3.1 The MRF class of attacks

A random field \( F = \{F_1, F_2, ..., F_m\} \) is a family of random variables defined on a set \( S \), in which each random variable \( F_i \) takes a value \( f_i \) in \( \mathcal{L} \).

\( F \) is said to be a Markov random field (MRF) on \( S \) with respect to a neighborhood system \( N \) if and only if the two following conditions are satisfied:

\[
P(f) > 0, \quad \forall f \in \mathcal{L}^m \tag{positivity} \]

\[
P(f_i|f_{S-i}) = P(f_i|f_{N_i}), \quad \forall i \in S \tag{Markov property} \]

where \( f = \{f_1, ..., f_m\} \) is a configuration of \( F \) (corresponding to a realization of the field), \( P(f) \) is the joint probability \( P(F_1 = f_1, ..., F_m = f_m) \) of the joint event \( F = f \), and

\[
f_{N_i} = \{f_i', i' \in N_i\} \tag{3} \]

denotes the set of values at the sites neighboring \( i \), i.e. the neighborhood \( N \) centered at position \( i \).

To exploit MRFs characteristics in a practical way we need to refer to the Hammersley-Clifford theorem \[13\] for which the probability distribution of a MRF has the form of a Gibbs distribution, i.e.:

\[
P(f) = Z^{-1} \times e^{-\frac{1}{\beta}U(f)} \tag{4} \]

where \( Z \) is a normalizing constant called the partition function, \( T \) is a constant called the temperature and \( U(f) \) is the energy function. The energy function

\[
U(f) = \sum_{c \in \mathcal{C}} V_c(f) \tag{5} \]

is a sum of cliques potentials \( V_c(f) \) over all possible cliques \( \mathcal{C} \), where a clique \( c \) is defined as a subset of neighboring sites in \( S \). The practical value of the theorem is that it provides a simple way of specifying the joint probability. \( P(f) \) measures the probability of the occurrence of a particular configuration: the more probable configurations are those with lower energies.

We can model geometric attacks with a random field \( F \) defined on the set \( S \) of the image pixels where the value assumed by each random variable represents the displacement associated to a particular pixel. Specifically, for each pixel we have two values for the two directions \( x \) and \( y \), for this reason each variable \( F_i \) is assigned a displacement vector \( f_i = (f_x, f_y) \in \mathcal{L} \times \mathcal{L} \).
As we said, an MRF is uniquely determined once the Gibbs distribution and the neighborhood system are defined. In the proposed approach, for each pixel \((x, y)\) only four neighbors of first order and the corresponding four pair-site cliques. The potential function we used is a bivariate normal distribution expressed by:

\[
V((x,y),(\tilde{x},\tilde{y})) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ -\left[ \frac{(f_x - f_{\tilde{x}})^2}{2\sigma_x^2} + \frac{(f_y - f_{\tilde{y}})^2}{2\sigma_y^2} \right] \right\} \tag{6}
\]

where \(f_x\) and \(f_y\) are the x,y components of the displacement vector \(f(x,y)\) associated to the pixel \((x, y)\), \((\tilde{x}, \tilde{y})\) is a point belonging to the 4-neighborhood of \((x, y)\), \(f_{\tilde{x}}\) and \(f_{\tilde{y}}\) are the x,y components of the displacement vector \(f(\tilde{x}, \tilde{y})\) associated to the pixel \((\tilde{x}, \tilde{y})\) and \(\sigma_x\) and \(\sigma_y\) are the two components of the standard deviation vector \(\sigma\) (these values are controlled by perceptual constraints).

We want to generate a displacement field according to the Gibbs probability distribution defined by equation (4) and the particular potential function expressed in (6).

To do so, the displacement field is initialized by assigning to each pixel \((x, y)\) in the image a displacement vector \(f(x,y)\) generated randomly (and independently on the other pixels) in the interval in \(L \times L\) with \(L = \{f \in \mathbb{Z} : -c \leq f \leq c\}\) (the value of \(c\) is determined by following perceptual considerations). The MF-DA field is then obtained by applying an iterative smoothing algorithm to the randomly generated field. More specifically, the technique we used is the Iterated Conditional Mode (ICM) algorithm detailed in [14]. When the ICM algorithm starts each pixel \((x, y)\) of the displacement field is randomly visited and its displacement vector updated by trying to minimize the potential function (6). After that, each iteration, each pixel is visited and the corresponding displacement updated, a new iteration starts. The algorithm ends when no new modification is introduced for a whole iteration, which is usually the case after 7-8 iterations.

We also considered a multiresolution version of the MF-DA, where the full resolution version of the the displacement field is built by interpolating the displacement field obtained by applying the MF-DA at a resolution level \(L\).

### 3.2 The perceptual analysis

In order to evaluate the potentiality of the MF-DA class of attacks, the perceptual impact of the distortion they generate must be taken into account. From a perceptual point of view, MRF DA’s have a different behavior for different values of \(L\). Specifically we found that, in case of images of size 512 \(\times\) 512, the larger perceptually admissible distortions are obtained by using \(L = 6\) \(\sigma = 1\) \(c = 6\), \(L = 5\) \(\sigma = 3\) \(c = 8\) and \(L = 4\) \(\sigma = 7\) \(c = 18\) (\(\sigma = \sigma_x = \sigma_y\)).

In figures (5) two examples of images distorted with an MF-DA attack applied at different levels of resolution are shown: in the Barbara image the MRF is applied at a lower level of resolution \((L = 6)\), while in the Lena image the distortion is generated at a higher level of resolution \((L = 4)\). In both cases by comparing the original image (on the left) with the attacked one (on the right), we can notice a slightly perceptible distortion, that is however not annoying due to the smoothness constraints of the field (the distortion is not visible if only the attacked image is provided so that the comparison with the original image is not possible).

**Remarks** The research described in this section has led to the following publication.

Fig. 5. Example of two images attacked with the MF-DA model: (a) original image; (b) attacked image with $L = 6$; (c) original image; (d) attacked image with $L = 4$
T2.2 - Validation of theoretical results on synthetic data

Abstract. The validity of the optimum embedding strategy under AWGN attacks described in section T1.3 (subsection 3) of the WP1 report is tested by applying it to synthetic data. The estimates of the false negative error exponent that we have got are in excellent agreement with the theoretical analysis.

1 Validation of the AWGN optimum embedder against synthetic data (year: Y1)

In order to check the validity of the theoretical analysis described in T1.3 of the WP1 report, we performed a set of experiments on synthetic data, computing the empirical values of

\[ -\frac{1}{n} \log(P_{fn}), \]  
(1)

and comparing them with the theoretically obtained expressions for the error exponent of the probability of false negative; be aware that the definition of error exponent of false negative probability \( E_{fn} \) is nothing but (1) when \( n \) goes to infinity, so the expected result is that when \( n \) is increased the empirically obtained results of (1) go to the theoretically obtained \( E_{fn} \).

Given the framework of the previous analysis, both the host and the noise signals used in these experimental results are Gaussian distributed, being the embedding strategy that described in T1.3 of WP1.1. Once the noise is added to the watermarked signal, the detection is performed, so one can estimate the probability of wrongly saying that the attacked signal is not watermarked. In order to the obtained results to be significant, the number of considered \( n \)-length vectors was made equal to

\[ \left\lceil \frac{1000}{e^{-n \cdot E_{fn}}} \right\rceil, \]

which means that the required number of vector observations is increased exponentially with the dimensionality of the problem \( n \). This constitutes a problem for the computation of (1) for high dimensionality vectors, as the number of the considered scalar variables is equal to \( n \left\lceil \frac{1000}{e^{-n \cdot E_{fn}}} \right\rceil \). Therefore, the only choice we have for being able to obtain significant values of (1) for large values of \( n \) is to consider very small values of \( E_{fn} \).

Fig. 1 shows an example of the experiments we carried out, where \( D_e = 2, \sigma_X^2 = 1, \lambda = 0.6 \) and \( \sigma_N^2 \), taking values in \{0.52, 0.53, 0.54, 0.55\}. The similarity among the considered values of \( \sigma_N^2 \) is due to the mentioned attempt of working with small values of \( E_{fn} \) in order to be able to produce meaningful results with a reasonable number of observations. Upon inspection of Fig. 1 the agreement between the theoretically foretold \( E_{fn} \) and its empirical approximation when \( n \) is increased clearly comes out,\(^1\) thus confirming experimentally the validity of the theoretical results.

\(^1\) In all the four cases plotted in Fig. 1 the value of \( n \) was not increased due to the mentioned exponential increase of the number of considered vectors.
Fig. 1. Theoretically obtained error exponent of the probability of false negative, and $-\frac{1}{n} \log(P_{fn})$ as a function of the number of dimensions $n$. $D_x = 2$, $\sigma^2_X = 1$ and $\lambda = 0.6$, and $\sigma^2_X$ equal to 0.52, 0.53, 0.54 and 0.55, respectively (from top to the bottom).
2.3 Validation of theoretical results on real data

Abstract. The actual de-synchronization capabilities of the geometric attacks studied within the project when they are applied to natural images watermarked with some of the most popular watermarking schemes proposed so far are measured.

1 Introduction

In this section of the report, we evaluate the de-synchronization capability of the new classes of attacks (LPCD and MRF) and we compare it with the Stirmark random bending attack. To do so, two very simple watermarking algorithms were implemented and the ability of the various DAs to inhibit watermark detection evaluated. The source image database used for the experiments includes the six standard images: Baboon, Barbara, Boats, Goldhill, Lena and Peppers.

The tested algorithms include:

- Blind additive Spread Spectrum in the frequency domain (BSS-F);
- Blind additive Spread Spectrum in the wavelet domain (BSS-W).

In both the systems the watermark consists of a sequence of \(n_b\) bits \(X = \{x(1), x(2), ..., x(n_b)\}\); each value \(x(i)\) being a random scalar that is either 0 or 1, with equal probability.

In the BSS-F algorithm the watermark is inserted into the middle frequency coefficients of the full frame DCT domain. The DCT of the original image is computed, the frequency coefficients are reordered into a zig-zag scan and the first \(L + M\) coefficients are selected to generate a vector \(W = \{t(1), t(2), ..., t(L), t(L + 1), ..., t(L + M)\}\). Then, in order to obtain a tradeoff between perceptual invisibility and robustness to image processing techniques, the lowest \(L\) coefficients are skipped and the watermark \(X\) is embedded in the last \(M\) coefficients \(T = \{t(L + 1), ..., t(L + M)\}\), to obtain a new vector \(T' = \{t'(L + 1), ..., t'(L + M)\}\) according to the following rule:

\[
T' = T + kPN \text{ if bit} = 0
T' = T - kPN \text{ if bit} = 1
\]  

(1)

where \(k\) is the embedding strength and PN is a uniformly distributed pseudo-random sequence of 1 and –1. Equation (1) refers to the embedding of one bit, the extension to multiple bits consists of applying equation (1) for each bit considering each time a different subset of \(T\) and a different PN sequence.

In watermark detection the DCT is applied to the watermarked (and possibly attacked) image, the DCT coefficients are reordered into a zig-zag scan, and the coefficients from the \((L + M)/h\) to the \((L + M)/h\) are selected to generate a vector \(T^* = \{t(L + 1), ..., t(L + M)\}\). For each bit the correlation coefficient between the corresponding subset of the \(T^*\) vector and a new PN sequence is evaluated and compared to a threshold (equal to 0) to recover the embedded bit.

The correlation coefficient is evaluated in the following way:

\[
r(A, B) = \frac{\sum_{i=1}^{n} (A(i) - \mu(A)) \cdot (B(i) - \mu(B))}{\sqrt{\left(\sum_{i=1}^{n} (A(i) - \mu(A))^2\right) \cdot \left(\sum_{i=1}^{n} (B(i) - \mu(B))^2\right)}}
\]  

(2)
where $A$ and $B$ are two vectors of same size $n$ and $\mu$ is the mean operator, and the decision rule states that:

\[
\begin{align*}
\text{bit} &= 0 \quad \text{if } r > 0 \\
\text{bit} &= 1 \quad \text{if } r < 0
\end{align*}
\]

(3)

In the BSS-W watermarking system the watermark is added to the DWT coefficients of the three largest detail (i.e. LH, HL, HH) subbands of the image. The embedding and decoding functions are implemented in the same way of the previous system but the watermark is inserted in the wavelet coefficients obtained with a one step wavelet decomposition.

2 Results

The six standard images were watermarked with the systems described above with different payloads and then attacked with RBA and the two new classes of attacks. Each image is attacked with a different realization of the field. In Table 1 the values of the parameters used for the experiments are shown. Fig. 1 and 2 show the ability of the RBA and of the two new DAs to inhibit correct decoding. The average of the bit error rate obtained for the six images is plotted versus different values of the payload for both the watermarking systems.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stirmark</td>
<td></td>
</tr>
<tr>
<td>$b$</td>
<td>0</td>
</tr>
<tr>
<td>$d$</td>
<td>0</td>
</tr>
<tr>
<td>$i$</td>
<td>0</td>
</tr>
<tr>
<td>$o$</td>
<td>0</td>
</tr>
<tr>
<td>$R$</td>
<td>0.1</td>
</tr>
<tr>
<td>MF-DA</td>
<td></td>
</tr>
<tr>
<td>$c$</td>
<td></td>
</tr>
<tr>
<td>$d_{im}$</td>
<td></td>
</tr>
<tr>
<td>DCT system</td>
<td></td>
</tr>
<tr>
<td>$k$</td>
<td>5</td>
</tr>
<tr>
<td>$L$</td>
<td>25000</td>
</tr>
<tr>
<td>$M$</td>
<td>16000</td>
</tr>
<tr>
<td>DWT system</td>
<td></td>
</tr>
<tr>
<td>$k$</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1. Value of the parameters used for the experiments

![Bit error rate vs Payload (DCT domain)](image)

Fig. 1. De-synchronization capabilities of the various DAs against the DCT domain system
Fig. 2. De-synchronization capabilities of the various DAs against the DWT domain system

For both the systems, the RBA attack is not able to prevent a correct watermark decoding, in fact the RBA plot is not visible in the figures because the bit error rate is always equal to zero. A more powerful class of DAs is the LPCD DAs that in both the systems gives a bit error rate much higher than the RBA attack. The MF-DA always results in a very high bit error rate also applying the attack to a lower level of resolution.

Remarks  The research described in this section has been partly published in the following paper

### Abstract
As a first step towards the development of a new watermarking system that is robust against watermark desynchronization, an image quality metric has been developed to assist the registration process between the watermark and the to-be-inspected image. An innovative quality metric has been developed to this aim, and tested by means of extensive psychovisual experiments. The validity of the proposed metric is confirmed by the experiments.

### 1 Introduction
As described in the project workplan the last phase of the project is devoted to the development of an innovative watermarking scheme that permits to cope with general DAs, with particular reference to local DAs such as LPCD and MRF.

The strategy that is going to be used works according to following paradigm.

An optimum watermarking scheme is selected among those developed in WP1. Regardless of the choice, the selected detection algorithm will define a sufficient statistic detection is based on. For instance, the optimum detector for discrete signals described in section 3 of the WP1.2 report is based on the mutual information between the searched watermark and the image under inspection. Alternatively, in the continuous universal case, the correlation coefficient between the watermark and the image may be considered.

In the presence of DAs, the watermark is registered onto the image trying to maximize the detector answer (e.g. the mutual information or the correlation coefficient). In such a phase, a constraint must be imposed on the smoothness of the displacement field used for the registration, that ensures that the geometric distortion used during the registration is perceptually admissible. This can be done either by constraining the maximization over the set of distortions for which the perceptual degradation is below a certain limit, or by adding a smoothness term to the objective function. In both cases, it is necessary that a metric is available to measure the perceptual degradation introduced when a given displacement field is applied to the image at hand. It is the scope of the following section to describe the work that has been carried out in this sense during the second year of the project.

### 2 A quality metric for geometrically distorted images (year: Y2)

The main function of the HVS when looking at an image is to extract structural information from the viewing field, therefore a measurement of structural distortions should be a good approximation of perceived image distortion.

Psychophysical studies show that human vision is sensitive to edges and bars in images, and structures of objects in images are typically outlined by edges and bars[15]. Hence, we expect that a measure that links the geometric distortions to the presence of edges and bars in the image is likely to provide an adequate measure. We decided to use Gabor filters to extract bar and edges information from the images and to use these features to evaluate the perceptibility of the distortions.

The idea underlying the quality metric we developed, is that a geometric distortion causes a degradation of the structure of the objects in the visual scene when the displacement field describing the distortion is orthogonal to the direction of the image bars and edges and is not homogeneous along this direction.
2.1 Overview of the metric

A 2D Gabor kernel for the edge detection in images, also called antisymmetric Gabor function, can be mathematically defined as:

\[ GaborE_{\lambda, \theta, \sigma, \gamma}(x, y) = e^{-\frac{x'^2 + y'^2}{2\sigma^2}} \sin \left( \frac{2\pi x'}{\lambda} \right) \] (1)

The corresponding equation for the bar detection, or symmetric Gabor function, is:

\[ GaborB_{\lambda, \theta, \sigma, \gamma}(x, y) = Bar_{\lambda, \theta, \sigma, \gamma}(x, y) - \overline{Bar}_{\lambda, \theta, \sigma, \gamma} \] (2)

where

\[ Bar_{\lambda, \theta, \sigma, \gamma}(x, y) = e^{-\frac{x'^2 + y'^2}{2\sigma^2}} \cos \left( \frac{2\pi x'}{\lambda} \right) \] (3)

and \( \overline{Bar}_{\lambda, \theta, \sigma, \gamma} \) is the mean value of the function defined in equation (3).

The parameters we used for the design of the filters are the following: \( \gamma = 0.5 \) and \( b = 1 \). Using these values the resulting filters for the edge and bar extraction, respectively, are the ones shown in Fig.1 (with \( \theta = 0 \) degrees and \( \lambda = 10 \) pixels).

![Fig. 1. Filters for the edges and bars detection with \( \gamma = 0.5, b = 1, \theta = 0, \lambda = 10 \): (a) GaborE_{\lambda, \theta, \sigma, \gamma}, (b) GaborB_{\lambda, \theta, \sigma, \gamma}.](image)

Once defined the parameters of the filters, and fixed a particular \( \theta \), we use the functions described in equations (1) and (2) to filter the original image and to find edges and bars in the direction orthogonal to \( \theta \). The filtering function we use is described by the following equation:

\[ I_f = \sqrt{I_{f,\text{bar}}^2(x, y) + I_{f,\text{edge}}^2(x, y)} \] (4)

where \( I_f \) is the filtered image and \( I_{f,\text{bar}} \) and \( I_{f,\text{edge}} \) in equation (4) are the original image convolved with the Gabor filters described respectively by equations (1) and (2).

To find the score associated to each pixel quantifying the perceivable degradation of the image at that pixel we need to link edges and bars of the image with the displacement field in the corresponding location. Specifically, the score associated to each pixel is defined by the following equation:

\[ Obj(x, y) = \sum_{\theta} (I_{f,\theta}(x, y))^\alpha \left( \frac{\partial D}{\partial d_\perp \theta}(x, y) \right)^\beta \] (5)
where $I_{f,\theta}$ is the filtered image described by equation (4) in the $\theta$ direction, $\alpha$ and $\beta$ are two constants whose values have been fixed with the experimental results, and the notation $\frac{\partial D_\theta}{\partial d_{\perp \theta}}$ indicates the gradient of the displacement field in the $\theta$ direction with respect to the direction orthogonal to $\theta$.

The overall score, quantifying the perceived distortion, is computed by using the Minkowsky relation, as follows:

$$\text{Score} = \left( \sum_{x,y} |\text{Obj}(x,y)|^p \right)^{\frac{1}{p}}$$  \hspace{1cm} (6)

where $p$ is typically a constant between 1 and 4 whose value has been set experimentally.

### 2.2 Metric tuning

Two sets of subjective experiments were carried out with different purposes. The first set of experiments, the training test, was performed to tune the proposed objective metric in the previous section with psychovisual data in order to transform them into a perceptual metric: the goal was to find a function that, given the score value given by equation (6), returns a numerical value that quantifies the image quality. This first set of experiments was also used to set the parameters of the model $\theta, \lambda, \alpha, \beta, p$.

The image database used for the test included twelve gray scale high quality images, 512 x 512 pixel in size, and was derived from a set of source images that reflects adequate diversity in image contents. The images of the database include pictures of faces, houses and natural scenes. Some images have high activity, while some do not have much structures and are mostly smooth.

To automatically generate the local geometric distortions to be applied to the images and to have a broad range of image impairments we used the Constrained LPCD model and the Markov Random Field model with different parameters in order to obtain different kinds of distortions going from invisible distortions to very annoying distortions, for a total of ten different distortions for each image and a total of 120 different images to be evaluated.

The subjective scaling method we used to measure the perceived quality of geometrically distorted images is the Absolute Category Rating (ACR) method, a category judgement where the test images are presented one at a time and are rated independently on a category scale.

Procedures for ACR experiment have been designed by following the ITU-T Recommendation P910,13 which suggests standard viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods.

The experiments were conducted in a dark room by using the VP800 video card of the Cambridge Research Systems together with a high resolution 21-inch digital monitor Mitsubishi DiamondPro 2070 with the external adaptor ViSaGe 71.02.00D2.14[11]. To have a correct color representation a luminosity calibration was previously carried out through a videocamera ColorCAL.14. Subjects viewed the monitor from an approximated viewing distance of 2 screen heights.

The tests involved a panel of fifteen subjects with a good vision, all naives with respect to image quality assessment methods and image impairments. Subjects were shown images in a random order, the randomization was different for each subject.

The subjective scores have to be analyzed with statistical techniques used in standard methods to yield results which summarize the performance of the metric. The averaged score values (MOS), that is the arithmetic mean of all the individual scores, are considered as the amount of distortions that anyone can perceive on a particular image. The MOS values obtained following the above approach were used to derive a perceptual metric by fitting them with a psychometric curve. Through psychometric mapping, a match between the human perception of geometric artifacts and the values provided by the objective metric defined by equation (6) is established. In particular, we used the Weibull function defined as follow:

$$y = -4 \left(1 - e^{-\left(\frac{z}{k}\right)^k}\right) + 5$$  \hspace{1cm} (7)
where \( c \) and \( k \) are the parameters to be estimated by fitting the objective metric values to the subjective data, and \( x \) is the objective metric used to measure the visual distortions, that is the score value described by equation (6). We opted for this psychometric curve since it provided the best fit for our data among the commonly used curves, i.e., Gaussian, logistic and Weibull curves.

The results of the subjective test are described in figure 2 which shows the scatter plot of the Mean Opinion Score versus the objective metric evaluated by using equation (6). Specifically, Fig.2.(a) shows the scatter plot for all the 120 images while the following three graphs present the results obtained for each class of images.

![Scatter plot of the MOS vs Objective Metric](image)

**Fig. 2.** Scatter plot of the Mean Opinion Score versus Objective Metric: (a) all images; (b) houses images; (c) natural images; (d) faces images.

The equation of the overall score associated to each distortion, using the values obtained through the first set of experiments, is given by the following formula:

\[
\text{Score} = \sum_{x,y} \sum_{\theta=0}^{\pi} I_{\epsilon,\theta}(x,y) \left( \frac{\partial D_{\theta}}{\partial d_{\theta}}(x,y) \right)^3
\]

and the final perceptual metric is described by the following equation:

\[
\text{GaborMetric} = -4 \left( 1 - e^{-\left( \frac{\text{Score}}{c} \right)^k} \right) + 5
\]
where \( c \) and \( k \) are parameters whose value is reported in table 1. The Weibull function in equation (9) describes our metric: for each distortion it returns a numerical score, going from 1 to 5, quantifying the dissatisfaction of the viewer observing the distorted image (with 1 corresponding to a bad image quality and 5 to an excellent image quality).

<table>
<thead>
<tr>
<th>All images</th>
<th>House images</th>
<th>Natural images</th>
<th>Face images</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c )</td>
<td>0.00229</td>
<td>0.00212</td>
<td>0.001152</td>
</tr>
<tr>
<td>( k )</td>
<td>0.3529</td>
<td>0.4061</td>
<td>0.2927</td>
</tr>
</tbody>
</table>

Table 1. Value of the parameters used for the Weibull fitting function

2.3 Metric validation

The second set of experiments, the validation test, was conducted to validate the proposed metric and to compare it with the state of the art quality metrics for geometric distortions.

For the validation step, once again we used the ACR test following the procedures previously explained. A new dataset [8] of twelve images was built according to the class of images explained above and new ten distortions for each image were generated by using the same models C-LPCD and MRF. The tests involved a panel of others fifteen subjects, all naives with respect to image quality assessment methods and image impairments.

The results of the test are shown in Fig.3 that describes the scatter plot of the MOS versus the perceptual metric described for all the images and for classes of images.

In order to provide quantitative measures on the performance of the proposed model, we followed the performance evaluation procedures employed in the video quality experts group (VQEG) Phase I FR-TV test.

To remove any nonlinearities due to the subjective rating process and to facilitate comparison of the models in a common analysis space, the relationship between objective data and the subjective ratings was estimated by using a nonlinear regression, the red plot in Fig.3.

Once the nonlinear transformation was applied, the objective metric performance was evaluated through three attributes, applied on the fitted values, as specified in the report of the VQEG group [16]. The first metric is the Pearson linear correlation coefficient between the objective/subjective scores that is a measure of prediction accuracy of a model. The second metric is the Spearman rank-order correlation coefficient between the objective/subjective scores that is considered as an evaluation of the prediction monotonicity. Finally, the third metric is the outlier ratio (percentage of the number of predictions outside the range of 2 times of the standard deviations) of the predictions after the nonlinear mapping, which is a measure of prediction consistency. The result of the performance evaluation of the proposed algorithm through these three metrics is shown in table 2.

<table>
<thead>
<tr>
<th>All images</th>
<th>House images</th>
<th>Natural images</th>
<th>Face images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson correlation coefficient</td>
<td>0.9109</td>
<td>0.9457</td>
<td>0.9028</td>
</tr>
<tr>
<td>Spearman correlation coefficient</td>
<td>0.9039</td>
<td>0.9347</td>
<td>0.9188</td>
</tr>
<tr>
<td>Outlier ratio</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Performance of the proposed perceptual metric

The experimental results show good performances of the metric, either applying the scheme to a generic image or applying the algorithm to a specific class of images. Specifically, by referring to table 2 and by looking at the scatter plot in Fig.3, the following considerations are in order:
Fig. 3. Scatter plot of the Mean Opinion Score versus Gabor metric: (a) all images; (b) house images; (c) natural images; (d) face images.

- The outlier ratio is always equal to zero, meaning that the metric maintains prediction accuracy over the range of image sequences, and both the Pearson and the Spearman coefficient are quite high for all the classes of images revealing a good prediction accuracy and monotonicity of the model.

- Applying the model per class of images instead to all the images together allows us to obtain a little improvement of the objective metric for the class of houses images but it is not so relevant considering the disadvantage of having different objective metrics for different classes of images.

- The performance of the proposed metric slightly decreases for the class of face images, this can be explained by the absence of a multi-scale filtering or because face images evoke some higher level visual processing in human vision not captured by the low level processing for edges and bars, on which our metric is based. In fact HVS consists of two major components: an early vision model and a visual attention model which indicates regions of interest of a scene through the use of Importance Maps. The attention model takes into account several factors which are known to influence visual attention and eye movements and these factors must be considered to obtain a more accurate indication of picture quality.

- After the nonlinear regression the relationship between the objective and subjective data is almost linear thanks to the fitting procedure that we used for the metric design and described in the previous section.
We also performed a comparison of the proposed technique with other metrics proposed in the literature. Specifically we considered the PSNR measurement and the SSIM-index[6], that are two widely used full-reference quality metric thanks to their simple formulation and computation, and the Gibbs metric and the metric based on the variance of the jitter noise[17], that are specifically designed to deal with geometric distortions.

From these comparisons we can derive that the proposed measure outperforms the quality metrics proposed so far.

2.4 Remarks

The research described in this section has been published in the following papers.

References