**Programming Research Group**

# AN IDEA FOR A BLIND WATERMARKING SCHEME RESISTANT TO *STIRMARK*

Andrew D. Ker

PRG-RR-01-14

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford OX1 3QD

**Abstract**

We present an idea for digital watermarking of still colour images. The scheme described is *blind*, so that the watermark detector does not require access to the unmarked image, or an undistorted copy of the watermarked image. We make use of the correlation of signals in the different colour components of a colour image, even after special desynchronisation attacks which usually defeat blind watermarks, using some components of the image to synchronise the others. We term this the dual channel approach to watermarking of colour images. A mathematical problem describing the difficulties with this approach is formulated, and a solution so simple as to be almost trivial is exhibited. We show how the solution can be used to motivate a practical watermarking scheme. An extremely crude implementation of this scheme is made. Despite its basic nature, this scheme performs well under some preliminary testing, exhibiting robustness to filtering attacks, JPEG compression, small amounts of rotation, scaling, and other linear transformations, and the *StirMark* tool even with greater than default parameters.

# 1 Introduction

The rise of computing technology and communications infrastructure has brought challenges along with the many manifest benefits. Whilst digitization of multimedia allows authors heretofore unparalleled ease of manipulation of their works, and publishers can take advantage of negligible costs of distribution compared to those for physical media, both authors and publishers are concerned that their copyright can be infringed more easily, or their works tampered with and passed off as genuine.

The field of digital watermarking aims to provide a tool to mark ownership of digital media. In this paper we focus on the watermarking of still colour images (other commonly watermarked media include moving video and audio, and much of the image watermarking literature concentrates on grayscale images), and the particular type of watermark, known as *blind*, which can be detected without access to either the unmarked image or an undistorted version of the marked image. We begin by describing some background to digital watermarking, and some terminology associated with it, in Section 2. The motivation for this work is given in Section 3, where we examine the well-known *StirMark* attack, which distorts a watermarked image only imperceptibly but renders the watermark inserted by most watermarking schemes undetectable. In Section 4 we describe a mathematical problem which, if a solution could be found, may provide a watermarking method resistant to this particular attack, at least for colour images. We note one simple, almost trivial, solution to the mathematical problem, which at first sight appears to have no relevance to watermarking, but show that in fact it can lead to a viable watermarking scheme.

In Section 5 we describe a very rough-and-ready implementation of this idea, and in Section 6 we show that despite its crudity the implementation functions creditably. In Section 7 we note briefly some related research by other authors, some of which uses quite similar ideas. Finally, in Section 8, we discuss how some of the shortcomings of this implementation may be solved, suggesting some of the many directions for further research prompted by the ideas introduced in this paper.

We stress that the details of the watermarking algorithm presented here is only of a prototypical nature — this makes its efficacy somewhat surprising — and that further work should produce a much more powerful watermarking method based on these ideas.

## 2 Watermarking Background

The aim of a watermarking scheme is to embed information into an item of media. It should provide an algorithm to translate any input media into an output which is "marked" in such a way that another algorithm can recognise and decode the mark. The efficacy of the watermarking scheme is measured by the *imperceptibility* of the inserted mark (to a human observer) and the *robustness* of the mark to other imperceptible manipulations of the marked data. A watermarking scheme which is not imperceptible will not be suitable for high fidelity applications. A watermarking scheme which is not robust is of very little use at all (excepting "fragile" watermarks designed to detect tampering); if an enemy can apply an imperceptible operation to the media and render the mark undetectable then all security endowed by the mark is lost. Imperceptibility and robustness are competing goals, because increasing robustness must mean more alteration to the original media, distortion which at some level will become perceptible.

The insertion and detection processes will be parameterised by a secret key, analogous to an encryption key in cryptography. Even with knowledge of the algorithm, it should not be possible to detect or remove the watermark without the correct key. The amount of information embedded is typically a few dozen bits or just a single bit (indicating only the presence of the mark), depending on the intended application.

In this paper we consider watermarking for colour images. The motivating application we have in mind is copy control: this involves placing a watermark detector in some (ideally every) piece of equipment used to for copying data of the appropriate type. The equipment would refuse to copy marked material[1]. Because a copy control device needs only to know if the mark is present, a single bit watermark suffices. In fact the copy control watermarking schemes currently in use do embed more than a single bit of information, but the idea explored in this paper could also be applied to higher capacity marks if necessary.

There are various types of watermarking scheme, which are usually distinguished by the amount of information available to the detector. A watermarking system for copy control would need to be *blind*[2] — the detector does not have access to any information about the image being tested other than the image itself and the secret key, in particular not

[1]Ideally the implementation of the copy control mechanism should include ways for those in possession of the image to exercise their fair use rights, although many such devices presently in use are deficient in this regard. Copy control may also work in conjunction with digital licensing schemes to permit copying in certain circumstances.

[2]Other authors use the term *oblivious*.

the original unmarked image — because at the time of deployment not every image to be filtered against would be known. The alternative types of watermarking system occur when the detector also has access to an undistorted copy of the watermarked image but not the original unmarked image (this sort of scheme might be called *semi-blind*), or when the detector has access to the original unmarked image (usually called *private watermarking* or *cover image escrow*).

Clearly the amount of information available to the detector is greatest for a cover image escrow detector and least for a fully blind detector; fully blind watermarking is the most challenging problem. Fully blind watermark schemes are also especially vulnerable to the attack described in the following section.

### Other Watermarking Schemes

We expect that the reader will be familiar with the standard watermarking literature, so do not give a comprehensive survey. Instead we highlight relevant features of the standard works. In Section 7 we discuss some recent works based on ideas similar to those presented here.

The vast proportion of watermarking schemes for still images take the form of insertion, via spread spectrum techniques, of additive or multiplicative noise into either the image direct (the spatial domain) or in some transform space of the image signal (often the frequency domain or related concepts such as the DCT or DWT domain). The location (in space, frequency, etc.) of the noise is determined by the secret key, and the content of the noise either by the watermark to be inserted or (in the case of single bit watermarks) pseudorandom noise generated from the secret key.

When the original image is available for detection, the detector can subtract this from the image under scrutiny (perhaps after applying corrective measures to remove distortion in the tested image), use the secret key to locate the correct places for the added noise, and either output the result as the embedded information or (in the case of a single bit watermark) correlate with the expected noise to produce a confidence level of the presence of a mark. When the original image is not available the additive noise will have been inserted in such a way as to alter some statistics of the image, exactly which statistic depending on the secret key. The detector will compute the relevant statistic and use it to output either the message of the watermark or a confidence level for the presence of a mark.

The very early watermarking schemes tended to rely on the least significant bits of pixel values in an image, or in quantization noise. Although the watermarking process is usually imperceptible the marks are fairly easily destroyed by filtering or requantization. The classic robust watermarking scheme is that of Cox *et al.* [CKLS96], which spreads a gaussian noise watermark into the low frequency components of the DCT transform of a grayscale image. This system requires the original unmarked image for detection. A generalisation of this system is presented in [FBS98], where a key-dependent transform space is used to enhance the security of the watermark.

In [PZ98] the authors introduce both a block-based DCT approach and one using a multiresolution wavelet transform. They also use a perceptual mask, which modulates the amplitude of the inserted watermark to place the most energy in precisely those parts of the host image which can take it without perceptible distortion. These systems are also private watermarking schemes needing the original image for detection.

The *Patchwork* algorithm, described in [BGM95] and explored in [GB98], is of interest here because we make use of a similar technique the watermarking process described in this paper. *Patchwork* uses as secret key two disjoint sets, of equal size, of pixels in an image; the brightness value of each pixel in the first set is increased a little, and decreased a little for each pixel in the second set. The individual alterations in brightness are not visible, but computing the sum of the brightness values of the first set of pixels and subtracting those for the second set gives a statistic which should be close to zero in an unwatermarked image, but significantly above zero in marked images. This is a blind watermarking system, but it suffers from poor robustness.

More recent work includes [LWB$^+$00], which uses a novel transform space to achieve robustness to rotation and scaling, and [LLHS99] which inserts two complementary watermarks in the DWT.

## 3 Attacks on Image Watermarking Systems

As one would expect, the progress made in understanding digital watermarking has included improvements in our understanding of effective attacks on watermarks. Most attacks aim to render the watermark undetectable (there are others, so-called "protocol attacks" such as [CMYY98], and watermark estimation attacks including [LvD98], which are outside the scope of this paper). The power of a watermark attack can be measured in a complementary way to that of a watermarking algorithm: the attack should introduce only imperceptible distortions to the image, but should damage as much of the information conveyed by the watermark as possible.

We note that there is an inherent asymmetry between watermarking schemes and attacks on them: for a watermarking scheme to be useful in practice it must resist *every* known attack, whereas an enemy would be able to pick his attack to suit the watermarking scheme in use. Only one successful attack will produce an unmarked version of the original media, which could in the worst case be expected to propagate everywhere via the internet. In the future we would expect a proliferation of attacks targeted at possible weaknesses of individual schemes. For the moment, robustness against general attacks is challenge enough.

Common Attacks

Early literature, including [CKLS96], [FBS98], [PZ98], and [GB98], identified a number of distortions which a watermark should survive. Typically they included: common image processing operations such as blurring or sharpening, lossy compression, addition of noise, scaling (with or without aspect ratio change), rotation or reflection, shear, cropping down of the image, conversion to analogue and then rescanning into digital form, conversion of the image depth (usually conversion of a grayscale image to black and white by dithering or halftoning), collusion attacks (a number of examples of the same image with different watermarks are averaged, or multiple watermarks are inserted into a single image), and special desynchronisation attacks (which are discussed later in this section).

Many of these attacks fall into the same broad category, whereby the image quality is degraded but its synchronisation — the location of the watermark within the tested image — unaffected. This applies to blurring, sharpening, noise addition, conversion to black and white and back (the halftoned or dithered image would be converted back to grayscale and low-pass filtered). It also applies to affine geometric transformations such as scaling, rotation, and shear, but only if the transformation parameters are known. In this case the transformation can be inverted, leaving only some resampling noise. In much of the early literature it was assumed that they *would* be known and inverted, usually because a "registration process" would be used before the image is tested.

Given a registration process, the schemes presented in [CKLS96], [FBS98], and [PZ98] claimed acceptable robustness against many of the above attacks. More modern schemes, including [WSK98], [LLHS99], and [XBA98], appear to have better performance; they are robust against larger classes of attacks or larger amounts of the same attacks. Note that it may depend on the application as to how much distortion we expect a watermarking scheme to be robust to. It is not clear whether, for example, robustness to JPEG compression with a "quality factor" of, say, 70%[3] is sufficient, or whether we require the mark to be still detectable at a quality factor of 5% (which usually results in a heavily distorted image). Similarly, the method proposed in [LLHS99] claims robustness even after a median filter with width almost 10% of the image size. It may be a bonus to have such robustness, but its worth is perhaps questionable; at that level of filtering the image is almost unrecognizable.

The image registration process is all very well for private or semi-blind watermarking schemes when there is a watermarked or original image to register with, but it cannot be used in blind watermarking. In the absence of a registration step even simple rotation and scaling attacks can be effective. Most of the early watermarking systems would fail even against rotation by $1^\circ$, or flipping, of the marked image, if registration was not implemented.

A number of partial solutions to the problems of geometric transformations have been proposed. A common approach to scaling is for the detector to estimate the scaling factor simply using the size of the tested image (and many watermarking schemes claim "robustness to scaling" based on methods amounting to no more than this). As noted by Lin *et al.* in [LWB+00], this does not present a satisfactory solution because cropping or padding of the watermarked image will mislead the detector as to the scaling applied. Some authors (including [HSG99] and [KJB97]) advocate using a brute force search by testing for a watermark at a large number scale factors or rotation angles. This will be computationally extremely expensive, and to test for all possible linear transformations is surely infeasible. Also one must apply special care when computing the probability of false positive detection results.

[3] The quality factor of JPEG compression controls the level of quantization in the DCT domain.

At least three techniques have been proposed to deal properly with rotation, scale and (to some extent) cropping of images without registration. Kutter [Kut98], amongst others, has proposed the use of a regular pattern inserted along with the watermark. Detection of the pattern allows for estimate of the geometric transformation it has undergone. This "registration pattern" must itself be robust enough to survive attacks, and may be especially vulnerable to targeted attacks. Alternatively, as in [LWB+00] or [OP97], a method such as the Fourier-Mellin transform or related space, could be used to embed the watermark. Such transforms are either invariant under — or vary in a simple way with — rotation, scaling and (to some extent) translation. However, these watermarking systems are not robust to shear or change of aspect ratio. Finally, in [AT00] a statistic of the image (based on the alignment and size of edge features) is proposed to allow for blind estimation of rotation and scaling factors. It is not clear how robust these statistics are, and particular vulnerability to cropping is expected.

Again, it may depend on the application as to whether not being robust to large geometric transformations renders the system useless. Arguably, since rotation by a large amount such as $90^\circ$ is not an imperceptible operation, we need not expect the watermark to survive it[4]. It may be acceptable for the watermark to survive only what we could rather loosely term "imperceptible geometric transformations" such as, say, rotation of less than $3^\circ$, and scaling, shearing, or cropping of up to 5%.

### The StirMark Attack

Unfortunately, the exists an imperceptible geometric transformation attack which defeats the vast proportion of watermarking schemes.

Motivated by the known vulnerability of spread-spectrum signals to jitter attacks, Petitcolas *et al.* introduced the attack, designed to interfere with the synchronisation or location of watermarks in images. *StirMark* was first described in [PAK98], with revisions being made and published on the website [SM]. Initially it simulated the distortion likely to be introduced by printing and rescanning an image: small amounts of rotation and shearing along with some smooth noise. More recent versions introduce further spatial displacement, shifting pixels by distances controlled by both low- and high-frequency sine waves, and to achieve a smooth displacement of pixels without too much blurring a non-linear resampling scheme is employed. The parameters of the distortions are randomised; the result is a small random "stirring" effect in the image. Finally, the whole image undergoes mild contrast adjustment and JPEG compression.

Unlike large-scale rotations or flipping, the geometrical distortions introduced by *StirMark* are more or less imperceptible. This exploits the human eye's poor spatial location facility in the absence of reference patterns[5]. An example of the effects of *StirMark* on an image

[4] In the extreme case, note that we cannot hope to have a watermarking scheme robust against every invertible operation. We would need one to completely stop potential pirates working around copy control procedures, but consider the situation whereby a pirate encrypts an image with a secret key and sends it to another; the recipient can decode the image but a watermark detector placed in their communication path will not have any hope of stopping them.

[5] The same characteristic of the human visual system enables the construction of "optical illusions" whereby the same line can appear bent or straight, or of different lengths, depending on its context.

are shown in Figure 1. Because of the relative simplicity of the attack the authors in [PAK98] summarise:

> We suggest that image watermarking tools which do not survive *StirMark* — with default parameters — should be considered unacceptably easy to break.

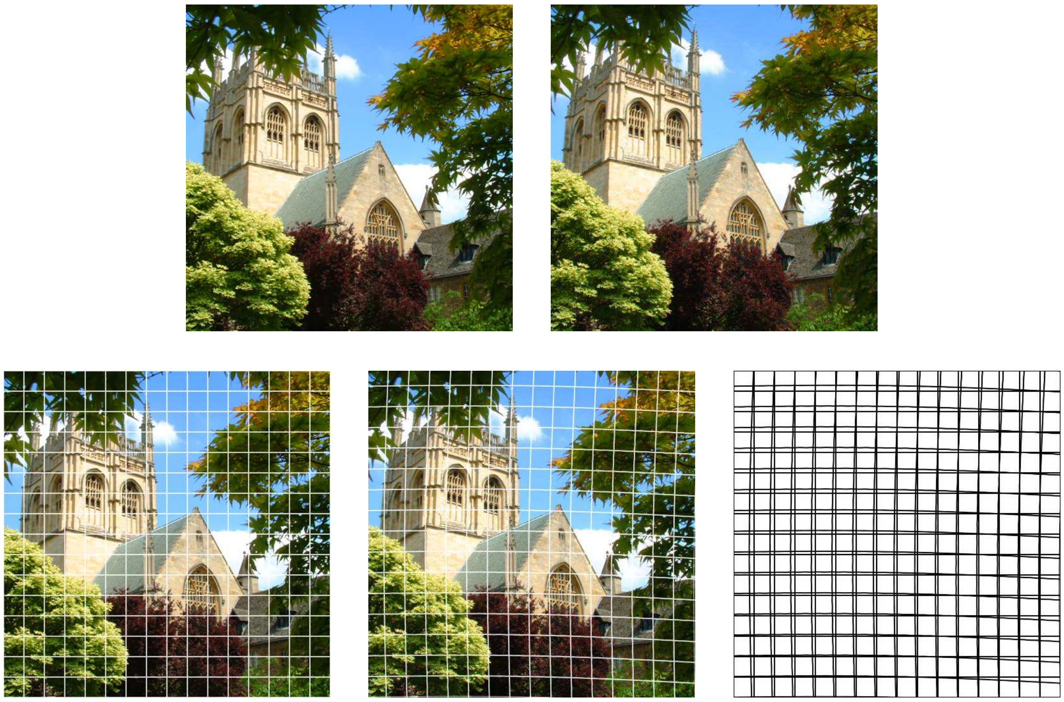
In practice, most if not all watermarking schemes in use at the time *StirMark* was described are defeated by it. The same applies to many more recent proposals. The attack is so devastating because all watermarking schemes must find some way to spread the energy of the watermark throughout the image. In the spatial domain *StirMark* simply moves pixels about by an amount very difficult to determine, so that finding the correct locations for a spatially spread watermark becomes impossible after the attack. Frequency domain watermarks are also destroyed, because the stirring effects smear energy from one frequency into those around it. The movement of pixels or frequency is not large, but attempts to make a watermark robust to *StirMark* by spreading coarsely enough (either spatially or frequentially), so that even with the disturbances most of the watermark information can be found, fails. The spreading has to be so coarse that not many locations exist to spread into (spatially, for example, it seems that one cannot do much better than dividing the image into 64 blocks), and the watermark ends up visible and easy to remove. One the other hand, note that locally *StirMark* has little effect. That is, if one can locate where a block of say 16 by 16 pixels has been moved to, the block itself will have undergone very little distortion (some small shear in practice).

The methods proposed to provide robustness to global rotation and scaling (such as [LWB+00] and [Kut98]) do not help against the *StirMark* attack. The use of templates to recover from global affine transformations is limited by the precision of the template: typically 9 template points are embedded, nowhere near enough to recover from a *StirMark* attack which has potentially almost unlimited degrees of freedom as mentioned in [DP00]. A rotation- or scale-invariant transform will not be *StirMark*-invariant (although for colour images we propose to produce such a transform in the next section).

For non-blind watermarking schemes there are successful countermeasures to *StirMark*. A fairly large number of papers have been produced with roughly the same idea, extending the registration process to correct for the warping introduced by *StirMark* (which is, note, invertible). The schemes are usually based on techniques for motion compensation in video images, and include [DBHC99] and [LK01]. In [JDJ99] it is noted that a few feature points of the image suffice for successful registration, reducing the storage data requirements for a detector using registration, but not enabling truly blind detection.

Once such a powerful attack against image watermarks has been identified, and given that all images can nowadays be converted into digital form — with the security issues this raises — with the utmost of ease, there appear two courses open to copyright holders. At their behest, governments could decree that the possession, distribution, or discussion of *StirMark*, and other related "circumvention devices", is illegal. As long as this proves sufficient discouragement to potential copyright violators, or if law enforcement is made sufficiently powerful to keep the program, or any description of its function, unavailable

Figure 1: The *StirMark* attack. Above, an image and a *StirMark*ed version (right). Below, as above but with a regular grid superimposed on the original, and the superimposition of the undistorted and distorted grid illustrating the level of pixel displacement.



in any medium including the internet, and as long as security researchers do not mind the suppression of discussion of attacks on watermarking technology, security may be achieved. However it appears, to this author at least, perverse to attempt to turn a weak security product into a strong one by legislation. Instead one should encourage research into a watermarking scheme which may resist attacks such as *StirMark*.

## 4 The Dual Channel Approach

Digital colour images are made up of pixels taken from a three dimensional colour space. This space is usually described by orthogonal red, green, and blue components, although for image processing it is common to work in one of a number of transformations of this space (either better to match the characteristics of the human eye, or for compression purposes).

The idea behind this paper is that *StirMark* must apply the same or very similar distortions to the red, green, and blue components of a colour image (or to the components of whatever transform space we choose to work in). If it did not, fairly horrible colour fringing would be visible; see Figure 2 for the results of independent *StirMark* distortions to

Figure 2: The effects of *StirMark* applied independently to red, green, and blue components of a colour image

the three channels. We aim to exploit this relationship, by using one or two of the colour components as synchronisation for the others. This is the "dual channel" approach: the colour image is split into two channels, one for synchronisation and one for the insertion of a watermark.

It would be simple if we could choose the content of the synchronisation channel. We would choose a regular pattern, and then match the corresponding channel of the tested image against it. We could use an unwarping algorithm such as that described in [LK01] to determine the warping of the synchronisation channel and, knowing that the watermarking channel has undergone the same transformation, perform the same inversion. But of course we cannot choose the content of either channel, we have to work with what we are given.

How, then, can we gain anything from the dual channel approach? To explore the problem, we reduce it to a more mathematical setting. One simplification is to restrict attention to one dimensional signals, hoping to generalise a solution for this case to a solution for two dimensional images. This is a reasonable step to take, because even in one dimension the problem of synchronisation of arbitrarily jittered signals presents difficulties; reducing to one dimension has not trivialised the problem.

We also consider continuous rather than discrete signals. This is motivated by the observation that warping really takes place in a continuous domain, with resampling required to approximate the same effects for a discrete signal. Hence a signal, for these purposes, is a function on $\mathbb{R}$. For now, we do not specify the type of signals under consideration (the codomain of the function).

The most general "warpings" — desynchronisation attacks in the style of *StirMark* — would be the maps $w : \mathbb{R} \to \mathbb{R}$ which are continuous, strictly increasing, and onto. This guarantees invertibility, so that the warping only displaces the signals and does not destroy information. The warped version of the signal $f$ would be $w\, ;\, f$, where ';' represents functional composition.

The mathematical representation of the practical watermarking problem is to find a transform $\mathcal{T}$ which takes two signals and produces a transform — which need not necessarily be a signal of the same type as those being transformed, any function is allowable — subject to some conditions. If we write $\mathcal{T}[f, g]$ for the transform of the signals $f$ and $g$, we require:

**Invertibility:** Given a transform $h$ and a signal $f$ it is possible to construct a signal $g$ such that $\mathcal{T}[f, g] = h$.

**Invariance Under Warping:** For all warping functions $w$ (or as many as possible)

$$\mathcal{T}[w\,;f, w\,;g] = \mathcal{T}[f, g].$$

**Robustness:** Informally, we need that closely related signals should give closely related transforms. One way to formalise this would be to require

$$f_n \to f \text{ and } g_n \to g \text{ implies } \mathcal{T}[f_n, g_n] \to \mathcal{T}[f, g]$$

where convergence of signals and transforms is pointwise.

Note that the transform need not be symmetrical in its treatment of $f$ and $g$; in particular there is no requirement for invertibility to hold for the first argument.

Three solutions are apparent:

a) We divide the signal into short time-frames, and within each frame use a Fourier transform to determine the peak frequency of the signal $f$. Then compute a sum-of-cosines approximation to $g$ over the frame in terms of multiples of this base frequency; the (approximate) representation of $g$ over each time frame is the result of the transform. Warping of the function $f$ will shift the peak frequency in each frame by the same amount as the local rate of warping, which will cancel out much of the effect on $g$.

   It is not clear that a transform of this sort would be robust, as determination of peak frequency may be very sensitive to noise. Further, its invariance under distortion is not exact (unless the warping applied is piecewise linear over each time-frame). Finally, it is not clear how this can be generalised well into two dimensions. We discarded it as a practical watermarking method.

b) We identify "feature points" of the signal $f$, perhaps local maxima of a smoothed version of it. We then determine a smooth warping which, applied to $f$, leaves those feature points regularly spaced. Finally, we apply that same warping to $g$ and output the warped signal as the transform.

   This solution is clearly invertible and robust, so long as the feature point extraction is robust to warping, and close to being invariant under warping because a number of points in the signal are repositioned canonically.

   Once again, however, we did not consider this for a practical watermarking scheme, because of the difficulties perceived in generalising to two dimensions. However we note that the watermarking schemes of Bas [Bas00] and Kutter *et al.* [KBE99] (discussed in Section 7) are in fact doing precisely this, although they do not present their algorithms in this way. They use Delauny triangulaton and Voronoi diagrams, respectively, to solve the problem of generalising to two dimensions.

c) A mathematical curiosity: just set

$$\mathcal{T}[f,g] = f^{-1}\,;g.$$

Invertibility follows simply, as does invariance under warping. In fact the transform is invariant under any invertible map on the signal domain. Suppose $w : \mathbb{R} \to \mathbb{R}$ is invertible, then

$$\mathcal{T}[w\,;f, w\,;g] = (w\,;f)^{-1};(w\,;g) = f^{-1};w^{-1};w;g = f^{-1};g = \mathcal{T}[f,g]$$

Finally, some elementary analysis shows that the transform is "robust" in the preservation of limit sense above, at least for real-valued signals, provided the signals are all continuous, and all necessary inverses exist.

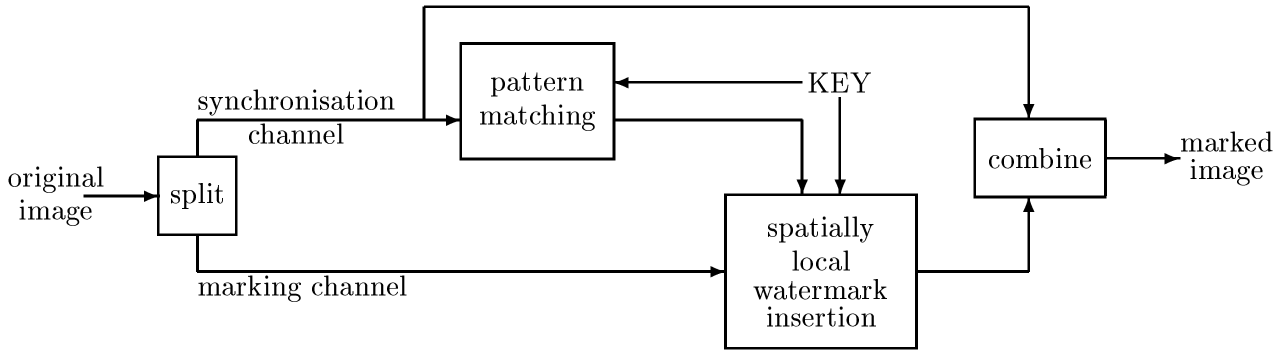
The third solution appears to help not at all: we cannot expect an arbitrary signal to be either 1-1 or onto (nor would it seem to any more likely for signals representing pictures), hence the inverse of a signal is not well defined. (Note however that this solution at least works without modification for signals of arbitrary dimension.)

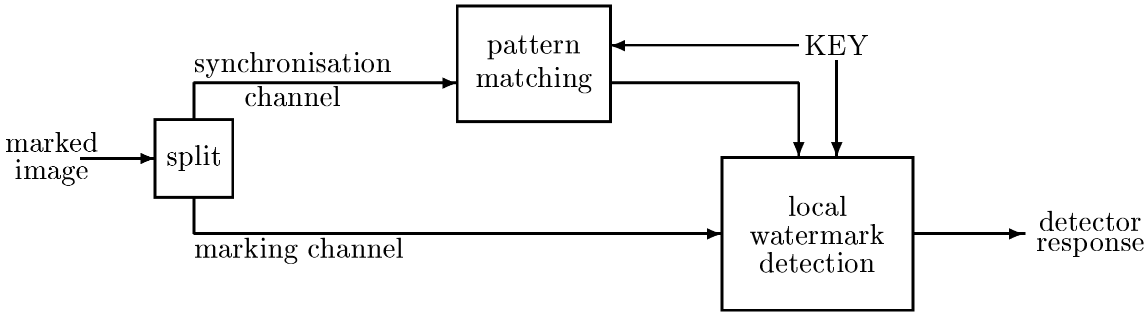
We show how these problems can be overcome to turn solution c) above into a practical watermarking scheme, robust to distortion. It is important that we did not specify the content of the signal representing the two channels (synchronisation and watermarking) making up the picture, although we had in mind that it should consist of a discrete map from spatial pixel locations to the value of the channel at that point.

Instead we choose to view each channel as a map from pixel locations to *features*; the signal describes the local content of the image at each point. The precise meaning of "local content" is not important for now. If the contents of the synchronisation channel are not so repetitive that identical local content occurs more than once then the signal will be 1-1 (we at worst expect a small number of identical local features, and could define $f^{-1}$ to choose randomly between them). The signal may well still not be onto, but we can work around this by setting $f^{-1}(x)$ to be the position of closest match, with respect to a suitable metric on "local content", for the feature $x$ in the channel $f$.

Using this paradigm for both the synchronisation channel and marking channel we arrive at a solution to the theoretical problem, which works equally well for two dimensional signals. Unfortunately a number of other weaknesses have been introduced, so that the solution to the mathematical problem is not exact. Firstly, if $f$ is a map from pixel locations to local content, then the warped version of the signal represented by $f$ is not quite equal to $w;f$ because the local content will itself be warped. Thankfully, in the case of imperceptible attacks like *StirMark*, the local effects of the warping are negligible (as long as the local content is genuinely local). Secondly, there is the possibility of multiple

Figure 3: The Dual Channel Watermark Insertion and Detection Schemes




features $x$ being matched to a single position in the synchronisation channel, or the local content of closely spaced positions in the marking channel overlapping. Because of these effects, the invertibility requirement cannot always be met exactly: given a transform $h$ and a signal $f$ it is only possible to construct a signal $g$ such that $\mathcal{T}[f, g]$ is approximately equal to $h$. This will not prove fatal, as long as the insertion and detection of the watermark signal (which will take place on the transformed signal) allows for noise in the transform, and it should do so if it expects to be robust to other distortions of the image as well as warping.

From here it is quite simple to turn the mathematical solution into a system for watermarking, albeit with significant details remaining. For the both insertion and detection of the watermark we break the image into synchronisation and marking channels. We use the watermarking secret key to generate a number of patterns, and a pattern matching algorithm to find the locations of the best matches for those patterns in the synchronisation channel. For watermark insertion, we make spatially local changes to the marking channel, possibly also based on the secret key, in the region around these locations, and combine the altered marking channel with the unaltered synchronisation channel to produce the watermarking image. For watermark detection we use the locations identified by pattern matching to recover the information from the spatially local watermarks. A diagrammatic representation of the processes is shown in Figure 3.

We would hope to make use of existing research to implement the pattern matching and local watermarking algorithms.

## 5 A Crude Implementation

To test the viability of this idea, a very simple implementation has been produced. Experimental results obtained using this implementation appear in the following section.

### Channel Splitting/Combination

Most colour images are stored as 24 bit RGB. We simply use the red components of each pixel of the colour image for the synchronisation channel, and the blue components for the watermarking channel. The green component we ignore throughout. The motivation is partly simplicity and partly because the blue component of an image is known to be perceptibly much less significant than the others[6]. It is also easier to work with signals taking one dimensional values, both as synchronisation and marking channels, for now.

### Pattern Matching and Pseudorandom Pattern Generation

Assuming the synchronisation channel is $m$ by $n$ pixels and represented by the signal $S_{ij}$ ($0 \leq i < m$ and $0 \leq j < n$), and given a $p$ by $q$ pattern similarly represented by $P_{ij}$ we compute $x$ and $y$ to maximise

$$\frac{\sum_{i=0}^{p-1} \sum_{j=0}^{q-1} P_{ij} S_{(i+x)(j+y)}}{\sqrt{\sum_{i=0}^{p-1} \sum_{j=0}^{q-1} S_{(i+x)(j+y)}^2}}.$$

This is equivalent to finding the origin of the $p$ by $q$ block of $S$ maximizing the simple (not DC-corrected) correlation coefficient with $P$. In fact the location we produce is given by $(x + p/2, y + q/2)$, where $x$ and $y$ maximize the above expression, so that the identified point is at the centre of the matching block.

We also included a fairly arbitrary filter, not allowing parts of the synchronisation channel very low in energy to be used in the pattern matching process. This is simply because such parts are too vulnerable to noise disrupting the matching process.

The number of patterns to be generated (i.e. the number of spatially-local watermarks to be inserted into the marking channel) is a variable parameter $N$. Initially we envisaged between 100 and 1000 as the likely range of values for $N$.

The key is used to seed a pseudorandom number generator, which produces the patterns to be matched. Each pattern is $3\mu$ by $3\mu$ pixels, consisting of 9 blocks, $\mu$ by $\mu$ pixels each, of constant intensity. $\mu$ is an integer parameter which could be modified. Using too small a value means that only edge features in the image would vary enough over such a

[6] For a discussion of why, with hindsight, this decision was probably very poor see the conclusions section.

short span of pixels to give a high correlation with many blocks, resulting in clustering of the locations output by the pattern matching part of the algorithm. This would cause many of the local watermarks to overlap and possibly interfere. Using too big a pattern may make the detector more sensitive to noise; at this stage we have not tested the performance of the detector with different parameters properly.

### Spatially Local Watermark Insertion

With the dual channel approach producing a list of locations in the image, the *Patchwork* algorithm serves as an excellent starting point for our watermark procedure. As noted in [NP98], the robustness of *Patchwork* suffers from the fact that the watermark added into the image is the small perturbation of individual pixels by small amounts. Because the noise added to neighbouring pixels is uncorrelated it amounts to low power white noise, and is easily removed by JPEG compression or small amounts of blurring.

We use a simple method to deal with this problem: instead of altering the value of single pixels we alter a square region around each one, inserting a pattern varying as a cosine wave (in both dimensions) with period the size of the square. This is more or less equivalent to adding noise to the $(1,1)$ component of the two dimensional DCT of the square around the point (although the function we used for this implementation is not exactly correct for the traditional DCT). This ensures that the noise inserted has its energy concentrated in the lowest possible frequency, hopefully the better to survive blurring and JPEG compression. (In more advanced implementations, we could use other low-frequency components of the DCT as well.) Following the *Patchwork* model, we will add multiples of this double cosine wave into the regions around half of the points located by the pattern matching, and subtract from the other half.

Given the $N$ locations identified by the pattern matching stage, say $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, let us also generate a set $\{z_1, \ldots, z_N\}$, where exactly half the $z_i$ are 1 and the other half $-1$. This can be generated from the secret key, although it makes no difference to security if we just set the $z_i$ to alternate between $\pm 1$. Representing the watermark channel by $W_{ij}$ ($0 \le i < m$ and $0 \le j < n$) before marking, and $W'_{ij}$ afterwards, we set:

$$W'_{ij} = W_{ij} + \alpha \sum_{t=1}^{N} z_t \chi(x_t - i, y_t - j) \cos\left(\frac{\pi(x_t - i)}{\nu}\right) \cos\left(\frac{\pi(y_t - j)}{\nu}\right)$$

where $\chi$ is the characteristic function

$$\chi(h,k) = \begin{cases} 1, & \text{if } -\nu \le h \le \nu \text{ and } -\nu \le k \le \nu \\ 0, & \text{otherwise.} \end{cases}$$

Here $\alpha$ and $\nu$ are parameters to the algorithm, the strength of the inserted watermark and half the width of the spatially local watermark respectively. In practise we clamped the absolute difference between $W'_{ij}$ and $W_{ij}$ at $\alpha$ (due to overlapping watermarks it could end up significantly — and very visibly — higher).

Watermark Detection

Since the inserted watermark affected not only the pixels at the points located by the pattern matching stage, but a square region around each, we must take this into account when computing the detection statistic. Because adding the two dimension cosine wave was (nearly) equivalent to adding to the $(1,1)$ component of the DCT of the square, we compute the detection statistic by recovering this $(1,1)$ component (modulo a constant scaling factor).

As above, we suppose that the $N$ locations identified by the pattern matching algorithm are $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, and that we have the same set $\{z_1, \ldots, z_N\}$ as used to insert the mark. Let $W_{ij}$ be the representation of the watermarking channel of the image presented to the detector.

For $(x, y)$ such that $0 \leq x < m$ and $0 \leq y < n$ we define

$$a_{xy} = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} W_{ij} \chi(x-i, y-j) \cos\left(\frac{\pi(x-i)}{\nu}\right) \cos\left(\frac{\pi(y-j)}{\nu}\right).$$

We calculate three statistics:

$$\bar{a} = \frac{\sum_{t=1}^{N} a_{x_t y_t}}{n} \qquad s_a^2 = \frac{\sum_{t=1}^{N} (a_{x_t y_t} - \bar{a})^2}{n-1} \qquad b = \sum_{t=1}^{N} z_t a_{x_t y_t}$$

Finally, the output of the detector is the value

$$d = \frac{b}{\sqrt{N s_a^2}}.$$

In order for a watermark detector to be of any use we must know how to interpret its detection values. For single bit watermarks, detection can be viewed as a hypothesis test. The null hypothesis must be that there is no watermark, and the alternative hypothesis that there is a watermark. We must examine the distribution of the detector statistic $d$ under these hypotheses.

Unfortunately the theoretical distributions for detector statistics are often not possible to calculate precisely. Usually it is necessary to assume that pixel values in an image are independently distributed, which is extremely unlikely, and even then we can often only calculate a limiting distribution, which we hope is valid for large values of the relevant parameter.

Here we assume that the pattern matching algorithm chooses uniformly from among the available pixel locations (taking into account that the locations it finds must be at least $3\mu/2$ from the edge of the image). Thus picking the $a_{x_t y_t}$ can be seen as random samples from a discrete distribution which we will call $A$.

It is easy to see that $\mathrm{E}[d|H_0] = 0$. Furthermore, because $s_a^2$ is an unbiased estimator for the variance of $A$, we know that, in the limit as $N \to \infty$, $\mathrm{Var}[d|H_0] \to 1$. Unfortunately we need to know the full distribution of $d|H_0$, or at least have an accurate estimate of the cumulative distribution function at the tails of the distribution, in order to reliably estimate the probability of a type-I error, a false positive watermark detection.

We could appeal to the Central Limit Theorem. That tells us that, asymptotically, $d|H_0 \dot{\sim} N(0,1)$. However the values of $N$ we use will not be so large (100-1000) that we can use the asymptotic distribution without further evidence that it is valid.

If $A$ formed a normal distribution then we could make use of a well-known statistics theorem (for example an adaptation of the results appearing in [Ric95, §6.3]) to show that $d|H_0 \sim t_{N-1}$. However $A$ is a very dubious candidate for a normal distribution, because some experimentation shows it to be limited in range, bimodal, and skew! Nonetheless we may hope that the approximation leads to a valid conclusion, because the $t$ statistic is known to be quite robust to non-normal population data. The $t$ distributions are shaped similarly to the standard normal distribution, but has more probability in the tails, and as $n$ tends to infinity, $t_n$ converges to the normal distribution. The extra weight in the tails of the $t$ distributions is a reflection of the additional uncertainly in using an estimate for the population variance to standardise the sample data.

In the absence of exact analysis we will have to use experimental data to provide evidence for the distribution of $d|H_0$, albeit with a suspicion that it may well be either the standard normal distribution or a thick-tailed variation.

Even more difficult is calculation of the distribution of the detector statistic on watermarked images, which is the alternative hypothesis of this test. One can show that $\mathrm{E}[d|H_1] = K(\nu)\alpha\sqrt{N}$, where $K(\nu)$ depends only on $\nu$, as long as none of the local watermarks overlap. In practice some overlapping will occur and the value of $d$ will be rather lower.

As long as we know that $d|H_1$ can be made to have a mean significantly above zero, there is not the same urgency to calculate an exact theoretical distribution for it (and we would probably not be able to anyway). We would not be trying to estimate the probability of type-II errors, since we would want to adjust to watermark parameters to make a type-II error impossible (always inserting a watermark strong enough to be recognisable). We are much more interested in how the detection statistic degrades after attacks designed to destroy the watermark, and this can only be gained by experimental evidence.

# 6 Experimental Results

Because the method for implementing a dual channel watermarking scheme outlined in the previous section is so basic, we did not perform totally rigorous testing at this stage. The partial results obtained, however, are perhaps surprisingly good given the poor algorithm and uninformed selection of parameters. It appears that this implementation is functioning well as a watermarking scheme in its own right.

Figure 4: The three images used in testing the watermark scheme: "Lenna" (256×256), "Merton" (384×384), and "Cow" (512×512)

Therefore these results serve as an indication of the merits (and drawbacks) of this approach, and give some confidence that a properly refined version of the scheme could be very effective. We hope to produce such a scheme after further research, and to test it fully.

### Test Pictures

Three pictures were used: the standard image "Lenna" (a detail taken from a nude photograph of a Swedish Playboy model particularly beloved of the watermarking community), scaled to 256 pixels square, a picture of Merton College, Oxford (taken by the author) scaled to 384 pixels square, and a bovine closeup (also by the author) at 512 pixels square. They are shown in Figure 4. Between them the images offer areas of almost flat colour (the sky in the Merton image), textures varying from very shallow (the background to Lenna) to very deep (the grass behind the cow), strong vertical and horizontal features (the Merton tower), and a good range of colours. They also span a reasonable range of sizes, although further work will be needed to see if smaller images can be marked reliably, and whether the computational complexity of larger images merits special consideration.

### Parameters for Marking

The pictures were marked using the basic implementation of the dual channel scheme presented in the previous section. The same parameters were used for each: $N = 100$, $\alpha = 15$, $\mu = 32$, $\nu = 16$. These parameters were picked using a very small amount of experimentation; further work will indicate how they may be optimized and should improve the performance of even this crude implementation. The maximum distortion in pixel values was therefore 15 (out of 256 levels of blue per pixel); this is quite a substantial alteration to the signal. The peak-signal to noise ratio of the blue channel is included in the table below. This cannot be compared directly with the PSNR of other watermarking schemes inserted in grayscale images (or the luminescence component of colour images) because the blue component is only a part of the whole signal. The author is not aware of a comparable measure of noise in colour images. In any case the PSNR is a poor measure of distortion because it fails to take into account the particular characteristics

Figure 5: The three watermarked images

of the human visual system which highlight or degrade noise in different contexts. The response of the detector to the marked images is also included in the table.

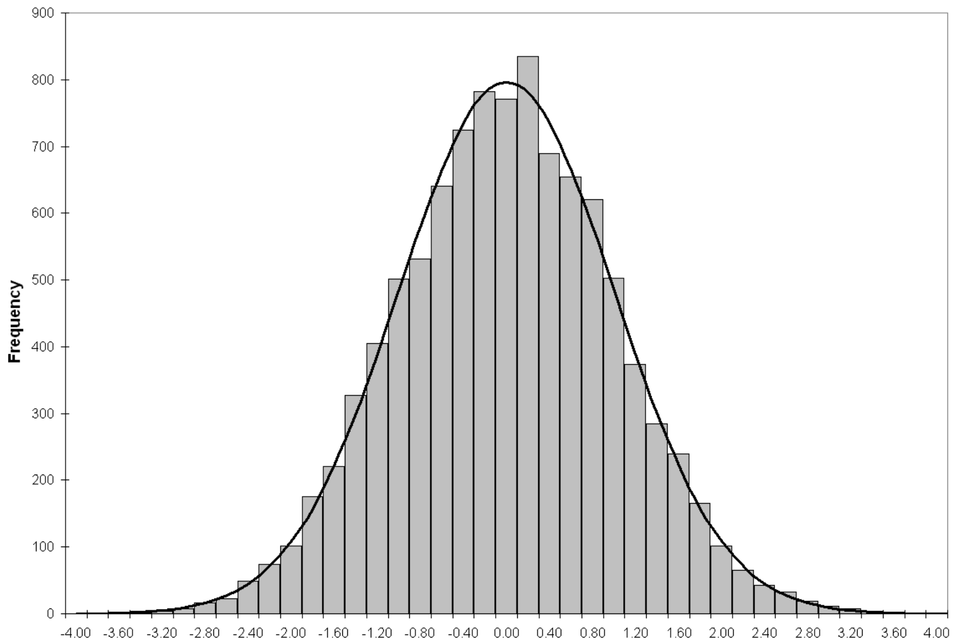
| Image | Lenna | Merton | Cow |
|---|---|---|---|
| Size | 256×256 | 384×384 | 512×512 |
| $\mathrm{PSNR}_{\mathrm{Blue}}$ | 28.50dB | 33.95dB | 36.14dB |
| Detector Response | 7.56 | 7.95 | 7.32 |

The three watermarked images are shown in Figure 5. Because of the lack of perceptual masking, and the selection of the blue channel for the watermark, the watermarks inserted under these parameters are probably more visible to the eye than those of other schemes. In particular the sky of the Merton image, and some of the background of Lenna, show faint but visible blue/yellow bands where the watermark has been located. We do not consider this to be a serious defect for two reasons. Firstly a mature implementation of the dual channel watermarking approach will very likely not use the plain blue channel for the watermarks, or at the least will use a perceptual mask to reduce the strength of the watermark in flat areas. Secondly, a simple experiment whereby the offensive parts of the watermark were removed by hand in an image processing package, using judicious local blurring, leads to only a minor degradation in detector response.

### Null Hypothesis Detector Response

The first thing to test is the response of the detector to unmarked images. The detector statistic has been designed to produce zero mean and unit variance in this case, but we have already commented that to estimate the probability type-I errors (false positives) the full distribution is needed. The calculations in Section 5 show that, while a theoretical derivation of this distribution seems unlikely, we might expect roughly a normal distribution (if we admit some rather implausible assumptions of independence). The use of an estimate for the population variance would be expected to lead to a heavier-tailed distribution than normal, such as a $t$ distribution, and because this may substantially affect the probabilities in the distant tails — precisely the area we are interested in to give

Figure 6: Histogram of 10000 observations of the watermark detector applied to an unmarked image, compared to the standard normal distribution




the probability of false positives — it is important to investigate whether this happens or not.

We performed an experiment to estimate the distribution, with the unmarked image Lenna being tested against 10000 watermark keys. The resulting distribution of scores, perhaps surprisingly, turned out to be a very close fit to the standard normal distribution. The histogram of the observed detector responses is plotted in Figure 6, along with the density function of the standard normal distribution. This gives a subjective measure of how close a fit the distributions are. Numerical evidence for the normality of the detector on unmarked images (on this unmarked image anyway) is as follows: from 10000 samples the observed mean was 0.00025, and the observation unbiased estimate of population variance was 1.0052. The maximum observed value was 3.98 (the next highest was 3.26). The Anderson-Darling normality test coefficient was 0.40, a p-value of 0.37 (see [Ste74] for a description of this test, and why it is appropriate in this case). None of these statistics is evidence for rejecting the hypothesis of a standard normal distribution. The only significant statistic was the kurtosis of the distribution, which was -0.087, with a p-value of 0.068. Kurtosis is a measure of spread, and this result is an indication that the tails of this distribution may be slightly *lighter* than normal.

For comparison, the same tests run on data generated from a $t_{99}$ distribution, roughly what we might expect if the distribution $A$ from the previous section were normal, produced an Anderson-Darling coefficient of 0.71 (a p-value of 0.066) and Kurtosis of 0.32

(p-value 0.0011). This is at least strong evidence that the distribution is much closer to standard normal than $t_{99}$.

Of course we would prefer to run tests on 10000 different images, rather than the same image with 10000 different keys, giving some confidence that the observed distribution of detector response was not an artifact of the Lenna image. Sadly, a large library of images was not available to the author. Hopefully this can be provided in future work. We did perform an additional experiment, testing the response to images marked with the wrong key. The results were similar.

If we believe this evidence, then we can make an informed choice about the thresholds to set for positive detection, based on the desired level of false detection rate. We use normal tail probabilities to give an upper bound of the probability of false positives for various detection thresholds. The results are summarised in the following table.

| P(False Positive) | Detection Threshold | P(False Positive) | Detection Threshold |
|---|---|---|---|
| $10^{-3}$ | 3.09 | $10^{-7}$ | 5.20 |
| $10^{-4}$ | 3.72 | $10^{-8}$ | 5.61 |
| $10^{-5}$ | 4.27 | $10^{-9}$ | 6.00 |
| $10^{-6}$ | 4.75 | $10^{-10}$ | 6.36 |

The table was computed using a numerical approximation in [AS70]. We emphasise that, if the null hypothesis distribution does indeed have lighter tails than standard normal then the probabilities of false detection will be *lower* than listed.

Which false positive probability is appropriate will depend on the use of the watermark. To "prove" ownership of an image to a judge or jury, a probability of $10^{-4}$ may well suffice. For an automated spider scanning files on the internet, expecting to scan millions of online images against a library of thousands of copyright images, many orders of magnitude lower would be required. For these experiments, we selected $10^{-7}$ as our target probability for false positives, aiming to insert a watermark of sufficient strength for the detection statistic to be above 5.20 after most of the attacks we wish to be robust to. This motivated the choice of $\alpha = 15$ and $N = 100$ as watermarking parameters.

### Experimental Evidence of Robustness

The three marked images were distorted in a variety of ways and tested by the watermark detector. The results are reported in the following tables. Most of the test images were generated by the automatic benchmarking facility of the *StirMark* 3.1 package, which performs a number of common attacks, all of them optionally followed by mild JPEG compression (we distil out the more interesting and representative results rather than reproducing the rather exhaustive full list). Other tests were hand produced using Adobe Photoshop version 5. Each table is followed with comments.

Blurring/Sharpening attacks
Key: MF=Median Filter, GB=Gaussian Blur, Sh=Sharpen, -$n$=filter of $n$ pixels square

| Image | Without JPEG | | | | With JPEG QF=90 | | | |
|---|---|---|---|---|---|---|---|---|
| | MF-2 | MF-4 | GB-3 | Sh-3 | MF-2 | MF-4 | GB-3 | Sh-3 |
| Lenna | 6.54 | 6.32 | 6.61 | 5.52 | 6.26 | 6.30 | 6.46 | 5.14 |
| Merton | 6.13 | 5.86 | 6.88 | 5.33 | 5.97 | 5.99 | 6.94 | 5.09 |
| Cow | 6.37 | 6.41 | 6.67 | 6.37 | 6.03 | 5.77 | 6.28 | 5.56 |

We would expect the watermark to be robust to such operations, because of its low frequency power spectrum. The above attacks were generated by *StirMark*. Some authors test against much higher levels of blurring so we performed some additional tests using Photoshop, finding that the watermark still produced a detection value of over 5.20 under a median filter of *radius* 5. At this level of filtering, the distortion is severe.

Random row and column removal

| | Without JPEG | | | With JPEG QF=90 | | |
|---|---|---|---|---|---|---|
| | 1row/1col | 1row/5col | 17row/5col | 1row/1col | 1row/5col | 17row/5col |
| Lenna | 6.66 | 6.33 | 5.00 | 6.38 | 5.78 | 5.12 |
| Merton | 7.71 | 6.40 | 5.99 | 6.41 | 6.00 | 5.80 |
| Cow | 7.18 | 6.79 | 6.25 | 6.49 | 6.48 | 5.84 |

Cropping

| | Without JPEG | | | | | With JPEG QF=90 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1% | 2% | 5% | 10% | 20% | 1% | 2% | 5% | 10% | 20% |
| Lenna | 6.75 | 6.24 | 5.91 | 5.71 | 4.45 | 5.96 | 5.43 | 5.74 | 5.00 | 3.95 |
| Merton | 7.25 | 7.18 | 6.88 | 5.48 | 5.35 | 6.23 | 5.88 | 5.95 | 4.41 | 4.29 |
| Cow | 7.09 | 6.82 | 6.45 | 6.10 | 4.92 | 6.48 | 6.23 | 5.60 | 5.49 | 4.66 |

The watermark seems very robust to both row and column removal, and small amounts of cropping. Because of the synchronisation channel, the effects of row and column removal is to mildly jitter some of the spatially local watermarks, without displacing them. Their low frequency means that we expect only a mild degradation in detector response. Similarly, cropping will delete the synchronisation necessary for some of the outer spatially local watermarks (in these tests, the middle of the image is retained in each case). It was rather disappointing that cropping by 20% caused the detector to fail most of the time; this is probably because the size of the patterns being matched against (96 pixels square) already ruled out the outer region of the image (48 pixels around each edge) for locating the watermarks. In the case of the Lenna image, cropping by 20% removed 36% of the information in the image but about 55% of the available positions for watermark location. Better performance against cropping can probably be obtained with a better choice of watermarking parameters.

Compression, by colour reduction or JPEG

| | Colour reduction | | JPEG Compression (quality factor) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 256 colours | 32 colours | 90 | 70 | 50 | 35 | 20 | 10 |
| Lenna | 6.17 | 5.15 | 6.37 | 5.30 | 6.17 | 5.70 | 4.85 | 3.20 |
| Merton | 7.08 | 4.55 | 6.32 | 6.66 | 6.32 | 5.12 | 5.38 | 2.13 |
| Cow | 6.39 | 3.92 | 6.60 | 6.22 | 5.89 | 5.77 | 4.63 | 2.83 |

The watermark proves robust to colour reduction to 8 bits, but fails when reduced as far as 5 bits. At 32 colours, a photographic image becomes rather cartoon-like, so this is not a serious weakness. Robustness to JPEG compression is good down to a quality factor of around 35, tailing off quite quickly below this level. Some authors quote robustness to JPEG compression by various bit rates for the compressed image, as opposed to by quality factors. The bit rate resulting from each quality factor vary from image to image, but in these cases we observed that robustness was achieved down to around 0.7 bits per pixel. The quality factor corresponding to 1 bit per pixel (quoted by some authors as a reasonable target for watermark robustness) was between 40 and 50, at which rates good robustness was achieved.

Rescaling

| | Shrinkage-enlargement factor | | | | |
|---|---|---|---|---|---|
| | 70% | 50% | 30% | 15% | 10% |
| Lenna | 7.00 | 6.70 | 6.47 | 5.68 | 4.26 |
| Merton | 7.03 | 6.60 | 6.88 | 5.61 | 6.20 |
| Cow | 6.93 | 6.93 | 6.30 | 5.70 | 4.50 |

Rescaling was performed using Photoshop's bicubic resampling algorithm. We see that, in this situation where the scaling factor is known and inverted, the watermark is extremely robust. Reducing an image to 15% (in each dimension) deletes over 97% of the information in it, but the watermark is still detectable. This is because of the low frequencies used.

*StirMark* Random Geometrical Distortions

| | Without JPEG | | | | With JPEG QF=90 | | | | Default Parameters |
|---|---|---|---|---|---|---|---|---|---|
| | `-b1` | `-b2` | `-b4` | `-b8` | `-b1` | `-b2` | `-b4` | `-b8` | |
| Lenna | 6.25 | 6.26 | 5.14 | 4.59 | 6.00 | 6.10 | 4.80 | 4.60 | 5.64 |
| Merton | 6.39 | 5.98 | 5.18 | 4.47 | 5.87 | 5.88 | 5.04 | 3.70 | 5.55 |
| Cow | 6.71 | 6.11 | 5.23 | 3.83 | 5.37 | 5.38 | 4.97 | 3.00 | 6.04 |

The `-b` options allow the variation of the "bending factor", the pixel displacement caused by *StirMark*. Default parameters are `-b2`, JPEG compression at 90%, and also some histogram equalisation which appears not to take place when the `-b` option is specified. We see that the watermarking scheme is robust to the standard parameters and also to higher bending factors, although defeated by as much as `-b8`, at which level the image is quite badly warped. Nonetheless, for images without many straight edges, this might be used as a viable attack and further research is indicated to see if the performance can be enhanced further against large bending factors.

Rotation

| | Rotated and cropped to square | | | | | Rotated, cropped, and rescaled | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0.25° | 0.5° | 1° | 2° | 5° | 0.25° | 0.5° | 1° | 2° | 5° |
| Lenna | 6.48 | 6.25 | 5.33 | 4.66 | 4.04 | 6.25 | 6.56 | 5.81 | 5.84 | 2.23 |
| Merton | 6.93 | 7.15 | 6.57 | 5.61 | 4.48 | 7.12 | 6.87 | 6.48 | 5.86 | 2.86 |
| Cow | 6.75 | 6.48 | 6.41 | 5.95 | 4.38 | 6.30 | 6.43 | 6.18 | 5.51 | 2.17 |

We would not expect the watermark to be robust against these attacks, because we did not build any rotation robustness into either the pattern matching algorithm or the spatially local watermark. Nonetheless, against small rotations of up to 1–2°, the detection value is still above the threshold, whether the image is rescaled back to its original size after rotation or not. Above 5°, the detector response was very poor. Nevertheless, we observe robustness to "imperceptible" rotations.

Uncorrected scaling and aspect ratio change

| | Scaling | | | | Aspect Ratio Change | | | |
|---|---|---|---|---|---|---|---|---|
| | 99% | 98% | 95% | 90% | 99% | 98% | 95% | 90% |
| Lenna | 6.00 | 6.06 | 4.75 | 3.46 | 6.67 | 6.51 | 5.60 | 3.82 |
| Merton | 6.51 | 5.90 | 4.84 | 4.30 | 6.90 | 6.58 | 6.17 | 6.11 |
| Cow | 6.47 | 6.67 | 5.65 | 3.55 | 6.79 | 6.84 | 6.06 | 5.02 |

Similarly we would not expect robustness to much scaling or change of aspect ratio, because neither the pattern matching nor watermark detection is scale-invariant. Although robust at scaling of 98%, the detection response quickly falls off at lower factors, especially in the smaller images. The effect of aspect ratio change is less severe then scaling, simply because aspect ratio change does not cause as much distortion (one dimension is left unaltered).

Uncorrected scaling is the weakness of this watermarking scheme which must be addressed most urgently, in further work. (We note, however, that very few published watermarking schemes are genuinely robust to scaling and particularly change of aspect ratio, without registration to invert the effects. Even robustness to 98% scaling, which we have observed with this scheme, is rare.)

Other imperceptible linear transformations

| | Shear by 5% | | Linear transformations | | |
|---|---|---|---|---|---|
| | one dimension | both dimensions | #1 | #2 | #3 |
| Lenna | 5.54 | 5.32 | 5.90 | 5.56 | 5.91 |
| Merton | 6.40 | 5.45 | 6.72 | 6.86 | 6.78 |
| Cow | 6.30 | 5.55 | 6.15 | 6.28 | 5.94 |

These are generated by *StirMark*. The linear transformations are given by the matrices

$$\begin{pmatrix} 1.010 & 0.013 \\ 0.009 & 1.011 \end{pmatrix}, \quad \begin{pmatrix} 1.007 & 0.010 \\ 0.010 & 1.012 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 1.013 & 0.008 \\ 0.011 & 1.008 \end{pmatrix}.$$

Robustness is observed in each case.

### Summary of Experimental Evidence

A more sophisticated implementation of the dual channel approach would merit more thorough testing than we have given here. Our results, however, indicate that this implementation is robust to blurring attacks, row and column removal, cropping down by about 10%, colour reduction, JPEG compression to well below 1 bit per pixel (quality factor of at least 35), rescaling down as low as 15% when the scaling factor is known, *StirMark* with default parameters and also up to increased bending factor of at least 4, rotation of 1–2°, scaling of 98% when the scaling is not inverted, aspect ratio change of 95%, and other imperceptible shearing and general linear transformations.

The scheme is not robust to cropping of 20%, JPEG compression at very low quality factors such as 20, rescaling down to 10% even when the rescaling is inverted, *StirMark* with a bending factor of 8, rotation of 5° or over, uncorrected scaling of 95% and aspect ratio change of 90%. See Section 8 for ideas on how robustness to rotation and scaling may be achieved.

We consider that the most significant achievement is the blind robustness to *StirMark*, which seems rare among watermarking schemes, and the fact that the detector response is degraded so little by the various attacks. The unmodified watermarked images give detector responses around 7.5, and most of the attacks are producing values around 6, only 20% lower. This contrasts sharply with the performance of most other watermarking schemes, where the detector response is usually extremely high in the unattacked case, and drops sharply after attack, suggesting that a combination of attacks may defeat the watermark detector. If robustness to rotation and scaling can be achieved, and no targeted attack designed especially to disrupt this sort of watermarking scheme comes to light, the dual channel approach may offer a blind watermarking system more powerful than heretofore available.

# 7 Related Research

We should say that a number of papers have appeared with similar goals to this one, sometimes using related methodology. We give a brief survey for comparative purposes[7].

In [NP98] a modification of the *Patchwork* algorithm is proposed to make for greater robustness, especially to lossy compression. The paper also contains an investigation of the theoretical distributions of the detector response in the unmarked and marked cases, although it relies on the Central Limit Theorem. Furthermore, it notes the extreme weakness of *Patchwork* to blurring and compression, suggesting as we have done that lower frequency watermarks are needed. It also notes the weakness to cropping and advocates line-by-line correlation of the tested image with the watermark sequence, to try to undo the effects of row or column removal. However it does not deal with any other geometric distortions.

Closely related work includes Bas' thesis [Bas00] (and related papers [BCM00] and [BCD99]). This work includes a number of watermarking algorithms which include the detection of feature points. The system with most in common with the ideas we describe in this paper, which is in truth a great deal more sophisticated, involves the identification of feature points, Delauny triangulation of the image based on these feature points,

[7] Because of the extremely high number of publications in the field of watermarking, the author may have missed out a paper using a closely related — or even identical — idea to that presented here. He would be grateful for any additional references.

and the insertion of spatial-domain watermarks into these triangles independent of their location and orientation. It also employs a perceptual mask to maximise energy in the imperceptible watermark, and Wiener filters to allow for blind detection. Sadly, at least in [Bas00], the performance of this watermarking system is rather under-reported, but the limited data available suggests that its performance is poorer than we might expect.

Another closely related idea is presented in [KBE99], where a small number of feature points are identified and used to segment the image via Voronoi diagrams, the watermark being inserted spatially into the separate segments in a fashion not dissimilar to the above paper. The authors even consider colour images, proposing the use of the blue component of the image to contain the watermark, exactly as we have used, but they do not explicitly the image into a synchronisation and watermarking channel. Aside from the more complicated use of segmentation (which ensures that the whole image is used to spread the spatial watermarks, but will be less robust under the incorrect detection of a small number of feature points) an important difference is that no mention is made of using the secret key to parameterise the location of the feature points. The paper is rather short and details are not given, but it appears that this would be particularly vulnerable to an attack aimed at those feature points. Furthermore, although "very promising" preliminary results are claimed, none are published. It seems surprising that no follow-ups have been published in the 2 years following this paper, but perhaps this author has simply missed them.

Another attempt to form blind watermarking schemes resistant to the *StirMark* attack is [RMvO99], which marks an image by introducing geometrical distortions. Although it is robust to *StirMark* attacks, the scheme is highly vulnerable to other attacks such as lossy compression.

One other paper with some significance to this work is [PBBC99], which exploits the correlation between the different colour components of a colour image. The use it makes of this correlation is quite different to that presented here, however, and only serves to grant greater resistance to lossy compression. The scheme is not robust to any geometrical transformations. Another paper dealing with colour images is [KJB97], which uses the luminescence component of the image as a crude perceptual mask to modulate alteration of the blue component. This is one of the early watermarking schemes and not robust to geometrical transformations.

# 8 Conclusion and Directions for Further Research

We have described a weakness in the *StirMark* attack, where the components of a colour image must undergo the same warping, and described how this may be exploited in a dual channel watermarking scheme. Our implementation of the dual channel scheme was very basic, but nonetheless performed well under testing.

Many ideas for further research have been indicated. We list some of the more obvious directions.

An easy improvement to make should be to select a better method of splitting the colour image into synchronisation and watermarking channels. Our choice, using the red and blue components respectively, may have been one of the worst possible, with hindsight. For one thing, we are wasting the information carried in the green component, which is the most significant perceptually. Also, although the blue component is the perceptually least significant, it is also the worst degraded by JPEG compression. We noted that alterations in the blue channel could cause visible blue/yellow banding, presumably because the eye is quite sensitive to that chromacity change. Finally, some very tentative experimentation suggests that the blue component is the one which is the most mangled by printing on a colour inkjet printer, and that our implementation does not survive the print-scan process mainly for this reason.

It is clear that the dual channel approach leads to a watermarking scheme which can fail in two ways. Either the spatially local watermarks can be removed, or the synchronisation disrupted so that the pattern matching produces the wrong locations. Fuller analysis of the performance of the detector described in this paper should include a breakdown of the effects of attacks on the two parts.

Some ideas to improve robustness of the watermarking scheme: we did not design it to be robust to rotation, and found that it was defeated by rotations of $5^\circ$. However it should be quite easy to build this robustness into the implementation. Just ensure that the patterns matched against are rotationally symmetric, and that the spatially local watermarks inserted are either rotationally symmetric or inserted in some other way to made their detection after rotation simple. We could make use of the work in [LWB$^+$00] for the latter.

Robustness to scaling would be harder. Making the pattern matching truly scale-invariant would probably be a computationally expensive process, but we may be able to ensure that the matching performed during both insertion and detection is performed at a canonical scale, perhaps using the techniques of [AT00]. There are a number of other ideas to chase up. Making the spatially local watermarks scale invariant is harder, but if the scale factor can be estimated using the pattern matching part of the watermark process then it would not be necessary. It would also not be very computationally expensive to perform tests for the local watermarks at a variety of different scales. The same cannot be said for the pattern matching, however.

Finally we may wish to consider efficiency. The time taken to test for a watermark in an image of $P$ pixels is proportional to $P$ (the total time for testing is also roughly proportional to $N$, the number of small watermarks inserted). In practice we found that it took 2–3 seconds to test the Lenna image, and this rose to around 10 seconds for the Merton image. For larger images this could become infeasible. A simple solution is to scale down the synchronisation channel of large images before pattern matching (both at insertion and detection).

## Acknowledgements

The author is supported by a Junior Research Fellowship at University College, Oxford.

## References

[AS70] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions.* Dover Publications, Inc., New York, 1970.

[AT00] Masoud Alghoniemy and Ahmed H. Tewfik. Geometric distortion correction in image watermarking. In Ping Wah Wong and Edward J. Delp, editors, *Security and Watermarking of Multimedia Contents II*, volume 3971 of *Proceedings of SPIE*, pages 82–89, 2000.

[Bas00] Patrick Bas. *Méthodes de Tatouage d'Images Fondées sur le Contenu.* PhD thesis, LIS de Grenoble, 2000.

[BCD99] Patrick Bas, Jean-Marc Chassery, and Franck Davoine. Geometrical and frequential watermarking scheme using similarities. In Ping Wah Wong and Edward J. Delp, editors, *Security and Watermarking of Multimedia Contents*, volume 3657 of *Proceedings of SPIE*, pages 264–272, 1999.

[BCM00] Patrick Bas, Jean-Marc Chassery, and Benoit M. Macq. Robust watermarking based on the warping of pre-defined triangular patterns. In Ping Wah Wong and Edward J. Delp, editors, *Security and Watermarking of Multimedia Contents II*, volume 3971 of *Proceedings of SPIE*, pages 99–109, 2000.

[BGM95] Walter Bender, Daniel Gruhl, and Norishige Morimoto. Techniques for data hiding. In Wayne Niblack and Ramesh C. Jain, editors, *Storage and Retrieval for Image and Video Databases III*, volume 2420 of *Proceedings of SPIE*, pages 164–173, 1995.

[CKLS96] Ingemar J. Cox, Joe Kilian, Tom Leighton, and Talal Shamoon. A secure, robust watermark for multimedia. In Ross Anderson, editor, *Proceedings of First International Workshop on Information Hiding*, volume 1174 of *Lecture Notes in Computer Science*, pages 183–206. Springer, 1996.

[CMYY98] Scott Craver, Nasir Memon, Boon-Lock Yeo, and Minerva M. Yeung. Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. *IEEE Journal of Selected Areas in Communications*, 16(4):573–586, 1998. Special Issue on Copyright and Privacy Protection.

[DBHC99] F. Davoine, P. Bas, P.-A. Hébert, and J.-M. Chassery. Watermarking et résistance aux déformations géométriques. In J.-L. Dugelay, editor, *Cinquièmes journées d'études et d'échanges sur las compression et la représentation des signaux audiovisuels (CORESA'99)*, 1999.

[DP00] Jean-Luc Dugelay and Fabien A. P. Petitcolas. Possible counter-attacks against random geometric distortions. In Ping Wah Wong and Edward J. Delp, editors, *Security and Watermarking of Multimedia Contents II*, volume 3971 of *Proceedings of SPIE*, pages 338–345, 2000.

[FBS98] Jiri Fridrich, 2 Lt Arnold C. Baldoza, and Richard J. Simard. Robust digital watermarking based on key-dependent basis functions. In D. Aucsmith, editor, *Proceedings of Second International Workshop on Information Hiding*, volume 1525 of *Lecture Notes in Computer Science*, pages 143–157. Springer, 1998.

[GB98] Daniel Gruhl and Walter Bender. Information hiding to foil the casual counterfeiter. In D. Aucsmith, editor, *Proceedings of Second International Workshop on Information Hiding*, volume 1525 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 1998.

[HSG99] Frank H. Hartung, Jonathan K. Su, and Bernd Girod. Spread spectrum watermarking: Malicious attacks and counterattacks. In Ping Wah Wong and Edward J. Delp, editors, *Security and Watermarking of Multimedia Contents*, volume 3657 of *Proceedings of SPIE*, pages 147–158, 1999.

[JDJ99] Neil F. Johnson, Zoran Duric, and Sushil Jajodia. Recovery of watermarks from distorted images. In Andreas Pfitzmann, editor, *Proceedings of Third International Workshop on Information Hiding*, volume 1768 of *Lecture Notes in Computer Science*, pages 318–332. Springer, 1999.

[KBE99] M. Kutter, S. Bhattacharjee, and T. Ebrahimi. Towards second generation watermarking schemes. In *Proceedings of the 6th International Conference on Image Processing (ICIP'99)*, 1999.

[KJB97] Martin Kutter, Frederic Jordan, and Frank Bossen. Digital signature of color images using amplitude modulation. In *Storage and Retrieval for Image and Video Databases*, volume 3022 of *Proceedings of SPIE*, pages 518–526, 1997.

[Kut98] M. Kutter. Watermarking resisting to translation, rotation and scaling. In A. G. Tesher, editor, *Multimedia Systems and Applications*, volume 3528 of *Proceedings of SPIE*, pages 423–431, 1998.

[LK01] P. Loo and N. G. Kingsbury. Motion estimation based registration of geometrically distorted images for watermark recovery. In Ping Wah Wong and Edward J. Delp, editors, *Security and Watermarking of Multimedia Contents III*, volume 4314 of *Proceedings of SPIE*, 2001.

[LLHS99] Chun-Shien Lu, Hong-Yuan Mark Liao, Shih-Kun Huang, and Chwen-Jye Sze. Cocktail watermarking on images. In Andreas Pfitzmann, editor, *Proceedings of Third International Workshop on Information Hiding*, volume 1768 of *Lecture Notes in Computer Science*, pages 333–347. Springer, 1999.

[LvD98] Jean-Paul M. G. Linnartz and Marten van Dijk. Analysis of the sensitivity attack against electronic watermarks in images. In D. Aucsmith, editor, *Proceedings of Second International Workshop on Information Hiding*, volume 1525 of *Lecture Notes in Computer Science*, pages 258–272. Springer, 1998.

[LWB$^+$00] Ching-Yung Lin, Min Wu, Jeffrey A. Bloom, Ingemar J. Cox, Matt L. Miller, and Yui Man Lui. Rotation-, scale- and translation-resilient public watermarking for images. In Ping Wah Wong and Edward J. Delp, editors, *Security and Watermarking of Multimedia Contents II*, volume 3971 of *Proceedings of SPIE*, pages 90–98, 2000.

[NP98] N. Nikolaidis and I. Pitas. Robust image watermarking in the spatial domain. *Signal Processing*, 66(3):385–403, 1998.

[OP97] J. J. K. O'Ruanaidh and T. Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. In *IEEE Signal Processing Society 1997 International Conference on Image Processing (ICIP'97)*, pages 536–539, 1997.

[PAK98] Fabien A. P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on copyright marking systems. In D. Aucsmith, editor, *Proceedings of Second International Workshop on Information Hiding*, volume 1525 of *Lecture Notes in Computer Science*, pages 218–238. Springer, 1998.

[PBBC99] A. Piva, M. Barni, F. Bartolini, and V. Cappellini. Exploiting the cross-correlation of RGB-channels for robust watermarking of color images. In *Proceedings of the 6th International Conference on Image Processing (ICIP'99)*, pages 306–310, 1999.

[PZ98] Christine I. Podilchuk and Wenjun Zeng. Image-adaptive watermarking using visual models. *IEEE Journal on Selected Areas in Communications*, 16:525–539, 1998.

[Ric95] John A. Rice. *Mathematical Statistics and Data Analysis.* Duxbury Press, Belmont, California, second edition, 1995.

[RMvO99] Peter M. Rongen, Maurice J. Maes, and Kees W. van Overveld. Digital image watermarking by salient point modification: Practical results. In Ping Wah Wong and Edward J. Delp, editors, *Security and Watermarking of Multimedia Contents*, volume 3657 of *Proceedings of SPIE*, pages 273–282, 1999.

[SM] StirMark homepage: `http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/`.

[Ste74] M. A. Stevens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69:730–737, 1974.

[WSK98] Houng-Jyh Mike Wang, Po-Chyi Su, and C.-C. Jay Kuo. Wavelet-based digital image watermarking. *Optics Express*, 3(12):491–496, 1998.

[XBA98] Xiang-Gen Xia, Charles G. Boncelet, and Gonzalo R. Arce. Wavelet transform based watermarking for digital images. *Optics Express*, 3(12):497–511, 1998.