BENCHMARKING TOOLS FOR FAIRLY COMPARING WATERMARKING ALGORITHMS

Stewart I. Fraser
Department of Engineering
University of Aberdeen
Aberdeen, AB24 3UE, UK
Email: s.i.fraser@abdn.ac.uk

Alastair R. Allen
Department of Engineering
University of Aberdeen
Aberdeen, AB24 3UE, UK
Email: a.allen@abdn.ac.uk

ABSTRACT

Fair benchmarking tools for comparing watermarking systems are introduced. Three different watermarking systems are fairly compared using these tools. The parametric values for each watermarking system are carefully selected so that the resultant watermarked images all have equal measures of visual degradation. The Watson Metric, which is based upon the Human Visual System (HVS), is used to compute the visual degradation. The watermarked images are subjected to signal processing attacks. The robustness of each watermarking system is analysed using the fair benchmarking tools. The effect of adding Error Correcting Codes (ECC) to these watermarking systems is also studied. Using fair benchmarking tools, it is shown that, in certain cases, what appears to be an increase in robustness is in fact a decrease in robustness.

KEY WORDS

Watermarking Techniques, Digital Image Watermarking, Benchmarking.

1 Introduction

The need to provide copyright protection to digital media is of great importance in today's connected world. This may be attributed to the widespread use of the internet which allows users to download perfect copies of digital media onto their computers. In order to combat this problem, various watermarking systems have been developed which allow authors of digital media to protect their copyrighted works.

There are many issues to consider when inserting a watermark for copyright protection, for example: (1) type of host image, (2) length of watermark, (3) type of watermark (e.g., one-bit watermark or binary payload watermark), (4) embedding strength, (5) attacks that the watermark is likely to undergo. All of these issues will have a bearing upon the robustness of a watermarking system.

In [1, 2], methodologies for fairly comparing different watermarking systems were outlined. It was suggested that the following attributes be measured: (1) visual degradation caused by the watermark, (2) robustness of the watermarking system to attack and (3) the reliability of system. In a watermarking system, there is a trade off between robustness and watermark perceptibility. Thus, for fair evaluation and comparison of different systems, the parameters of each system need to be set so that the resultant marked images have equal amounts of visual degradation. In this paper, the visual degradation is measured using the Watson Metric [3, 4, 5]. This metric is based upon the Human Visual System (HVS); thus, it is better than the pixel based Peak Signal to Noise Ratio (PSNR) metric. The robustness of the different watermarking systems is measured by computing the Normalised Correlation (NC) between the original and recovered watermarks. To ensure statistically valid results, many tests are run using different watermarks and different insertion keys. Performing these steps ensures a fair and unbiased comparison of the different watermarking systems.

The reliability of the watermarking systems under test was computed using Receiver Operating Characteristic (ROC) graphs. ROC graphs [6] are very useful in assessing the overall behaviour and reliability of a watermarking system. Fair detector thresholds were set using Probability of False Positive (P_{fp}) calculations [7]. By considering P_{fp} while computing detector thresholds, the chance of obtaining a *false positive* reading for watermark presence becomes statistically unlikely.

This work was extended by adding Error Correcting Codes (ECC) to the watermarking systems under investigation. Many authors [8, 9, 10, 11] have stipulated that various forms of ECCs have been incorporated into their watermarking systems in order to increase robustness. However, many of these authors do not compare the performance of watermarks with ECCs against watermarks without ECCs. This may lead one to believe that adding ECC to a watermarking system will automatically result in better performance. To this end, Bose, Chaudhuri and Hocquenghem (BCH) codes [12] were added to the watermarking systems under investigation and their effect studied.

2 Methodology

A fair comparison of three different watermarking systems was carried out. These comprised the spatially based Bruyndonckx *et al.* [13] watermarking system, the Discrete Cosine Transform (DCT) based Koch *et al.* [14] system, and the Discrete Wavelet Transform (DWT) based Xie *et al.* [15] system. All three of these systems are blind, each embed a binary watermark within the host image and it is the goal of all three to provide copyright protection.

The Bruyndonckx algorithm works in the spatial domain and selects non-overlapping blocks of 8×8. Elementary perceptual calculations are performed in these blocks by classifying pixels into zones of homogeneous luminance. One bit of a binary watermark $\in \{0, 1\}$ is embedded in the relationship between mean values in these zones of homogeneous luminance.

The Koch algorithm takes the DCT of non-overlapping 8×8 blocks. Two random DCT coefficients from the middle frequency range are selected. The relationship between these two coefficients is altered to encode a binary watermark $\in \{0, 1\}$.

The Xie algorithm runs a non-overlapping 1×3 window across the low frequency sub-band of the DWT. The coefficients in these windows are sorted and the median value is quantised to represent a watermark bit $\in \{0, 1\}$.

2.1 The Watson metric

The three watermarking systems were fairly compared by setting their parametric values so that the marked images that they produced were of equal visual degradation. This visual degradation was quantitatively measured using the Total Perceptual Error (TPE) measurement calculated from the Watson Metric [3]. The Watson Metric is based upon the Human Visual System (HVS). This makes it a better quality metric than the commonly used Peak Signal to Noise Ratio (PSNR) metric which merely measures the pixel based difference between the original and the watermarked image (it does not take the HVS into consideration). The Watson metric weights the errors for each DCT coefficient in each block by its corresponding sensitivity threshold (which is a function of contrast sensitivity, luminance masking and contrast masking). The higher the TPE value, the more the image has been degraded. A more detailed description of the Watson Metric can be found in [4, 5].

2.2 BCH codes

The following BCH [12] coding strategies were used: BCH(80,52,9), BCH(80,38,13), BCH(80,24,19), BCH(320,203,27), BCH(320,113,51), BCH(320,24,79) as well as the uncoded case. The BCH codes are in the form of BCH(*n*,*k*,*d*), where *n* is the total length (*i.e.*, the watermark), *k* is the length of the message (that the user wishes to embed) and *d* is the minimum distance. A BCH code can correct up to *t* errors, where $d = 2t + 1$.

2.3 Signal processing attack

Watermarking systems which attempt to provide copyright protection must be resilient to different forms of attack. Such attacks include, amongst others, compression, noise addition, filtering, cropping, resizing and rotation. These attacks can be grouped into two categories: (1) signal processing attacks and (2) geometrical attacks. Signal processing attacks (*e.g.* noise addition, compression and filtering) reduce watermark energy within an image. After such an attack, a decoder can locate the pixels that have been marked but it cannot necessarily detect the watermark correctly (due to the low energy of the watermark). Geometrical attacks (*e.g.* cropping, resizing and rotation) attempt to desynchronize the watermark at the decoder. When desynchronization occurs, the decoder cannot find the pixels that have been marked and thus cannot detect the watermark. This paper focuses upon signal processing attacks only. A more detailed discussion of geometrical attacks and resynchronization techniques can be found in [16, 17, 18, 19]. The signal processing attack of JPEG compression is used in this paper.

2.4 Measuring robustness

Each JPEG quality attack level was performed fifty times upon each host image. This was repeated for all three watermarking systems under test. For each JPEG quality attack level, a different seed value and a different binary watermark $\in \{0, 1\}$ were embedded into the test image. For all fifty attacks, the original and recovered messages were compared by computing the Normalised Correlation (NC):

$$\text{NC} = \frac{m^* \cdot m}{||m^*|| \cdot ||m||} \tag{1}$$

where m is the original message and m^* is the recovered message (convert unipolar vectors, $m \in \{0, 1\}$, to bipolar vectors, $m \in \{-1, 1\}$, in this equation). These fifty NC values were averaged, resulting in a single NC value for a particular JPEG quality level in a particular image for a specific watermarking system. A graph of NC against

JPEG quality was plotted for each watermarking system and each host image.

2.5 Probability of false positives

Different detector thresholds are calculated for different message lengths that result in Probability of False Positive (P_{fp}) values of similar magnitude. These detector threshold values are then used to fairly analyse the NC against JPEG quality graphs for different message lengths. As described in [7], the P_{fp} for a binary vector can be calculated via:

$$P_{fp} = \sum_{n=\lceil N_w(T+1)/2 \rceil}^{N_w} \left(\begin{array}{c} N_w \\ n \end{array} \right) 0.5^{N_w} \qquad (2)$$

where N_w is the message length, T is the chosen detector threshold and:

$$\left(\begin{array}{c} N_w \\ n \end{array} \right) = \frac{N_w!}{n!(N_w - n)!}. \qquad (3)$$

2.6 ROC graphs

Comparing different watermarking systems using a fixed threshold for the normalised correlation measurement could give misleading results (especially if the visual degradations had not been set equal nor P_{fp} calculations considered) [1]. Different watermarking systems have very different *modis operandi* and as such it is not prudent to set the same threshold value for every watermark detector. A more sensible approach is to treat each watermarking system as a separate entity and calculate an individual threshold value for each detector. ROC graphs [6] address this problem by continuously varying the threshold over a wide range of values. This has the benefit of giving a fair comparison of watermarking systems that have very different embedding/detection methods. The performance of the detector can be quantified by calculating the area[1] beneath the ROC curve. For an ideal detector, the area would be 1.00. For a worthless detector (a detector that randomly selects if an image is watermarked or not), the area would be 0.50.

3 Results: 80 bit watermarks

Results for 80 bit watermarks are presented. The 80 bit watermarks consist of the following: (1) uncoded 80 bit message, (2) BCH(80,52,9): message of 52 bits with 28 ECC bits, can correct 4 errors, (3) BCH(80,38,13): message of 38 bits with 42 ECC bits, can correct 6 errors, (4) BCH(80,24,19): message of 24 bits, 56 ECC bits, can correct 9 errors. Three different greyscale images, each 256×256, are used as hosts for the watermarks. These are the Pentagon image, the Fishingboat image and the Lena image (each widely used in image processing literature). The watermarked images are attacked with JPEG compression (a signal processing attack, see Section 2.3). Graphs of NC against JPEG quality are plotted for each watermarking system and each host image. The parameters for all three watermarking systems are chosen so that the resultant TPEs, measured via the Watson Metric (block size of 8×8), are all 0.002. This value indicates that the watermarks are imperceptible to a human viewer.

Table 1 shows the embedding strengths used for each algorithm and the resulting visual degradations. As was suggested in [2], it is safe to assume that a binary watermark is detected if 80% of the bits match the original. This is equivalent to a NC value of 0.60. For a binary vector of length 80 bits, the probability of randomly obtaining a false positive reading when using a detector threshold value of 0.60 is is 2.5×10^{-8} (equation 2). Thus, for message lengths of 52, 38 and 24, detector thresholds are chosen that result in P_{fp} values of similar magnitude. These detector threshold values and their corresponding P_{fp} values are shown in Table 2. In Figure 1, these different detector threshold values for different message lengths are repre-

Algorithm	Image	PSNR (dB)	TPE	Embedding strength
Bruyndonckx	Lena	46.75	0.002	7.50
	Fishingboat	46.18	0.002	7.50
	Pentagon	46.83	0.002	7.50
Koch	Lena	43.59	0.002	5.00
(JPEG quality	Fishingboat	43.05	0.002	7.50
setting of 90)	Pentagon	42.12	0.002	7.50
Xie	Lena	48.81	0.002	0.10
(4 wavelet levels)	Fishingboat	47.29	0.002	0.18
	Pentagon	50.55	0.002	0.25

Table 1. Visual quality of images with 80 bit watermarks

Algorithm	Coding strategy	Detector threshold	P_{fp}
Bruyndonckx	uncoded	0.60	2.5×10^{-8}
	BCH(80,52,9)	0.75	3.5×10^{-8}
	BCH(80,38,13)	0.85	3.3×10^{-8}
	BCH(80,24,19)	1.00	6.0×10^{-8}
Koch	uncoded	0.60	2.5×10^{-8}
	BCH(80,52,9)	0.75	3.5×10^{-8}
	BCH(80,38,13)	0.85	3.3×10^{-8}
	BCH(80,24,19)	1.00	6.0×10^{-8}
Xie	uncoded	0.60	2.5×10^{-8}
	BCH(80,52,9)	0.75	3.5×10^{-8}
	BCH(80,38,13)	0.85	3.3×10^{-8}
	BCH(80,24,19)	1.00	6.0×10^{-8}

Table 2. Detector thresholds for different message lengths

[1] done using the MATLAB *trapz* function

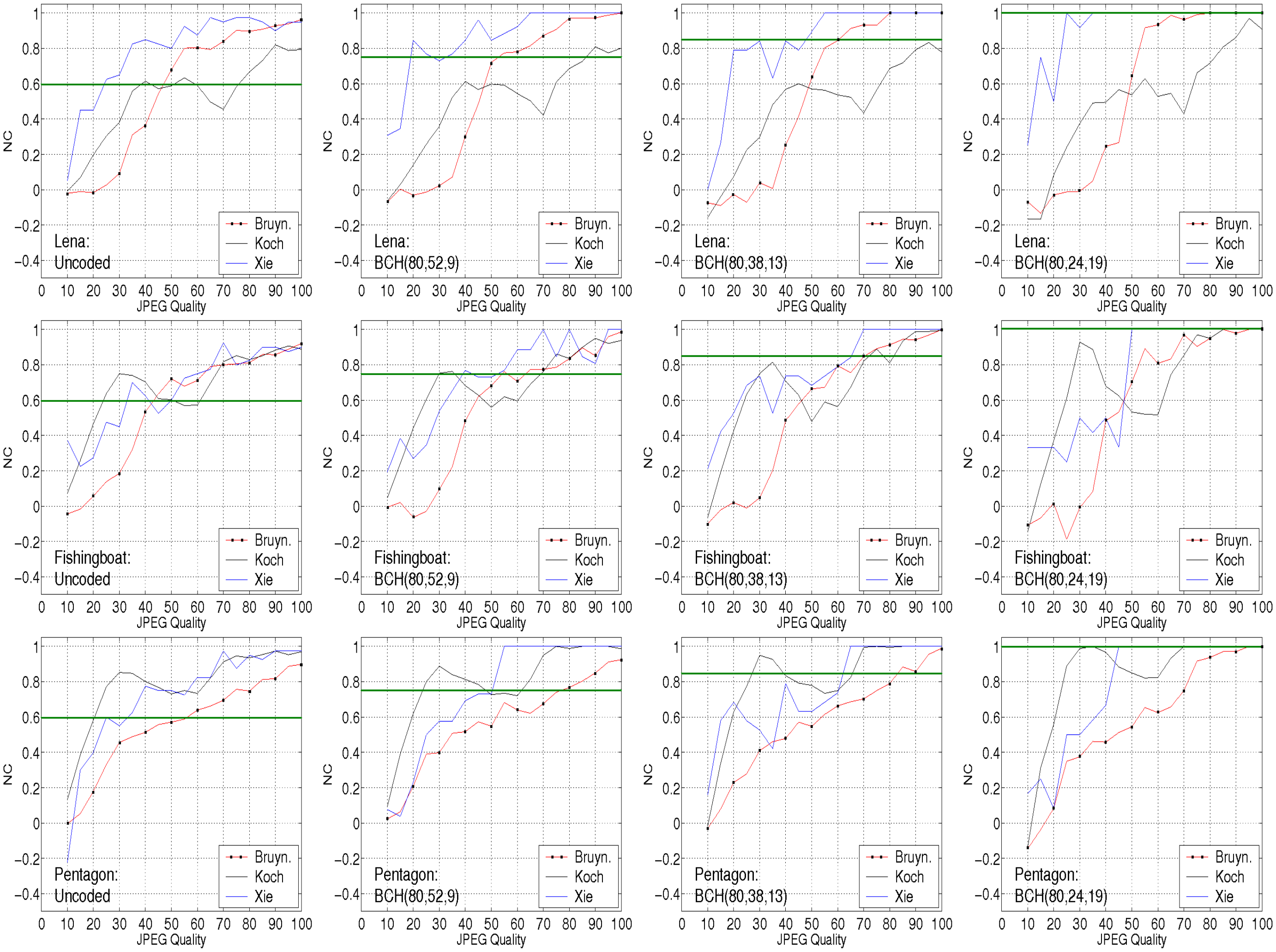

Figure 1. Results for different watermarking systems using 80 bit watermarks

sented by the "thick" horizontal lines at NC values of 0.60, 0.75, 0.85 and 1.00. Looking at the uncoded Lena graph in Figure 1 in conjunction with the detector thresholds in Table 2, it can be seen that the Xie algorithm is the most robust. For all JPEG quality values 25 and greater, the Xie algorithm is returning NC values greater than 0.60. The next most robust algorithm is the Bruyndonckx followed by the Koch algorithm. The Bruyndonckx algorithm is returning NC values greater than 0.60 for JPEG quality attacks of 50 and greater. The Koch algorithm is returning NC values equal to or greater than 0.60 for JPEG quality attacks of 75 and greater. Looking at the Lena:BCH(80,24,19) graph in Figure 1 (top right), the results are very different. Here, the Xie algorithm reaches the target NC of 1.00 at JPEG quality 35, the Bruyndonckx reaches NC of 1.00 at JPEG quality 80 and the Koch algorithm never reaches a NC value of 1.00. The two graphs between these two extremes, Lena:BCH(80,52,9) and Lena:BCH(80,38,13), follow a pattern of the Xie algorithm being the most robust, then the Bruyndonckx algorithm and then the Koch algorithm. It can be seen that the uncoded watermarks are performing the best; as more BCH coding is added, the results deteriorate.

The Bit Error Rate (BER) of this channel (watermarked image) is too high to reap the benefits of BCH coding (with the watermark embedding strengths used in Table 1). The BER of the channel could be reduced by increasing the length of the watermark or embedding the watermark with greater strength. However, both of these options would reduce the quality of the watermarked image. It should be noted that as the level of BCH coding increases in the Lena image, the NC values are increasing. However, because fair benchmarking tools have been used, it is clear that this is an *apparent* increase in robustenss only, the uncoded watermarks are performing better (although returning lower NC values).

Similar patterns of results are obtained for the Fishingboat and Pentagon images; as the level of BCH coding increases, the performance of a particular watermarking system decreases. The image specific nature of watermarking systems can be seen here too. In the Pentagon image, the Koch algorithm performs far stronger than it did in the Lena image. The Picture Information Measure (PIM) [20] can be used to ascertain the complexity of an image (the higher the PIM value of an image, the more inhomogeneous it is). A PIM value is calculated for each 8×8 sub-block via:

$$PIM = \left(\sum_{i=0}^{L-1} h(i) \right) - \max_i [h(i)] \tag{4}$$

where *L* is the number of grey levels in a sub-block and *h(i)* is a histogram for grey level *i* in a sub-block. The following PIM results were obtained: Lena (18848), Fishingboat (21513) and Pentagon (28442). It can be seen that Lena is the smoothest image, then Fishingboat and Pentagon is the busiest image. From this, it can be ascertained that the Xie and Bruyndonckx algorithms perform best in smooth images whereas the Koch algorithm performs best in busy images. Table 3 outlines the reliability of the watermarking systems via an analysis of the area under the ROC curves. All systems can be said to be reliable as the areas under the curves are much closer to 1.00 than they are to 0.50. Overall, it can be seen that the reliability of the transform based watermarking systems, Koch and Xie, are better (higher) than that of the spatially based Bruyndonckx system.

4 Results: 320 bit watermarks

The same three watermarking systems are tested again, however, watermarks of 320 bits are used rather than 80 bits. The 320 bit watermarks consist of the following: (1) uncoded 320 bit message, (2) BCH(320,203,27): message of 203 bits with 117 ECC bits, can correct 13 errors, (3) BCH(320,113,51): message of 113 bits with 207 ECC bits, can correct 25 errors, (4) BCH(320,24,79): message of 24 bits, 296 ECC bits, can correct 39 errors. Only the Lena image is used to embed the watermarks in these tests. The parameters for the watermarking systems were set so that the TPE of all the watermarked images was 0.006. For this level of TPE, it is possible for a human viewer to note a difference when a watermarked image is closely compared with an original image. Thus, the 320 bit watermarks have been inserted stronger than the 80 bit watermarks. Table 4 shows the parametric values used in all the systems to produce watermarked images of equal degradation. Table 5 shows the detector thresholds chosen for each coding strategy that result in P_{fp} values of similar magnitude. From the graphs in Figure 2, it can be seen that the Xie algorithm is again the most robust (in the Lena image). In the **Bruyndonckx system**, the BCH decoded messages only improve on the uncoded message NC values at high JPEG qualities (low compression rates). At low JPEG qualities, the performance of the BCH decoded messages is worse than the uncoded messages. In the **Koch system**, the BCH decoded messages do not improve upon the uncoded message at all; the more BCH coding, the worse the performance. In the **Xie system**, it can be seen that BCH(320,203,27) and BCH(320,113,51) can survive strong JPEG compression attacks (similar to the uncoded case) yet return slightly higher NC values than the uncoded case. In particular, BCH(320,113,51) return perfect NC values (1.00) for all attacks down to JPEG quality 65. BCH(320,29,79) makes the Xie algorithm *less* robust to strong JPEG attacks, but when an attack is survived, it returns perfect NC values of 1.00. If different detector threshold values had not been used in this analysis, a developer may wrongly conclude that adding any of these BCH codes to the Xie system would improve its robustness.

Algorithm	Image	ROC Area			
		Uncoded	BCH (80,52,9)	BCH (80,38,13)	BCH (80,24,19)
Bruyn- donckx	Pent.	0.938	0.946	0.923	0.890
	Fish.	0.883	0.845	0.832	0.811
	Lena	0.862	0.832	0.780	0.804
Koch	Pent.	0.995	0.991	0.977	0.962
	Fish.	0.990	0.982	0.973	0.950
	Lena	0.968	0.952	0.931	0.916
Xie	Pent.	0.933	0.913	0.893	0.927
	Fish.	0.949	0.915	0.967	0.929
	Lena	0.908	0.967	0.973	0.958

Table 3. Area under ROC graphs for 80 bit watermarks

Algorithm	TPE	Embedding strength	Block size	JPEG setting	Wavelet levels
Bruyndonckx	0.006	7	8×8	-	-
Koch	0.006	5	8×8	90	-
Xie	0.006	0.3	1×3	-	3

Table 4. Visual quality of Lena with 320 bit watermarks

Algorithm	Coding strategy	Detector threshold	P_{fp}
Bruyndonckx	uncoded	0.40	$< 2.3 \times 10^{-7}$
	BCH(320,203,27)	0.40	$< 2.3 \times 10^{-7}$
	BCH(320,113,51)	0.50	1.1×10^{-7}
	BCH(320,29,79)	0.90	8.1×10^{-7}
Koch	uncoded	0.40	$< 2.3 \times 10^{-7}$
	BCH(320,203,27)	0.40	$< 2.3 \times 10^{-7}$
	BCH(320,113,51)	0.50	1.1×10^{-7}
	BCH(320,29,79)	0.90	8.1×10^{-7}
Xie	uncoded	0.40	$< 2.3 \times 10^{-7}$
	BCH(320,52,9)	0.40	$< 2.3 \times 10^{-7}$
	BCH(320,38,13)	0.50	1.1×10^{-7}
	BCH(320,24,79)	0.90	8.1×10^{-7}

Table 5. Detector thresholds for different message lengths

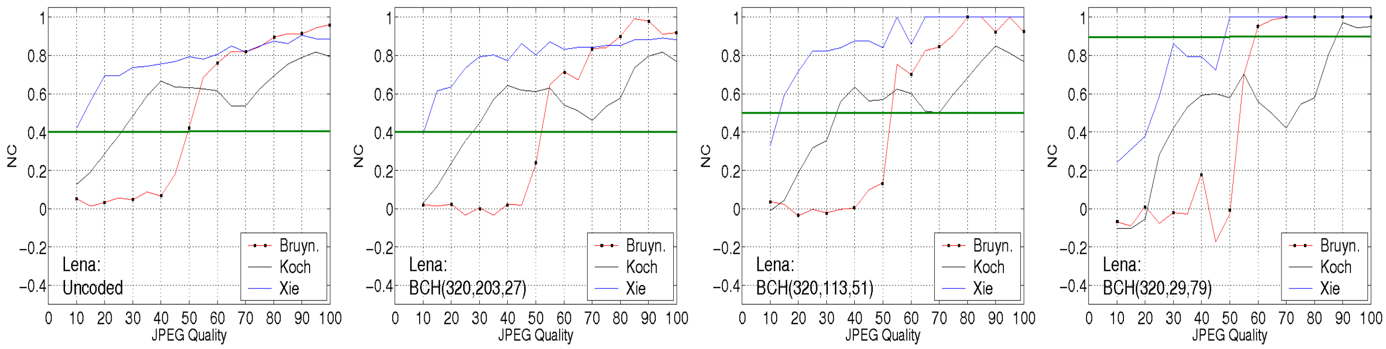

Figure 2. Results for different watermarking systems using 320 bit watermarks

Coding strategy	Image	Bruyndonckx	Koch	Xie
Uncoded	Lena	0.841	0.971	0.991
BCH(320,203,27)	Lena	0.748	0.925	0.960
BCH(320,113,51)	Lena	0.726	0.897	0.999
BCH(320,29,79)	Lena	0.717	0.864	0.909

Table 6. Area under ROC graphs for 320 bit watermarks

Analysis of the ROC values in Table 6 show that all three of the systems are reliable. In general, as more BCH code is added to any of these systems, the reliability of that system decreases. However, as reported in Table 3, the Xie and Koch systems are more reliable (in the Lena image) than the Bruyndonckx system.

5 Conclusion

In conclusion, benchmarking tools (*e.g.* NC against attack strength, Watson metric, detector thresholds based on fair P_{fp} values, ROC graphs and PIM) have been used to fairly compare different watermarking systems (with and without ECCs). Such tools (or their equivalent) should be used by designers to compare the robustness of novel watermarking algorithms with existing algorithms.

References

[1] S. Katzenbeisser and F. A. P. Petitcolas. *Information hiding: Techniques for steganography and digital watermarking*. Artech House Books, 1999.

[2] M. Kutter and F. A. P. Petitcolas. A fair benchmark for image watermarking systems. In *Security and Watermarking of Multimedia Contents*, volume 3657, pages 226–239. SPIE, 1999.

[3] A. B. Watson. DCT quantization matrices visually optimized for individual images. In *Human Vision, Visual Processing and Digital Display IV*, volume 1913, pages 202–206. SPIE, 1993.

[4] A. Mayache, T. Eude, and H. Cherefi. A comparison of image quality models and metrics based on human visual sensitivity. In *IEEE Intl. Conf. Image Processing, ICIP'98*, pages 409–413, Chicago, IL, USA, October 1998.

[5] S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun. Attack modelling: Towards a second generation watermarking benchmark. *Elsevier: Signal Processing*, 80(6):1177–1214, May 2001.

[6] A. R. van Erkel. Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. *European Journal of Radiology*, 27:88–94, 1998.

[7] D. Kundur and D. Hatzinakos. Digital watermarking using multiresolution wavelet decomposition. In *IEEE Intl. Conf. Acoustics, Speech and Signal Processing, ICASSP'98*, volume 5, pages 2659–2662, Seattle, WA, USA, May 1998.

[8] L. M. Marvel, C. G. Bonclet Jr., and C. T. Retter. Spread spectrum image steganography. *IEEE Trans. Image Processing*, 8(8):1075–1083, August 1999.

[9] J. Z. Wang and G. Wiederhold. WaveMark: Digital image watermarking using Daubechies' wavelets and error correcting coding. In *Multimedia Systems and Applications*, volume 3528, pages 432–439. SPIE, 1998.

[10] A. Herrigel, J. K. Ó Ruanaidh, H. Petersen, S. Pereira, and T. Pun. Secure copyright protection techniques for digital images. In D. Aucsmith, editor, *Information Hiding*, volume 1525 of *Lecture Notes in Computer Science (LNCS)*, pages 169–190. Springer-Verlag (Berlin), 1998.

[11] J. J. Eggers, J. K. Su, and B. Girod. Robustness of a blind image watermarking scheme. In *IEEE Intl. Conf. Image Processing, ICIP'2000*, pages 17–20, Vancouver, Canada, September 2000.

[12] V. Pless. *Introduction to the theory of error-correcting codes*. Wiley (New York), 1989.

[13] O. Bruyndonckx, J. J. Quisquater, and B. Macq. Spatial method for copyright labeling of digital images. In *IEEE Workshop on Nonlinear Signal and Image Processing*, pages 456–459, Neos Marmaras, Greece, 1995.

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

[15] L. Xie and G. R. Arce. Joint wavelet compression and authentication watermarking. In *IEEE Intl. Conf. Image Processing, ICIP'98*, pages 427–431, Chicago, IL, USA, October 1998.

[16] J. K. Ó Ruanaidh and T. Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Processing*, 66(3):303–317, May 1998.

[17] C-Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, M. L. Miller, and Y. M. Lui. Rotation, scale and translation resilient watermarking for images. *IEEE Trans. on Image Processing*, 10(5):767–782, May 2001.

[18] P. Bas and J-M. Chassery. Robust watermarking based on the warping of pre-defined triangular patterns. In *Proc. of SPIE, Electronics Imaging 2000, Security and Watermarking of Multimedia Contents II*, pages 99–109, San Jose, CA, USA, 2000.

[19] S. Voloshynovskiy, F. Deguillaume, and T. Pun. Multibit digital watermarking robust against nonlinear geometrical distortions. In *Proc. IEEE Intl. Conf. on Image Processing, ICIP'2001*, pages 999–1002, Thessaloniki, Greece, October 2001.

[20] J-W. Shin and D-S. Jeong. A new watermarking method using entropy-based region segmentation. In *Proc. of SPIE*, volume 3528, pages 531–538, 1999.