Robust and Semi-fragile Watermarking Techniques for Image Content Protection

by
Xi Zhao

Supervisor: Prof. Anthony TS Ho
Co-Supervisor: Prof. Steve Schneider

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
Faculty of Engineering and Physical Sciences
Department of Computing

July 2011
“Learning is like the horizon; there is no limit.”

My favourite Chinese proverb
Abstract

With the tremendous growth and usage of digital images nowadays, the integrity and authenticity of digital content is becoming increasingly important and of major concern to many government and commercial sectors. In the past decade or so, digital watermarking has attracted much attention and offers some real solutions in protecting the copyright and authenticating the digital images. Recently, image forensics techniques, based on a passive statistical analysis of the image data only, is an alternative approach to the active embedding of data associated with digital watermarking. In this thesis, the concept of digital watermarking is described, and a number of algorithms used for robust, fragile and semi-fragile image watermarking are reviewed in detail. We discuss the concept of image forensics briefly and conclude image forensics techniques into two main areas: image source detection and image manipulation detection. Four novel robust and semi-fragile transform based image watermarking related schemes are introduced. These include wavelet-based contourlet transform (WBCT) for both robust and semi-fragile watermarking, slant transform (SLT) for semi-fragile watermarking as well as applying the generalised Benford's Law to estimate JPEG compression, then adjust the appropriate threshold for improving the semi-fragile watermarking technique.

In this thesis, the proposed WBCT for robust watermarking is evaluated and compared with two other Discrete Wavelet Transform (DWT) based algorithms with results achieving high degree of robustness against most non-geometrical and geometrical attacks, while maintaining an excellent perceptual quality. For semi-fragile watermarking, the proposed SLT as a block-based algorithm achieves more accuracy for copy & paste attacks with non-malicious manipulations, such as additive Gaussian noise when compared with existing Discrete Cosine Transform (DCT) based and Pinned Sine Transform based schemes. While for the proposed WBCT method, good performance is achieved in localising the tampered regions, even when the image has been subjected to non-malicious manipulations such as JPEG/JPEG2000 compressions, Gaussian noise, Gaussian filtering, and contrast stretching. The average false negative rate is found to be approximately 1% while maintaining an average false positive rate below 6.5%. We also propose the use of generalised Benford's Law model as an image forensics technique for semi-fragile watermarking. This model can improve the lower tampered detection rate caused by the predetermined threshold in semi-fragile watermarking. The threshold is typically fixed and cannot be easily adapted to different amounts of errors caused by unknown JPEG compression. Our proposed method can adaptively adjust the threshold for images based on the estimated Quality Factor (QF) by using the generalised Benford's Law with overall average QF correct detection rate of approximately 99% when
5% of the pixels are subjected to image content tampering, as well as compression using different QFs (ranging from 95 to 65).
Acknowledgements

I wish to express my warmest appreciation to a number of people who have been supportive and contributed many efforts to my work.

I would like to give whole-hearted thanks to my supervisor Professor Anthony TS Ho, who offered me plenty helpful guidance on my research. When I was doing my Master degree, I was attracted by his interesting lecture. Without the input from him, I would not have known which directions I should go and which areas I would be interested, and a change in my career. He is one of few people, who have been made the biggest influence in my life.

A special thank to my co-supervisor Professor Steve Schneider, who kindly provided many supports. I would also thank to Professor Yun Q. Shi from New Jersey Institute of Technology Newark, USA, Dr Chang-Tsun Li from University of Warwick, UK and Professor Yaochu Jin from University of Surrey, UK who have given me many relevant suggestions for my research.

I would also like to express my gratitude to all members in the Watermarking and Multimedia Security group for the advice and support that they have given me.

Finally, I would also like to thank my wife, parents and parents-in-law for the supports and understanding they provided.
Contents

Abstract ii

Acknowledgements iv

List of Figures vii

List of Tables x

1 Introduction 1
  1.1 Background and Motivation .................................................. 1
  1.2 Aims and Objectives .............................................................. 4
  1.3 Contributions ........................................................................ 4
  1.4 Structure of the Thesis .......................................................... 5

2 Literature Review 7
  2.1 Digital Watermarking Overview ............................................. 8
    2.1.1 Requirements for Digital Image Watermarking ................. 9
  2.2 Robust, Fragile and Semi-fragile Image Watermarking .......... 13
    2.2.1 Robust Watermarking .................................................... 14
    2.2.2 Fragile Watermarking .................................................... 19
    2.2.3 Semi-fragile Watermarking .......................................... 21
  2.3 Image Forensics ................................................................. 24
    2.3.1 Classification of Image Forensics .................................... 24
    2.3.2 Benford’s Law for Image Forensics .................................. 28
  2.4 Summary ................................................................. 29

3 Novel Robust Watermarking Algorithm Based on WBCT 30
  3.1 Wavelet-based Contourlet Transform (WBCT) ....................... 31
  3.2 Coefficient Relations with JPEG Compression of Original Images 33
  3.3 Watermark Embedding Process ........................................... 33
  3.4 Watermark Detection Process ............................................. 35
  3.5 Experimental Results ............................................................ 36
    3.5.1 Imperceptibility ............................................................ 37
    3.5.2 Robustness ................................................................. 37
  3.6 Summary ................................................................. 42

4 Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT 45
4.1 Novel SLT Based Semi-fragile Watermarking Algorithm ..............................46
   4.1.1 The Slant Transform (SLT) .................................................................46
   4.1.2 SLT Watermark Embedding Process .................................................47
   4.1.3 SLT Watermark Detection, Authentication and Restoration ............48
   4.1.4 Experimental Results ........................................................................49
   4.1.5 Further Experimental Results .............................................................52
4.2 Novel WBCT Based Semi-fragile Watermarking Algorithm .......................56
   4.2.1 Parent and Children Coefficient Relations before and after Non-
       malicious Manipulations .......................................................................57
   4.2.2 Coefficient Differences Analysis .........................................................60
   4.2.3 WBCT Watermark Embedding Process ..............................................60
   4.2.4 WBCT Watermark Detection Process ................................................62
   4.2.5 WBCT Watermarked Image Authentication Process ...........................64
   4.2.6 Experimental Results ........................................................................65
4.3 Comparison Between SLT and WBCT Based Schemes ...............................70
4.4 Summary ..................................................................................................72

5 Proposed Image Forensics Technique for Semi-fragile Watermarking 75
5.1 Three Common Approaches Using Pre-determined Threshold in Semi-
    fragile Watermarking ..............................................................................76
5.2 Issue of Threshold in a Simple DCT Semi-fragile Watermarking Algorithm
    5.2.1 The DCT Watermark Embedding Process ........................................77
    5.2.2 The DCT Watermark Detection and Authentication Process ............78
    5.2.3 The Importance of the Threshold .......................................................79
5.3 Generalized Benford's Law vs. Watermarked Images ................................80
5.4 The Improved Authentication Method .......................................................85
5.5 Experimental Results (for a Simple DCT Semi-fragile Watermarking Scheme) 87
5.6 Improved Results (for Both SLT and WBCT Semi-fragile Watermarking
    Schemes) ..................................................................................................91
    5.6.1 Benford's Law for SLT-based Semi-fragile Watermarking ...............91
    5.6.2 Benford's Law for WBCT-based Semi-fragile Watermarking ............95
5.7 Summary ..................................................................................................96

6 Conclusion and Future work 98
6.1 Conclusion ..............................................................................................98
6.2 Future Work ............................................................................................100
## List of Figures

1.1 Four proposed image watermarking and forensics schemes ....................... 4  
2.1 A typical watermark embedding process. ..................................................... 9  
2.2 A typical blind watermark detection process ............................................. 9  
2.3 A typical non-blind watermark detection process. ..................................... 9  
2.4 Trade-off among the imperceptibility, robustness and capacity ................. 12  
2.5 An illustration of LSB watermark bits embedding ..................................... 13  
2.6 Difference between block and non-block based watermark embedding process. .......................................................... 14  
2.7 Schematic diagram for robust watermarking .............................................. 14  
2.8 Kay and Izquierdo watermark embedding process [72]. ............................ 15  
2.9 Kundur and Hatzinakos's DWT watermark embedding process [91]. ...... 17  
2.10 Three levels and one level of DWT decomposition .................................... 18  
2.11 Wang et al.'s watermark embedding process [99]. ................................. 19  
2.12 Fridrich's fragile watermark embedding process [105]. ............................ 20  
2.13 MSBs and LSB of pixel value 221 in 8 bits binary sequence. ..................... 20  
2.14 Fridrich's fragile watermark detection process [105]. ............................... 20  
2.15 Main functions of semi-fragile watermarking ....................................... 22  
2.16 Ho et al.'s PST-based watermark embedding process [13]. ...................... 23  
2.17 The Lincoln composite. .............................................................................. 26  
2.18 Comparison between Benford's Law and the 1st digits of DCT coefficients of 1338 images [134] ................................................................. 28  
3.1 The framework of the WBCT. ................................................................. 32  
3.2 (a) Original image (b) coefficient map after level-2 WBCT applied. ......... 32  
3.3 Parent-children relationship for DWT and WBCT. .................................... 32  
3.4 Six standard test images ............................................................................. 34  
3.5 Percentage of invariant energy relations after JPEG compression. ............ 35  
3.6 The proposed WBCT watermark embedding process. ............................... 35  
3.7 A demonstration of parent with their corresponded children coefficients in three level WBCT. ................................................................. 36  
3.8 The proposed WBCT watermark detection process. .................................. 36  
3.9 'Goldhill' - the original image and the watermarked image. 'San Diego' - the original image and the watermarked image. .......................... 38  
3.10 Performance of our method after JPEG2000 compression .................... 39  
3.11 Performance of our method after Gaussian white noise ......................... 39  
3.12 Performance of our method after Salt & Pepper noise ......................... 40  
3.13 Performance of our method after contrast stretching ............................. 40
List of Figures

3.14 Performance of our method after pixel circular shifting .................. 41
3.15 Performance of our method after scaling ..................................... 41
3.16 Attacked watermarked image ‘Goldhill’ - (a) JPEG Compression QF=20, (b) 400 pixel random shifted, (c) 20 degree of contrast stretching, (d) Gaussian white noise, variance=0.01 ..................... 44

4.1 Proposed SLT watermark embedding process ............................... 48
4.2 Our proposed SLT watermark detection, authentication and restoration process ................................................................. 49
4.3 Demonstration of the image Trucks in SLT semi-fragile watermarking scheme ................................................................. 51
4.4 Three law enforcement images .................................................. 53
4.5 Mid-frequency bands for embedding watermarks in SLT, PST and DCT based schemes ................................................................. 53
4.6 The watermarked, tampered, and authenticated images for image ‘Car1’  ................................................................. 54
4.7 The average percentages of invariant relations after JPEG compression . 58
4.8 The average percentages of invariant relations after JPEG2000 compression ................................................................. 59
4.9 Scatter plots showing the distributions of the coefficient differences before and after manipulations of image Goldhill (a) JPEG (QF=70) (b) JPEG2000 (QF=70) (c) Gaussian noise (σ=0.0003) (d) Gaussian filtering (e) Contrast stretching (f) copy and paste (64 x 128) ................... 61
4.10 WBCT semi-fragile watermark embedding process ............................. 62
4.11 WBCT semi-fragile watermark detection and image authentication process ................................................................. 63
4.12 The histogram of differences before and after JPEG compression (a) QF=90 (b) QF=70 (c) QF=50 ................................................................. 63
4.13 The histogram of differences before and after JPEG2000 compression (a) QF=90 (b) QF=70 (c) QF=50 ................................................................. 64
4.14 The histogram of differences before and after (a) Gaussian noise (b) Gaussian filtering (c) contrast stretch ................................................................. 64
4.15 Example of XOR fusion process ................................................................. 65
4.16 PSNR with different number of subband embedded randomly ..................... 66
4.17 P\text{FNR} with different number of subband embedded randomly ..................... 66
4.18 The watermarked, tampered, and authenticated images for image ‘Car2’ ................................................................. 67
4.19 Performance of false positive rate after JPEG compression ..................... 68
4.20 Performance of false positive rate after JPEG2000 compression ..................... 69
4.21 ‘Lena’ image (a) watermarked image (b) copy-and-pasted then JPEG compressed (QF=70) (c) authenticated image (d) authenticated image with morphological operations ................................................................. 72
4.22 ‘Goldhill’ image (a) watermarked image (b) four areas have been modified with JPEG compression (QF=70) (c) authenticated image (d) authenticated image with morphological operations ................................................................. 73
4.23 ‘Trucks’ image (a) watermarked image (b) copy-and-pasted image with Contrast stretching (c) authenticated image (d) authenticated image with morphological operations ................................................................. 74

5.1 The pre-determined threshold during the watermark embedding process. 77
5.2 An illustration of the DCT watermark embedding process. ..................... 78
5.3 An illustration of the DCT watermark detection and authentication processes. ................................................................. 79
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>The relationship among threshold $P_{FPR}$ and $P_{FNR}$</td>
<td>79</td>
</tr>
<tr>
<td>5.5</td>
<td>Different thresholds for $QF = 75$</td>
<td>80</td>
</tr>
<tr>
<td>5.6</td>
<td>Different thresholds for $QF = 95$</td>
<td>81</td>
</tr>
<tr>
<td>5.7</td>
<td>1st digit of block-DCT coefficients</td>
<td>82</td>
</tr>
<tr>
<td>5.8</td>
<td>1st digit of $8 \times 8$ Block-DCT coefficients</td>
<td>83</td>
</tr>
<tr>
<td>5.9</td>
<td>1st digit of JPEG coefficients ($QF = 100$)</td>
<td>83</td>
</tr>
<tr>
<td>5.10</td>
<td>1st digit of JPEG coefficients ($QF = 75$)</td>
<td>84</td>
</tr>
<tr>
<td>5.11</td>
<td>1st digit of JPEG coefficients ($QF = 50$)</td>
<td>84</td>
</tr>
<tr>
<td>5.12</td>
<td>Improved authentication process</td>
<td>86</td>
</tr>
<tr>
<td>5.13</td>
<td>Estimating the QF of a watermarked image</td>
<td>87</td>
</tr>
<tr>
<td>5.14</td>
<td>Ten types of test images with and without attacks</td>
<td>88</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Example of three parameters [134] .................................................. 29

3.1 PSNR comparison between our and Wang’ method ......................... 37

3.2 Performance of our method under different attacks ......................... 42

3.3 Comparison of our method’s performance with two other Methods ...... 43

4.1 Comparative performance in term of $P_{FPR}$ (%) the watermarking schemes against JPEG compression with varying quality factor .......... 50

4.2 Comparative performance in term of $P_{avg}$ (%) the watermarking schemes against copy and paste attack ...................................... 51

4.3 Comparative performance in term of $P_{avg}$ (%) the watermarking schemes against copy and paste attack (20% tampering) followed by JPEG compression with varying quality factor ........................................ 52

4.4 New comparative performance of the watermarking schemes against copy and paste attack (20%), detection rate (%) ..................... 54

4.5 New comparative performance of the watermarking schemes against copy and paste attack (20%) and JPEG compression (QF=85) ........ 55

4.6 New comparative performance of the watermarking schemes against copy and paste attack (20%) and JPEG compression (QF=75) ........ 56

4.7 New comparative performance of the watermarking schemes against copy and paste attack (20%) and Additive Gaussian noise (variance=0.003) .................................................. 56

4.8 New comparative performance of the watermarking schemes against copy and paste attack (20%) and Additive Gaussian noise (variance=0.005) .................................................. 57

4.9 The average percentages of invariant relations after three types manipulations .......................................................... 59

4.10 Performance of false positive rate after Gaussian noise, Gaussian filtering and contrast stretching ............................................. 68

4.11 Performance after copy and paste with JPEG compression ............ 70

4.12 Performance after copy and paste with JPEG000 compression ........ 70

4.13 Performance after copy and paste with three signal processing ........ 71

4.14 Comparison Between SLT and WBCT Semi-fragile ...................... 71

5.1 Mean standard deviations of 1338 images ..................................... 82

5.2 Mean standard deviations of 1338 JPEG compressed images ........... 85

5.3 Average $x^2$ of 1338 compressed watermarked images .................. 85

5.4 JPEG compression only ............................................................... 90

5.5 Combined (5%) copy and paste attack and JPEG compression ....... 90

5.6 Combined (20%) copy and paste attack and JPEG compression ....... 91

5.7 Combined (30%) copy and paste attack and JPEG compression ....... 91
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.8</td>
<td>Combined (20%) copy and paste attack and Median filtering (3 x 3) and JPEG compression</td>
<td>92</td>
</tr>
<tr>
<td>5.9</td>
<td>Combined (20%) copy and paste attack and Average filtering (3 x 3) and JPEG compression</td>
<td>92</td>
</tr>
<tr>
<td>5.10</td>
<td>Combined (20%) copy and paste attack and Gaussian lowpass filtering (3 x 3) and JPEG compression</td>
<td>93</td>
</tr>
<tr>
<td>5.11</td>
<td>Combined (20%) copy and paste attack and Histogram equalization and JPEG compression</td>
<td>93</td>
</tr>
<tr>
<td>5.12</td>
<td>The improved performance of the SLT semi-fragile watermarking scheme against copy and paste attack (20%) by using Benford’s Law</td>
<td>94</td>
</tr>
<tr>
<td>5.13</td>
<td>The improved performance of the SLT semi-fragile watermarking scheme against copy and paste attack (20%) and JPEG compression (QF=85) by using Benford’s Law</td>
<td>94</td>
</tr>
<tr>
<td>5.14</td>
<td>The improved performance of the WBCT semi-fragile watermarking scheme against copy and paste attack (20%) and JPEG compression (QF=90) by using Benford’s Law</td>
<td>96</td>
</tr>
<tr>
<td>5.15</td>
<td>The improved performance of the WBCT semi-fragile watermarking scheme against copy and paste attack (20%) and JPEG compression (QF=70) by using Benford’s Law</td>
<td>96</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background and Motivation

Nowadays, the rapid development of technologies has led to the significant increase of
digital information, particularly multimedia such as image, audio and video content.
Such technological advances have led to easier illegally share, distribution, and copy
of Intellectual Property (IP). Subsequently, the copyright infringement issue has been
identified as a "hot topic". According to a report from Oxford Economics in 2009,
the UK film industry loses 531 million pounds per annum, as a direct result of copy­
right theft. Examples of this can include recording films at the cinema, illegal sale or
purchase of copyrighted DVDs, household copying, file-sharing and downloading, and
streaming material from unauthorised web servers [1]. Furthermore, the music industry
is also affected. According to the US copyright industry group International Intellectual
Property Alliance, two billion music tracks were illegally downloaded in Spain in 2008,
compared to 2.2 million that were purchased legally [2]. Moreover, billions of digital im­
ages are widely available and can be accessed easily and quickly via almost any website
containing graphics, or image search engines.

The primary reason for the requirement of authenticating images stems from the in­
creasing amount of doctored images that are presented as accurate representations of
real-life events, but are later discovered to be faked. The history of manipulating images
goes back to almost as far as photography itself, and with the ease of use and availability
of image editing software, it has become ubiquitous in the digital age. Image authenti­
cation schemes attempt to restore trust in the image by accurately validating the data,
positively or negatively. Especially for law enforcement scenarios, images captured at
the scene, such as for crime scene investigation and traffic enforcement, can potentially
be used as evidence in the court of law. If an image presented in court as evidence from
a crime scene is to be effectively used by the jury, the integrity of the information must not be in question. The Crime Scene Investigators (CSIs) gathers forensic evidence, such as take the crime scene photographs. However, after the collection of evidence, there is no other way of examining the crime scene as a whole, apart from analysing the collected exhibits and photographs [3]. In order to maintain the integrity of the images, not only is it essential to verify that the photographic evidence remains unchanged and authentic, but any manipulated regions should also be localised to help identify which parts of the image cannot be trusted. With the tremendous growth and usage of digital cameras and video devices, the requirement to verify the digital content is paramount, especially if it is to be used as evidence in court [4].

An obvious requirement, therefore, is the development of solutions for copyright protection and image authentication for digital content. Cryptography and digital watermarking are two commonly used technologies. The traditional cryptography can, for example, be utilised for message authentication by generating and embedding a digital signature into a message, in an effort to prevent the sending of forged messages [5]. In addition, according to Friedman [6], digital signatures can be embedded into images by applying cryptography if the signature is meta-data. In all cases, the use of cryptography is constrained by the fact that it can be lost easily during the image format conversion process, which subsequently invalidates the authentication process. Moreover, such solutions do not prevent or track the content against illegitimate reproduction after it has been decrypted [7].

Digital watermarking has attracted much attention in the past decade or so, which is the process of embedding relevant information (such as a logo, fingerprint and serial number), into a media. This technique can be applied to different media types such as video, audio and image content. An example of visible digital watermark is the translucent logos that are often seen embedded at the corner of videos or images, in an attempt to prevent copyright infringement. However, these visible watermarks can be targeted and removed rather simply by cropping the media, or overwriting the logos. Subsequently, the field of digital watermarking is primarily focused on invisible watermarks, which are imperceptible and operate by tweaking the physical data of the media [7, 8]. The employing of digital watermarking technique is also recommended by the UK government agency, the Department for Culture, Media and Sport (DCMS), which expresses the following views: “To develop and adopt pre-competitive standards and unique identifiers, which are open and interoperable, to cover hardware and software for secure delivery of music, including encryption, watermarking and usage rules of music on-line” [9]. “The UK Film Council’s position is that an effective deterrence policy needs to be based on a blend of educative, technological and legislative interventions with the latter firmly enforced. This policy should be accompanied by detailed exploration by all parties of
the potential of watermarking and other technologies to facilitate identification of illegal activity" [10].

There are three different classifications associated with digital watermarking, depending on the applications: “robust”, “fragile” and “semi-fragile”. Robust watermarking is primarily designed to provide copyright protection and proof of ownership for digital images. The most important property of robust watermarking is its ability to tolerate certain signal processing operations that usually occur during the lifetime of a media object, as well as preventing any more deliberate attacks. Fragile and semi-fragile digital watermarking techniques are often utilised for image content authentication applications to verify or authenticate the integrity of the digital media content. Fragile watermarking schemes are designed to detect any possible manipulations that affect the watermarked image pixel values [11, 12]. In comparison, semi-fragile schemes make it possible to verify the content of the original image, as well as permit alterations caused by non-malicious (unintentional) modifications such as system processes [13–15]. Moreover, semi-fragile watermarking is more focused on detecting intentional attacks than validating the originality of the image [16, 17]. During the image transmission, the mild signal processing errors caused by signal reconstruction and storage, such as transmission noise or JPEG compression, are permissible. However, the image content tampering such as copy-and-paste attack will be identified as a malicious attack. Additionally, in the literature, a significant amount of research has been focused on the design of semi-fragile algorithms that could tolerate JPEG compression and other common non-malicious manipulations [18–24]. However, watermarked images could be compressed by unknown JPEG compression rates of various quality factors (QFs). As a result, in order to authenticate the images, these algorithms have to set a pre-determined threshold that could allow them to tolerate different QF values when extracting the watermarks. If the QF could be estimated, then appropriate thresholds could be adapted for each test image, before initialising the watermark extraction and authentication process. This adaptive threshold could decrease the false alarm and missed detection rates.

In contrast to authenticating the image using active watermarking technique, the field of image forensics has been considered as a passive approach, and attracted much attention recently [25–28]. The significant difference is that image forensic techniques seek to authenticate images based solely on the image data that are used for image statistical analysis. As such, no embedded information is loaded into an image, and so the security risks and robustness issues associated with a payload, are avoided. As described in [29, 30], image forensics could also identify anomalies that might exist due to non-malicious processing (such as a change in file format) or intentional, malicious modifications (such as cloning or creating composites). It could also identify the difference between natural and unnatural images. A natural image possesses its original
characteristics, such as shape, contrast and size. While this image would become unnatural if any of these characteristics were to be changed. Hence, there is a growing need to develop advanced image forensic techniques that could help to further analyse different natural and unnatural images, particularly, those have not been watermarked initially.

1.2 Aims and Objectives

The main aims and objectives of this research are to develop and analyse advanced image watermarking and forensic schemes. This thesis begins with the fundamental concept and applications of digital watermarking, followed by a detailed review of existing active watermarking techniques, such as robust, fragile and semi-fragile methods as well as passive image forensics techniques. As shown in Figure 1.1, we propose four novel robust and semi-fragile transform based image watermarking and forensics schemes in this thesis. These include Wavelet-based Contourlet Transform (WBCT) for both robust and semi-fragile watermarking, Slant Transform (SLT) for semi-fragile watermarking, and also incorporating the generalised Benford’s Law, to estimate the JPEG QF are introduced to adaptively determine the threshold for improving semi-fragile watermarking.

![Figure 1.1: Four proposed image watermarking and forensics schemes](image)

1.3 Contributions

During my research on digital watermarking and image forensics, two main contributions have been made. Firstly, the characteristic of WBCT coefficients, parent and children relationship, has been analysed and evaluated. We found the relationships between parents and their correlated four children coefficients maintain invariant if the image has been undergone some manipulations, such as JPEG/JPEG2000, pixel shifting, mean filtering, histogram equalisation, median filter, Gaussian filtering and sharpening. Therefore, we utilised this invariant property of WBCT for robust watermarking. Moreover, we further
analysed the parent and children relationship, and found that this invariant property could be destroyed if the image has been tampered, such as copy-and-paste attack. Therefore, we applied it into semi-fragile watermarking for image authentication by verity if any parent and children relationships has been destroyed. Secondly, we discussed the issue of "predetermined" threshold for current semi-fragile watermarking schemes. We adapted an image forensics method, the generalised Benford's Law to estimate JPEG compression rate of the test image. Hence, the threshold could be adaptively adjusted before initialise the authentication process that could reduce the error detection rates.

Three book chapters, one international journal and six international conference papers have been accepted and published, as listed in Appendix 1.

1.4 Structure of the Thesis

The rest of thesis is organised as follows:

- Chapter 2 describes the concept of digital watermarking in various forms, such as visible and invisible watermarking, blind and non-blind watermarking schemes. We also discuss classifications of digital watermarking techniques such as robust, fragile and semi-fragile watermarking extensively. Moreover, the field of image forensics is also discussed.

- Chapter 3 describes a novel robust watermarking algorithm using the WBCT that exploits the energy relations between "parent" and "children" coefficients. The concept and advantages of the proposed algorithm, and its embedding and detection processes are described in detail. The experimental results are analysed and evaluated by comparing it with two discrete wavelet transform (DWT) domain based algorithms.

- Chapter 4 describes two novel semi-fragile watermarking algorithms using the SLT as a block based method, and WBCT as a non-block based method. Both algorithm's watermark embedding, detection, and authentication processes and their results including a performance analysis of false positive and false negative rates are discussed in detail. The SLT method will be compared with two existing transforms, Discrete Cosine Transform (DCT) and Pinned Sine Transform (PST). The performance of the WBCT semi-fragile watermarking will be evaluated against various attacks. Moreover, we analyse and compare the differences between the two proposed semi-fragile watermarking schemes.
• Chapter 5 discusses the limitation of predetermined threshold in current semi-fragile watermarking schemes. The background of Benford's Law, generalised Benford's Law and their relationship with the watermarked image, JPEG compressed watermarked image are described. We propose a framework incorporating the generalised Benford's Law to detect unknown JPEG compression QFs in semi-fragile watermarked images to adjust the appropriate threshold. Furthermore, we apply generalised Benford's Law to the proposed SLT and WBCT based semi-fragile watermarking schemes to improve accuracy rates in authenticating and localising the tampered regions.

• Chapter 6 presents the conclusion of the thesis and presents some directions for future work of our robust, semi-fragile watermarking and image forensics research.
Chapter 2

Literature Review

In this chapter, we give a background review of digital watermarking and image forensic techniques, highlighting the concept and related proposed schemes. We briefly describe visible and invisible watermarking, blind and non-blind watermarking schemes, imperceptibility, robustness, capacity and security as four requirements for digital image watermarking, and the differences between spatial domain based and transform domain based watermarking schemes. We also discuss classifications of digital watermarking techniques such as robust, fragile and semi-fragile watermarking. The field of image forensics is then reviewed. This chapter will also provide an overview of four proposed digital watermarking and image forensic schemes, which will be discussed with more details in Chapters 3, 4 and 5.
2.1 Digital Watermarking Overview

The concept of watermarking has been used in many different forms and can be traced back to thousands of years ago. For instance, in the late 13th century in Italy, a thin, translucent layer was sewn with wire onto a paper mould to form a watermark [31]. Historically, postage stamps and currencies were commonly watermarked. Indeed, the currency watermark is still used today when printing banknotes. A digital watermark can be either visible or invisible. An example of digital visible watermark is the translucent logos that are often seen embedded at the corner of videos or images, in an attempt to prevent copyright infringement. However, these visible watermarks can be targeted and removed rather simply by cropping the media, or overwriting the logos. Subsequently, the field of digital watermarking is primarily focused on embedding invisible watermarks, which operate by tweaking the content of the media imperceptibly. As the watermark cannot be seen, there must exist a robustness property that ensures the watermark data survives if the image is altered. Typical applications of digital watermarking can include broadcast monitoring, owner identification, proof of ownership, transaction tracking, content authentication, copy control, device control, legacy enhancement and content description [7, 8].

Figure 2.1 illustrates a typical watermark embedding process. The watermarked work is produced by an embedding algorithm that is traditionally comprised of three inputs: the original work, the watermark and a key. A blind watermark detection process is shown in Figure 2.2. The watermark is extracted from the watermarked work by using a detection algorithm in conjunction with the same key that was originally used to embed the watermark. In contrast, Figure 2.3 illustrates a non-blind (or informed) watermark detection process that extracts the watermark. Here, the original work has to be provided as a reference source in order for the detection algorithm to function. Therefore, the selection of a blind or non-blind watermarking detection system typically depends on whether the original work is accessible or not [8].

The original work is the host signal which is employed into a digital media such as, video [32–34], audio [35, 36], image [37–39], halftone image [40–42], binary text [43, 44], 3D meshes [45–47], holography [48, 49], optical [50, 51] and network protocol [52, 53]. The watermark is a binary sequence of data produced from a logo image, fingerprint, serial number, owner’s name or ID, or indeed anything that could identify the ownership of the media. The key is used to increase the security of the procedure; it prevents the possibility of a hacker modifying or removing the watermark as this can only be achieved if the key is known.
2.1.1 Requirements for Digital Image Watermarking

In this subsection, four important properties for digital watermarking are discussed. These are imperceptibility, robustness, capacity and security.

- Imperceptibility
  The embedded watermark should be imperceptible from the watermarked work. The degradation from original work to watermarked work is permitted that maintaining image fidelity of the original work. Therefore, in order to evaluate the similarity between the original and watermarked image, objective and subjective evaluation methods will be needed. One of the standardised subjective methods for image and video is the Double Stimulus Impairment Scale (DSIS) [54]. The watermarked images are visually reviewed by observers who have no background knowledge of image processing and are not experienced assessors. The watermarked images are compared against the original image and scored with a scale
from one to five (1 = very annoying, 2 = annoying, 3 = slightly annoying, 4 = perceptible, but not annoying, 5 = imperceptible). To measure the quality of a watermarked image objectively, a well-known statistical metric called the Peak Signal-to-Noise Ratio (PSNR), in Equation 2.1, is commonly used to evaluate the quality of the watermarked image by comparing it with the statistics of the original image. The quality of the watermarked image can be considered as acceptable if the PSNR value is above 30dB [55].

$$PSNR = 10 \log_{10} \left( \frac{MAX^2}{\frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (img(i,j) - img.w(i,j))^2} \right)$$  \hspace{1cm} (2.1)$$

where \(m\) and \(n\) are the size of the image (e.g. 512 x 512), \(img\) is the original image, \(img.w\) is the watermarked image, and \(MAX\) is the maximum possible gray value of the image (e.g. 255 for a grayscale image).

Other alternative or supplementary methods of evaluating the similarity between the original image and the watermarked image are weighted PSNR (wPSNR) [56] and Mean Structural Similarity Index (MSSIM) [57]. The wPSNR is defined as follows:

$$wPSNR = 10 \log_{10} \left( \frac{MAX^2}{\sqrt{MSE \times NVF}} \right)^2$$  \hspace{1cm} (2.2)$$

where NVF is exploited as a Gaussian model to estimate the degree of texture in the image by providing a value between 0 and 1 (where 0 = highly textured and 1 = smooth) [58]. In contrast to PSNR and wPSNR, the MSSIM method, in Equation 2.3, separates the luminance, contrast and structure of the images for similarity measurement. This metric is based on the degradation of structural information and attempts to measure the attributes that reflect the key structure and objects of importance in an image. The MSSIM values exhibit much better consistency with the qualitative visual appearance [57].

$$MSSIM (X,Y) = \frac{1}{M} \sum_{j=1}^{M} SSIM (x_j, y_j)$$  \hspace{1cm} (2.3)$$

where \(SSIM (x,y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma\), the luminance comparison function is \(l(x,y)\), the contrast comparison function is \(c(x,y)\), the structure comparison function is \(s(x,y)\), \(\alpha > 0\), \(\beta > 0\) and \(\gamma > 0\) are parameters used to adjust the relative importance of the three components. \(X\) and \(Y\) are the original and the watermarked images, respectively, \(x_j\) and \(y_j\) are the image contents at the \(j^{th}\) local window, and \(M\) is the number of local windows of the image [57]. However, due to its popularity and simplicity for calculation, the PSNR technique is used as
an objective metric in this thesis for assessing the image quality between original and watermarked images.

• Robustness
Robustness is an important property for robust watermarking schemes. The watermark that is embedded into the image should be robust (to varying degrees according to the application) to tolerate different forms of attack or image processing operations, when the watermarked image is transmitted. These image manipulations or attacks can be categorised into non-geometrical and geometrical groups. Non-geometrical distortion is derived from lossy compression algorithms such as JPEG or JPEG2000, as well as noise addition, image filtering and contrast stretching, while geometrical distortion includes rotation, scaling, cropping, translation, and shifting pixels. These distortions are often implemented to simulate possible attacks to analyse the performance trade-off of the proposed algorithms by the researchers in the community [7]. Maintaining the robustness of the watermark is much more difficult and challenging when considering geometrical attacks. This is due to the fact that each individual pixel location of the watermarked image is likely to be shifted or translated. A possible approach is to find an invariant property of an image that can be used in the watermark embedding process to enhance the robustness against different attacks. More details on the use of invariant properties for watermarking are described in Chapters 3 and 4.

• Capacity
Capacity refers to the maximum amount of watermark bits that can be embedded into the original image. The number of watermark bits embedded into the image data can affect the overall perceptual quality of the image. Figure 2.4 illustrates the performance trade-offs concerned with watermarking; specifically, the imperceptibility of the watermarked image, the robustness of the watermark, and the capacity of the watermark data. Typically, if the quality of the watermarked image increases, the robustness and capacity of the watermark data would decrease. Similarly, if the robustness increases, the quality of the watermarked image would decrease because more watermark bits will be used. Finally, if the capacity of watermark data increases, the quality of the image and its robustness would decrease.

• Security
The approach to security in digital watermarking is mainly focused on malicious removal or modification of the watermark bits. The watermark security can be defined as “the inability by unauthorised users to have access to the raw watermarking channel” [59]. The watermarking systems could be compromised if an
attacker manages to obtain the secret key. In this case, the attacker will have access to parameters such as the watermark embedding locations, random frequency of the watermark bits, and the threshold for embedding the watermark bits. The secret key can be predicted by gathering the characteristics of a set of watermarked images and analysing their similarities, to evaluate whether the same secret key and watermark bits have been used repeatedly [60]. Some of the problems associated with secret key leakage have been studied by a number of researchers [61–63].

Spatial Domain Based Watermarking Schemes
Digital image watermarking techniques can be divided into two categories: spatial and transform domain. In the spatial domain, the watermark bits are embedded into the original image by modifying the image image intensity directly. These types of techniques are easy to implement and have higher embedding capacity than transform domain based watermarking techniques. However, the watermarks embedded with spatial domain based watermarking techniques are not robust against various manipulations, such as JPEG compression, additive noise, filtering and geometric distortions.

The least significant bit (LSB) watermarking is one of the first watermarking schemes operating in the spatial domain [64]. As seen from Figure 2.5, the image intensity 135, 98 and 221 are first converted into 8 bits binary sequences. The three watermark bits 0, 1 and 1 are then inserted by replacing the LSB of each binary sequence. Finally, the modified binary sequences are converted back to 134, 99 and 221, respectively. In essence, the original intensity can only be changed either by ±1 pixel value, or not at all (if the LSB already matches the watermark bit) when they are watermarked using this technique. As the changes are relatively minor, the quality of the watermarked image is high, especially if limited amount of changes occur. However, the watermark bits can be easily attacked by randomly overwriting the LSB values of the watermarked image, thereby destroying the watermark data.
Chapter 2. Literature Review

Image intensity

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>135</td>
<td>98</td>
<td>221</td>
</tr>
</tbody>
</table>

Image intensity in binary

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10000111</td>
<td>01100010</td>
<td>11011101</td>
</tr>
</tbody>
</table>

Watermark bits

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Watermarked image intensity in binary

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10000110</td>
<td>01100011</td>
<td>11011101</td>
</tr>
</tbody>
</table>

Watermarked image intensity

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>134</td>
<td>99</td>
<td>221</td>
</tr>
</tbody>
</table>

Figure 2.5: An illustration of LSB watermark bits embedding.

Transform Domain Based Watermarking Schemes

In contrast, transform domain schemes operate by embedding the watermark data into frequency coefficients of the image. Transform domain techniques have been comprehensively studied in the context of image coding and compression as well as digital watermarking [8, 65]. An image can be represented as frequency coefficients by mapping the image intensity. Low frequencies represent the overall shapes and outlines of features in the image, and its luminance and contrast characteristics, and high frequencies represent sharp edges and crispiness in the image, but contribute little spatial-frequency energy [66]. As some of the coefficients are not significantly distorted after some attacks (such as JPEG compression, additive noise, and filtering), robust watermarking schemes are commonly implemented in the transform domain [67–69]. One of first robust watermarking scheme based on spread spectrum was first proposed by Cox et al. [37]. In general, robust watermarking in transform domain can be classified into two groups: block based and non-block based algorithms. Figure 2.6 shows the difference between block and non-block based in watermark embedding process.

2.2 Robust, Fragile and Semi-fragile Image Watermarking

As mentioned in Chapter 1, there are three different classifications associated with digital watermarking, depending on the applications: robust, fragile and semi-fragile. Robust watermarking has been used extensively in the past decade, and is primarily designed to provide copyright protection and proof of ownership for digital images. In contrast to the applications of robust watermarking, fragile and semi-fragile techniques are geared towards image authentication and localisation of tampered regions. In this section, we discuss the concepts, characteristics, differences and related algorithms of these three classes.
2.2.1 Robust Watermarking

The most important property of robust watermarking is its tolerate to certain signal processing operations that usually occur during the lifetime of a media object. A schematic diagram illustrating the main functions of robust watermarking is shown in Figure 2.7. The sender watermarks the original work via a watermark embedding process, and then sends the watermarked work to the recipient. The recipient extracts the watermark via a watermark detection process. During the transmission of the watermarked work, the image is open to manipulations, meaning the integrity of the watermark data could be compromised.

**Figure 2.7:** Schematic diagram for robust watermarking

**Block Based Robust Watermarking**

The most common image transform framework, the Discrete Cosine Transform (DCT), is frequently used for block based robust image watermarking [70–74]. Some researchers have also proposed adaptive watermarking schemes in the DCT domain based on image content. Perceptual models have been utilised to analyse the individual image content.
before the watermark bits are embedded into the DCT coefficients of each block, which could lead to an optimisation of the imperceptibility of the watermarked image [75].

Just Noticeable Distortion (JND) is designed to determine the maximum strength of a watermark signal that can be inserted into an image, and to improve the quality of the watermarked image. Kay and Izquierdo [72] proposed a robust content based image watermarking scheme by estimating a JND mask from image characteristics such as texture, edges and smoothness, from both the spatial and DCT domains. The watermark embedding process is shown in Figure 2.8. The original image is first divided into non-overlapping blocks of 8×8 pixels, then the JND mask is calculated from both spatial and DCT domains of each block. To embed the watermark bits, the selected DCT coefficients are modified according to the JND mask by using a key. An inverse DCT transformation is then applied to the modified DCT coefficients, and the blocks are merged back into a watermarked image.

In each block, the JND is derived in Equation 2.4.

\[
JND = \left( D_T - \frac{1}{2} (D_E + D_U) \right) + (128 - \bar{I})^2
\]  

where \( D_T = \log \left( \sum_{i=0}^{63} v_i^2 - v_0^2 \right) \) is the texture information retrieved directly from the DCT coefficients of a block and \( v_i, i = 0, \ldots, 63 \) are the 64 DCT coefficients of the considered block. Edges and smooth areas \( D_E = \frac{64 \times P_E}{\max(P_E)} \) are extracted from the pixel domain, \( P_E \) is the cardinality of the set of pixels within the block and at edge locations. The uniformity \( D_U \) in a block is defined as the number of pixels belonging to a uniform area that is extracted by the Moravec corner detection operator [76] for each block. \( \bar{I} \) is the mean of the luminance values of each block. In order to insert the watermark, the modified DCT coefficients \( v'_i \) are derived in Equation 2.5.

\[
v'_i = v_i + \alpha \times JND \times |v_i| x_i
\]  

\[ (2.5) \]
where $\alpha$ is a scaling parameter, and $z$ is the watermark, consisting of a sequence of real pseudo-random numbers [72]. Another type of JND mask is generated from Watson's visual model which was utilised in Podilchuk and Zeng's image adaptive watermarking scheme [74], and then further adopted and extended by Hernandez et al. [73], Li and Cox [73] and Li et al. [77]. In Watson's visual model, the JND values are extracted by calculating the luminance and contrast masking of each block DCT coefficients of the image.

Wong et al. [78] proposed an iterative watermark embedding algorithm for JPEG compressed images capable of embedding multiple watermarks within the DCT domain with different keys. In addition, Dong et al. [79] also proposed two algorithms that embedded a multi-bit watermark in the DCT domain of the image. Their first algorithm utilised an image normalisation technique which was robust to general geometric transformation attacks. Their second algorithm utilised a resynchronisation scheme based on a mesh model to combat nonlinear geometric attacks. Moreover, in Yeo and Kim's scheme [80], a generalised patchwork algorithm (which is the combination of the additive patchwork algorithm and the multiplicative patchwork algorithm) was employed to embed the watermark bits in the DCT domain. Their experimental results showed that their method was robust against JPEG compression attacks and some signal processing attacks, such as median filtering, FMLR (frequency-mode Laplacian removal), Gaussian filtering, sharpening, and random bend attacks.

In order to maintain quality of the watermarked image and the watermark robustness, many researchers experimented with applying different image transform techniques for block based robust image watermarking schemes. Some examples include robust image watermarking in the Fast Hadamard Transform (FHT) which resulted in a much shorter processing time and simpler hardware implementation [81]. Ho et al. [68] used Slant Transform (SLT) which provided significant advantage for watermark insertion and retrieval for images with complex textures such as satellite images. Singular Value Decomposition (SVD) based scheme was found to be robust against typical attacks, such as low-pass and high-pass filtering [82]. A Curvelet Transform based scheme was proposed to overcome inherent limitations of traditional multiscale representations watermarking schemes in [83]. Xie et al. [84] chose the middle subband for embedding watermarks to achieve both good imperceptibility and robustness in Ridgelet Transform domain.

**Non-block Based Robust Watermarking**

One of the most popular image transform domains for non-block based robust watermarking is the discrete wavelet transform (DWT) [67, 85–91]. In contrast to DCT, the original image is not divided into blocks in the DWT domain. DWT is one of
the most computationally efficient frequency transforms that utilises the human visual system (HVS). Moreover, in DWT based watermarking schemes, it is possible to embed watermarks with more energy, thereby significantly increasing their robustness [92]. A DWT domain based robust watermarking algorithm was proposed by Kundur and Hatzinakos [91], and their watermark embedding process is shown in Figure 2.9. In their scheme, both the original and logo images were first transformed into the wavelet domain and then decomposed into three levels (Figure 2.10(a)), and one level (Figure 2.10(b)), respectively. The DWT coefficients of the logo image were next embedded into the coefficients of the original image by using the multi-resolution fusion technique integrated with a model of HVS. Finally, the modified coefficients are transformed back into the spatial domain to create the watermarked image. In the experimental results, the authors claimed their proposed scheme is highly robust to compression and additive noise attacks and resilient to moderate linear mean filtering.

![Figure 2.9: Kundur and Hatzinakos's DWT watermark embedding process [91].](image)

Two other schemes for DWT watermarking were proposed by Xia et al. [85] and Zhu et al. [93]. Both algorithms used a Gaussian sequence of pseudo-random real numbers as the watermark data, instead of a logo image. Xia et al. [85] proposed a watermarking scheme that decomposes the image into two levels in the DWT domain. The watermark was embedded into the middle and high-pass sub-bands in the wavelet domain of the image, denoted by $HL_1$, $LH_1$, $HH_1$, $HL_2$, $LH_2$ and $HH_2$, as shown in Figure 2.10(a). Xia et al. claimed that their algorithm could tolerate additive noise, rescaling/stretching, compression attacks and that the algorithm was also more robust than the DCT approach. Zhu et al. [93] proposed a unified DWT watermarking approach that decomposed the original image into four levels in the DWT domain. Next, the watermark was embedded into the high-pass sub-bands in the wavelet domain of the image, denoted by $HL_1$, $LH_1$ and $HH_1$ as shown in Figure 2.10(a).

In recent years, many researchers have attempted to develop watermarking algorithms based on the combination of two or more image transform techniques, with the aspiration of improving the schemes [69, 94–96]. For example, Yang and Zhang [95] improved the algorithm proposed by Lin et al. [97] in the distributed discrete wavelet transform
Chapter 2. Literature Review

(DDWT) domain. The DDWT technique is adapted largely from the DWT approach, along with SVD for robust watermarking. A fragile watermark is adaptively embedded into the spatial domain of the watermarked image. By embedding both robust and fragile watermarks into a single image, the method is capable not only of identifying the ownership, but also of authenticating the integrity of the image to deduce whether it has been tampered or not. Mabtoul et al. [96] implemented a robust watermarking scheme based on Kingsbury's Complex Dual Tree Wavelet Transform (DT-CWT) [98]. The aim of designing this scheme was to overcome the drawback of DWT caused by the lack of shift invariance and poor directional selectivity for diagonal features. Mabtoul et al. claimed that the DT-CWT approach was more robust and effective than the DWT approach.

Wang et al. [99] proposed a novel feature-based watermarking scheme for Discrete Fourier Transform (DFT) embedding. Their watermark embedding process is shown in Figure 2.11. In this scheme, the Local Characteristic Region (LCR) was first extracted from the original image. In the LCR extraction process, a set of feature points was obtained by employing the Harris-Laplace detector to the original image. Then LCRs were constructed from these characteristic scales of the feature points and their locations. The extracted LCRs were found to increase the robustness against various attacks such as signal processing and affine transformations. In order to embed the watermark bits into the DFT domain of the LCR, the zero-padding operation was applied to map the LCR circle areas into blocks of $n \times n$ pixels. Finally, after modifying the coefficients to embed the watermark bits in the DFT domain of these blocks, a zero-removing operation was applied to map these $n \times n$-pixel blocks back into LCR circle areas to create the watermarked image. From the simulation results, the authors claimed that their proposed scheme was robust against common signal processing operators, such as median filtering, sharpening, noise adding, JPEG compression, rotation, scaling, translation, row or column removal, cropping and random bend attack. In Chapter 3, we will discuss our
proposed wavelet-based contourlet transform (WBCT) non-block robust watermarking scheme.

\[\text{Original image} \rightarrow \text{zero-padding} \rightarrow \text{DFT} \rightarrow \text{Inverse DFT} \rightarrow \text{Extract LCR} \rightarrow \text{Watermark embedding} \rightarrow \text{zero-removing} \rightarrow \text{Watermarked image}\]

**Figure 2.11:** Wang et al.'s watermark embedding process [99].

### 2.2.2 Fragile Watermarking

As mentioned in Chapter 1, fragile watermarking can be used to detect any small manipulations made to the original image [100]. Hence, any attacks that ultimately alter the intensity of an image can be detected, and the tampered regions can be located accurately when applying fragile watermarking schemes [101]. Many fragile watermarking algorithms are intentionally designed for use in the spatial domain (typically by altering the Least Significant Bits (LSB) of the image), as this domain is widely documented as being relatively fragile and sensitive to small changes [102–104]. Therefore, it is possible to exploit the inherent weakness of the LSB schemes, and implement a fragile watermarking scheme in the spatial domain. Fridrich [105] proposed a spatial domain based fragile watermarking scheme that could localise tampered regions of a watermarked image, by adapting Wong’s method [106]. The watermark embedding process is shown in Figure 2.12. The original image is first divided into non-overlapping blocks of 8 \(\times\) 16 pixels. In each block, the seven Most Significant Bits (MSB) of each pixel are extracted, and a cryptographic hash function is applied as illustrated in Figure 2.13. The logo is also divided into blocks of 8 \(\times\) 16 pixels and each block contains information about the original block position, image index, original image dimensions (resolution), camera ID and author ID (PIN). The seven MSBs of each block are hashed and then corresponding logo block are subjected to an Exclusive-OR (XOR) operation and then encrypted using a key. Finally, the LSBs of the original image are replaced with the result of the XOR operation and encrypted watermark bits, and creating the watermarked image. In the authentication process, the LSBs of the test image are extracted, and the seven MSBs from each block are hashed as shown in Figure 2.14. For each block, the LSBs are decrypted with a key, along with its corresponding hashed seven MSBs using the XOR operation. Finally, the authentication process is achieved by comparing each block of
Zhang and Wang [107] proposed a statistical scheme of fragile watermarking scheme that embeds a folded version of the authentication data derived from five most significant bits of the original image along with other additional data into the image with acceptable watermarked image quality (PSNR at 37.9dB). Their results showed their algorithm could localise the tampered pixels accurately. They further improved their method in [107] that could restore the tampered image content after localising the tampered area without any errors [108]. He et al. [109] proposed a conventional self-embedding fragile watermarking scheme based on adjacent-block based statistical detection method (SDM) that could withstand copy-paste and collage attacks. Their algorithm could identify the tampered blocks with a probability of detection accuracy of greater than 98% even the tampered area was almost 70% of the host image.
Fragile watermarking scheme can also be applied in transform domain. Li and Shi [11] proposed a fragile watermarking algorithm in DWT domain to achieve the requirements of high security, low distortion, and high accuracy of tamper localization for authenticating JPEG2000 images. Their algorithm could also tolerate vector quantization attack, Holliman-Memon attack, collage attack and transplantation attack. Aslantas et.al [110] proposed intelligent optimization algorithms (IOA) to improve fragile watermarking schemes in DCT domain. They used IOA which included Genetic Algorithm (GA), Clonal Selection Algorithm (CSA), Particle Swarm Optimization (PSO), and Differential Evolution (DE) to correct rounding errors caused by transforming an image from the frequency domain to the spatial domain with the objective of improving DCT-based fragile watermarking. The experimental results showed that the CSA produced better PSNR results whereas DE has lower computational time among these four intelligent optimization algorithms. Yeh and Lee [111] proposed reversible fragile watermarking by utilizing the pyramidal structure method. They selected appropriate embedding areas by analysing the pyramid-structure of the image for embedding watermark bits in the wavelet domain. The experimental results showed that their scheme could successfully localise even when 50% of the watermarked image was tampered and detecting counterfeiting attacks.

### 2.2.3 Semi-fragile Watermarking

Semi-fragile watermarking techniques for image content authentication have recently attracted much attention [91, 112, 113]. This is due to the fact that comparing to fragile watermarking, semi-fragile watermarking is not as sensitive as fragile watermarking. Semi-fragile schemes make it possible to verify the content of the original image, as well as permitting alterations caused by non-malicious (unintentional) modifications such as mild JPEG compression. A schematic diagram illustrating the main functions of semi-fragile watermarking is shown in Figure 2.15. The sender watermarks the original image via a watermark embedding process, and then sends the watermarked image to the recipient through the transmission channel. The recipient authenticates the test image by way of a watermark detection and authentication process. During the image transmission, the mild signal processing errors caused by signal reconstruction and storage, such as transmission noise or JPEG compression, are permissible. However, the image content tampering such as copy-and-paste attack will be identified as a malicious attack.

Many semi-fragile watermarking techniques have been already proposed by researchers. Lin et al. [114] proposed embedding algorithm that first applied DCT to $16 \times 16$-pixel blocks of the cover image, then embedded the watermarks in the middle to low frequency (except DC coefficient) of each block. Their scheme could identify the tampered area
with 75% accuracy under moderate compression and with near 90% accuracy under light compression. Ho et al. [13] proposed a semi-fragile watermarking scheme in the Pinned Sine Transform (PST) domain. In their watermark embedding algorithm, as shown in Figure 2.16, the original image is applied by using PST to get the pinned and boundary fields in 8 by 8 blocks. The watermark bits were then inserted into middle to high frequency of each block in the pinned field. The scheme also used a self-restoration method, originally proposed by Fridrich and Goljan [115] to recover the tampered regions. Their scheme could tolerate some common image processing manipulations such as JPEG and wavelet compression, and the detection rate was higher than DCT-based scheme. The algorithm has been further improved by using irregular Sampling instead of the LSB method [21], which aimed to improve the robustness of tampering restoration. We have adapted their scheme and proposed the Slant Transform (SLT) based semi-fragile watermarking scheme, which will be discussed in more detail in Section 4.1. Kundur and Hatziankos [24] proposed a DWT based algorithm called telltale tamper-proofing, which made it possible to determine tampered regions in multi-resolutions. Unlike other schemes that use DCT, this method does not require a block division process to detect the tampered regions due to the localisation ability of the wavelet transform. The localization ability of the wavelets in both spatial and frequency domains would potentially indicate a good candidate for semi-fragile watermarking.

Maeno et al. [116] presented two algorithms that focused on signature generation techniques. The first algorithm used random bias to enhance the block based DCT watermarking scheme proposed by Lin and Chang [18]. The second algorithm used nonuniform quantisation on a non-block based semi-fragile watermarking scheme in the wavelet domain. Their experimental results showed their method was fragile to malicious manipulations, but robust to non-malicious manipulations such as JPEG and JPEG2000 compression. Ding et al. [117] also proposed a method by using DWT. In their algorithm, chaos was used to generate a pseudo-random sequence as a watermark, in an
effort to improve the overall security. This was an improvement to the more traditional methods of generating a pseudo-random sequence. The sub-bands (\(HL_2, LH_2, HH_2\)) were used for embedding the watermark after applying a 2-level wavelet decomposition of the original image. The normalized cross-correlation (NC) was used to evaluate their algorithm by comparing the original watermark with the extracted watermark after applying JPEG compression and Additive white Gaussian noise (AWGN) manipulations. Ni et al. [118] proposed a robust lossless data hiding technique that could be employed into their semi-fragile watermarking scheme. The different bit-embedding strategies for groups of pixels with different pixel grayscale value distributions and error correction codes were utilized in their scheme. They analyzed their results under both lossless and lossy JPEG compression. If the watermarked image has experienced losslessly compression, the watermark bits can be extracted correctly and the image will be classified as authentic and the original image can be recovered exactly. If this losslessly compressed watermarked image has been further undergone lossy compression, the original image will not be able to recover and will be rendered authentic as long as the compression is not so severe that the content has been changed.

It was observed that Wavelet transforms are not optimal in capturing the contours or edges of the host image [119], which are vital to image authentication. To overcome this drawback, several multiscale and directional transforms have been proposed and proven to be more efficient than wavelets for capturing smooth contours and edges in natural images. Some examples include steerable pyramid [94], ridgelet [120], curvelet [119], bandlet [121], contourlet [122] and the WBCT [123]. In this thesis, we propose a novel WBCT based robust watermarking scheme. This will be discussed in more detail in Chapter 3. We then exploit the advantage of WBCT, and incorporate it into our proposed novel WBCT-based semi-fragile watermarking algorithm, which will be discussed in more detail in Section 4.2.

Figure 2.16: Ho et al.'s PST-based watermark embedding process [13].
2.3 Image Forensics

Recently, there is a significant amount of interest in identifying reliable techniques capable of accurately proving the authenticity of an image, without the requirement of actively inserting a digital watermark or signature into the data. Whilst the watermarking schemes have been shown to be useful for copyright protection and verifying the integrity of the image, there always exists the underlying risk that the watermark data might be forcibly or accidentally removed. When this happens, the image is effectively stripped of its identity. Forensic techniques aspire to achieve similar objectives but do not rely on the strength of embedded data. Instead, the ambition is to prove the authenticity of an image based solely on the data provided. In the following subsections, we first classify the image forensic techniques into two main areas: image source detection and image manipulation detection. We then discuss one of the image forensic techniques - Benford's Law, which can be utilised for improving the detection rates of semi-fragile watermarking schemes in Chapter 5.

2.3.1 Classification of Image Forensics

Many researchers have classified image forensics into different categories according to the applications. For instance, Fridrich et al. classified image forensics into six categories, which are source classification, device identification, device linking, anomaly investigation, processing history recovery, and forgery detection [29]. Moreover, the field of image forensics has also been divided into image source identification, identification of synthetic images and image forgery detection by Memon et al. [124]. Furthermore, Farid et al. have classified image forensics into five areas, which are pixel based techniques, format based techniques, camera based techniques, physically based techniques and geometric based techniques [30]. In general, image forensics can be divided into two main areas: image source detection and image manipulation detection.

Image Source Detection

Image source detection is the task of successfully linking suspect images to its source such as camera, scanner and printer devices. In this area, from the results of analysing captured images, the brand or model of device could be potentially classified, such as a Canon or Nikon camera. In addition, image source detection techniques could also be used to identify the source of the image by locating unique anomaly features within the image, and therefore determine which device is used in capturing the image. One of the earliest reported approaches for digital camera identification was based on the characteristic of the imaging sensor in the device [125]. The imaging sensor is arguably the most important component of the image acquisition process, as it captures the light
intensity of the scene on a pixel-by-pixel basis, and converts it into an electrical signal. From here, the signal will pass through a Colour Filter Array (CFA), however, it is possible that the imaging sensor operates with an element of noise, caused by hot or dead pixels. Errors such as these dead pixels can often be seen in the final image, even if the image has been lossy compressed. As these errors are likely to be slightly different for different devices, the technique is useful for reliably linking images to the source sensor and therefore the source camera that captured the image. However, most modern and high-end digital cameras are able to detect deficiencies in the processing such as these dead pixels, and often remove them altogether. As the forensics scheme relies on the existence of such pixels, it can only be targeted towards less advanced cameras that do not correct these types of errors.

Another prominent research area of image source detection has been proposed by Lukáš et al. [25, 126], and later by Khanna et al. in 2009 [27]. The technique relied on sensor pattern noise, which is a deterministic component that remains consistent for all images that the sensor captures. Pattern noise can be sub-divided into two categories: fixed pattern noise (FPN) and photo-response non-uniformity noise (PRNU). The FPN is an additive noise that is suppressed to varying standards by many camera manufacturers. This noise is amounted with the camera’s exposure and temperature [25]. For these reasons, it is not reliable for camera identification purposes as it can vary inconsistently. PRNU, on the other hand, is a multiplicative noise and contains a property refereed to as pixel non-uniformity (PNU), which is defined as the sensitivity differences to light at each pixel. The PNU is a direct result of the manufacturing process and is therefore not influenced by exposure and light. Indeed, the PNU noise remains the same for each image captured, meaning this component is extremely useful for determining the source camera. To perform the classification, a reference pattern for the camera must first be identified. The pattern noise obtained from a suspect image can now be compared with the pattern noise obtained from the device itself. If the correlation is identical, then there can be little doubt that the image is originated from the device, as the chances of two camera’s producing the same pattern noise are extremely remote.

Li [127] proposed an enhanced sensor pattern noise approach based on a hypothesis to improve the device identification rate of the identifier. The hypothesis is: the stronger a signal component is, the more likely it is associated with strong scene details, and thus the less trustworthy the component should be. The experimental results of his enhanced sensor pattern noise method illustrated a greater performance than the original sensor pattern noise based image forensic schemes.

In addition, Bateman et al. [128] demonstrated the benefits of using Statistical Process Control (SPC) for analysing image data on a range of different digital cameras. Based on
their research, an anomaly in the camera processing elements was identified for iPhone 3G devices, whereby the brightness of the images was found to be fluctuating. By analysing the latest iPhone 3GS model under the same conditions, they proved that the newer model does not show this property, indicating that reliable detecting iPhone 3G and iPhone 3GS devices is possible based on the captured images.

**Image Manipulation Detection**

Image manipulation detection focuses on the detection of images that have been manipulated. These manipulations include malicious (intentional) modifications, such as image content copy-and-paste attack and non-malicious (unintentional) modifications, such as format changing, image enhancement and compression. One of the most common content malicious manipulations is *splicing*, which involves removing content from one image and overwriting it with something similar from another image to form a composite. This type of modification can be dated back over 150 years; a famous example of which is the Abraham Lincoln portrait \[129\]. In this example as shown in Figure 2.17, a portrait of John Calhoun was manipulated such that it appeared as if the portrait was of Abraham Lincoln. In fact, Lincoln never posed for the portrait, and the image was actually constructed by flipping and resizing Lincoln's head from a head-shot photograph taken by Matthew Brady such that it resembled the same proportions as the Calhoun portrait. Calhoun's face was then replaced by Lincoln's face to produce a composite image.

![Figure 2.17: The Lincoln composite.](image)

An image forensic technique proposed by Farid et al. found that composite images could also be identified by studying the light reflected into the subjects eyes \[130\]. The positioning of white dots (caused by flash photography) indicated the direction of the light when the image was captured. When images were spliced together, these issues were often overlooked. When several people all appear in the scene the correlation of the light direction will match almost exactly. However, when a person has been spliced into the image from another image, the direction of light on the subjects eyes will not
match. By studying the light pattern, it is often a fairly trivial process to determine whether the image is genuine or not. Similarly, Johnson and Farid [26] discusses how lighting observations would be applied more generally to images. They explained that the light striking a surface was dependant on the position of the light source. As such, an estimate of the direction of the light source could be derived from an image by reviewing a given object’s 2-D surface contour, such as a human jawline and chin. The lighting of an object can ultimately be compared against that of other objects in the photo, and if there exists a mismatch in lighting directions, then the image is likely to be faked.

Furthermore, the sensor pattern noise technique (PRNU), as discussed earlier for image source detection, can also be adapted to authenticate images. As the complete pattern noise exists for every pixel in an image, a manipulated image can be derived when the pattern noise is not present at a particular region of interest. It is important to note, however, that the PRNU noise will not be present in highly saturated areas of clean images, and is also highly suppressed in dark areas, as the noise is multiplicative. Therefore, a region that does not contain the pattern noise should be checked to ensure that neither of these two properties hold true before classifying the image as tampered. Further details of how this can be achieved statistically are discussed in [131]. However, Li et al. [132] demonstrated that the sensor pattern noise of the original images can be modified or replaced. Therefore, the detection accuracy could be reduced if the image forensic investigation process relies on analysing sensor pattern noise only.

Image forensic techniques can be also used to detect the non-malicious (unintentional) modifications. For example, in some cases, images are enhanced through image editing software to provide better visual clarity, even though the content itself will remain true. Similarly, the images could have been compressed to minimise storage. Moreover, processing history recovery could be used to detect non-malicious (unintentional) modifications, such as JPEG/JPEG2000 compression ratio [29]. Zhang et al. [133] proposed a double compression detection technique for JPEG2000 compressed images. Double compression occurs when an image is saved twice in the same image format with different or similar compression. In their scheme, they applied the DWT to a JPEG2000 compressed image, and extracted the High/Low and Low/High sub-bands of the DWT coefficients. A histogram was then formed by applying the Fast Fourier Transform (FFT) to these extracted coefficients. By analysing the sharp peaks and valleys of this histogram, the test image could be classified according to whether or not it has been subjected to a double JPEG2000 compression. Fu et al. [134] proposed an image forensic technique to detect the Quality Factor (QF) of unknown JPEG compressed images by using the Benford’s Law. The application of Benford’s Law will be discussed in more detail in the next section as this is one of the contributed areas of this thesis.
2.3.2 Benford’s Law for Image Forensics

Benford’s Law was introduced by Frank Benford in 1938 [135] and developed by Hill [136] for analysis of the probability distribution of the first digit (1 – 9) of numbers from natural data in statistics. Benford’s Law has also been applied to accounting forensics [137, 138] and image processing [139], [140]. The basic principle of Benford’s Law is given as follows:

\[ P(x) = \log_{10} \left(1 + \frac{1}{x}\right), x = 1, 2, ... 9 \]  

(2.6)

where \( x \) is the first digit of the number and \( p(x) \) is the probability distribution of \( x \).

Fu et al. [134] observed that the 1st digits of DCT coefficients (uncompressed) of 1338 images were found to be obeying the Benford’s Law, as shown in Figure 2.18.

They also analysed the 1st digits of JPEG coefficients (compressed) of images with different Quality Factors (QF), and found that 1st digits of JPEG coefficients obeyed their proposed generalised Benford’s Law as given in the following shown equation:

\[ p(x) = N \log_{10} \left(1 + \frac{1}{s + x^q}\right), x = 1, 2, ... 9 \]

(2.7)

where \( N \) denote normalisation, and \( s \) and \( q \) are model parameters. Table 2.1 illustrates an example of three best fitted parameters to their corresponded QFs by using the curve fitting tool in Matlab [134]. Moreover, the probability distributions were not
following the generalized Benford’s Law if the image had been compressed twice with different quality factors. Thus, by utilizing this property, the QF of the image could be estimated. In Chapter 5, we propose a framework that further explores the generalised Benford's Law as an image forensics technique, to accurately detect the unknown JPEG compression in semi-fragile watermarking images to improve the detection rate.

Table 2.1: Example of three parameters [134].

<table>
<thead>
<tr>
<th>Quality Factors</th>
<th>N</th>
<th>q</th>
<th>s</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.456</td>
<td>1.47</td>
<td>0.0372</td>
</tr>
<tr>
<td>90</td>
<td>1.255</td>
<td>1.563</td>
<td>-0.3784</td>
</tr>
<tr>
<td>80</td>
<td>1.324</td>
<td>1.653</td>
<td>-0.3739</td>
</tr>
<tr>
<td>70</td>
<td>1.412</td>
<td>1.732</td>
<td>-0.337</td>
</tr>
<tr>
<td>60</td>
<td>1.501</td>
<td>1.813</td>
<td>-0.3025</td>
</tr>
<tr>
<td>50</td>
<td>1.579</td>
<td>1.882</td>
<td>-0.2725</td>
</tr>
</tbody>
</table>

2.4 Summary

In this chapter, The fundamental concept of digital watermarking was described. We presented requirements such as imperceptibility, robustness, capacity and security for digital watermarking, characteristic of spatial and transform domain based watermarking schemes. We also reviewed different watermarking algorithms in the literature according to three classifications: “robust”, “fragile” and “semi-fragile”. Finally, we reviewed image forensics techniques briefly, and outlined image forensics into two main areas: image source detection and image manipulation detection, as well as the use of Benford’s Law for image forensics.
Chapter 3

Novel Robust Watermarking Algorithm Based on WBCT

One of the important applications of digital watermarking technology is copyright protection and ownership identification for digital images. To achieve this goal, robust watermarking has been rapidly developed in the past decade or so. Robust watermarking is designed to survive various non-geometric manipulations such as JPEG compression, additive noise and filtering as well as some geometric distortions such as rotation and scaling. In Chapter 2, a number of different transforms and algorithms used for robust image watermarking have been reviewed. These include block based DCT, non-block based DWT and other state-of-the-art watermarking algorithms operating in the transform domain. In contrast to conventional transform domain methods, a new adaptive robust watermarking algorithm using the non-redundant contourlet transform known as the WBCT is presented in this chapter. From experiments, we exploit the energy relations between parent and children coefficients, which are invariant before and after JPEG compression. Results show that even for QF set as low as 10, the percentages of invariant energy relations of all test images were above 75% after JPEG compression. This invariance feature is therefore very useful for robust image watermarking. The results of WBCT are evaluated and compared with two other DWT based algorithms achieving a high degree of robustness against most non-geometrical and geometrical attacks, while maintaining an excellent perceptual invisibility.
3.1 Wavelet-based Contourlet Transform (WBCT)

The contourlet transform can be realised efficiently via a double-iterated filter bank structure. In the double filter bank, the Laplacian Pyramid (LP) [141] is first used to capture the point discontinuities. In the LP stage, the image is decomposed into a low-pass and a set of band-pass sub-bands. Each band-pass image is then further decomposed by a directional filter bank (DFB) [142] into a number of sub-bands to capture the directional information and link-point discontinuities into linear structures. Subsequently, the image is decomposed into several directional sub-bands at multiple scales.

Eslami and Radha [123] developed a WBCT, also as non-redundant contourlet transform, by replacing the LP with a wavelet, followed by implementing a directional filter bank (DFB) into the wavelet sub-bands to extract the directional information. As shown in Figure 3.1, at each level in the wavelet decomposition, the three high-pass bands corresponding to the LH, HL, and HH bands can be obtained. DFB is applied with the same number of directions to each band at a given level. As a result, each LH2, HL2 and HH2 are decomposed into four directional sub-bands, and each LH1, HL1 and HH1 are decomposed into eight directional subbands.

WBCT was developed as an improvement to the wavelet transform that is inefficient when extracting smooth contours. It has the multiscale and time-frequency localisation property of wavelets, but it also provides a high degree of directionality and anisotropy [122]. The main advantage of WBCT is that a non-redundant multi-resolution and multidirectional expansion of images can be achieved. The transform has been successfully applied in image coding [123] and image fusion [143]. As an example, the WBCT coefficient map of the image 'peppers' is shown in Figure 3.2 to demonstrate the property of this transform.

Eslami and Radha [123] stated that the WBCT parent-children relationship was different from the relationship that exists in conventional wavelet domains. In a conventional wavelet-domain, the parent-children links are always in the same direction among the three wavelet directions, as shown in Figure 3.3(a). WBCT coefficients, on the other hand, comprise four children in two separate directional sub-bands for each LH, HL and HH sub-bands, as shown in Figure 3.3(b). In Figures 3.3(a) and 3.3(b), the blank square is the parent coefficient and the four white squares (arrowed) are their children. This special relationship and characteristic of WBCT form the fundamental basis for our novel robust watermarking algorithm.
Chapter 3. Novel Robust Watermarking Algorithm Based on WBCT

---

**Figure 3.1:** The framework of the WBCT.

**Figure 3.2:** (a) Original image (b) coefficient map after level-2 WBCT applied.

**Figure 3.3:** Parent-children relationship for DWT and WBCT.
3.2 Coefficient Relations with JPEG Compression of Original Images

In this section, we first investigate the characteristics of the energy relations of the original images between the parent and the children coefficients before and after JPEG compression (the JPEG compression attack is one of the most common attacks in robust watermarking) in a three level WBCT. As per the pseudo-code below, suppose parent coefficients in the original image and JPEG compressed image are denoted as $P$ and $P'$, respectively, and its corresponding four children as $C_i$ and $C'_i$, where $i = 1, 2, 3, 4$.

\[
\text{if } (|P| > \text{mean}|C_i| \land |P'| > \text{mean}|C'_i|) \lor (|P| < \text{mean}|C_i| \land |P'| < \text{mean}|C'_i|) \text{ then }
\]

We assume the energy relations are invariant before and after compression.

\[
\text{end if}
\]

Six standard $512 \times 512$ test images (Figure 3.4) are used in this experiment to determine the invariant energy relationship as shown in Figure 3.5. As the quality factor (QF) decreases from 90 to 10, the average percentages of invariant relations also gradually decrease. For $QF = 90$, it reaches above 95%, and for $QF = 10$, although the image is distorted significantly, it still maintains above 75%. From Figure 3.5, it can be observed that highly textured images such as ‘San Diego’, ‘Bridge’ and ‘Baboon’ all performed relatively better than the other images. In addition, we also analysed 1000 standard grayscale test images and the results showed similar characteristics. Overall, an improved performance can be achieved for all images by exploiting the modulation of their energy relationship.

3.3 Watermark Embedding Process

In this section, we describe the watermark embedding process as shown in Figure 3.6. Three level WBCT is first applied to the image. We then randomly select a number of parent coefficients (black squares) with their corresponded children coefficients (white squares) using a key, as shown in Figure 3.7. The total number of these parent coefficients is equal to the length of the watermark, which is a pseudo-random binary sequence $\{-1, 1\}$. For each selected parent coefficient, we then embed the watermark bits by modulation as expressed in Equation 3.1.
Figure 3.4: Six standard test images

\[
P' = \begin{cases} 
  P, & (|P| \geq |C_{avg}|) \land (w = 1) \\
  P, & (|P| < |C_{avg}|) \land (w = -1) \\
  P + K1 \times (|C_{avg}| - P), & (|P| < |C_{avg}|) \land (w = 1) \land (P \geq 0) \\
  P - K1 \times (|C_{avg}| - |P|), & (|P| < |C_{avg}|) \land (w = 1) \land (P < 0) \\
  P + K2 \times (P - |C_{avg}|), & (|P| \geq |C_{avg}|) \land (w = -1) \land (P \geq 0) \\
  P - K2 \times (|P| - |C_{avg}|), & (|P| \geq |C_{avg}|) \land (w = -1) \land (P < 0) 
\end{cases} 
\]

where \( P' \) is the watermarked parent coefficient, \( P \) is the original parent coefficient, \( C_{avg} \) is the average of four children coefficients, \( w \) is the watermark bit, \( K1 \) and \( K2 \) are thresholds to determine the trade-off between imperceptibility and robustness. Note that we only modulate the parent coefficients selected from the LH3, HL3 and HH3 bands.
Chapter 3. Novel Robust Watermarking Algorithm Based on WBCT

Figure 3.5: Percentage of invariant energy relations after JPEG compression.

subbands. After the embedding steps, the watermarked image is reconstructed using the inverse WBCT transform.

Figure 3.6: The proposed WBCT watermark embedding process.

3.4 Watermark Detection Process

In Figure 3.8, we present our proposed algorithm for watermark detection. First, the WBCT transform is performed on the watermarked image before the tree structures are selected using the key. The absolute value of the parent is compared with the absolute
value of the average of the children, and if the former is greater than or equal to the latter, then the watermark bit ‘1’ is obtained, otherwise, ‘-1’ is obtained. The process is repeated for every tree structure to retrieve and construct the entire watermark data. A normalised correlation is used to determine whether a watermark is present or not, by comparing it to a pre-specified threshold. The normalized correlation is computed in Equation 3.2.

\[
NC(w, \bar{w}) = \frac{\sum w(n) \bar{w}(n)}{\sqrt{(\sum w(n) \sum \bar{w}(n))}}
\] (3.2)

where \(w\) is the given watermark, and \(\bar{w}\) is the extracted watermark. If \(NC \geq \tau\), then the watermark is present in the image. We adapt the threshold based on the false positive probability [144]. Based on empirical results for our algorithm, \(N_w = 512\) is the number of the watermarks, \(\tau\) is chosen to be 0.23 for a false positive probability of \(1.03 \times 10^{-7}\), which has been proved in [144].

![Figure 3.7: A demonstration of parent with their corresponded children coefficients in three level WBCT.](image)

3.5 Experimental Results

In this section, six grayscale images ‘Lena’, ‘Peppers’, ‘Goldhill’, ‘San Diego’, ‘Bridge’, and ‘Baboon’ (each of size 512 \( \times \) 512) are used for our experiments to evaluate our proposed WBCT watermarking method.
3.5.1 Imperceptibility

In Table 3.1, the PSNR is used to evaluate the perceptual distortion of these images before and after watermark embedding. High PSNR values indicate that the watermarked data is highly imperceptible. We compare these results with Wang and Lin [145], as they proposed a watermarking method based on a wavelet tree quantisation and obtained a strong robustness to several different attacks. Our proposed contourlet method achieves higher PSNR values than the other two wavelet methods. Figure 3.9 illustrates the original and watermarked images for ‘Goldhill’ and ‘San Diego’, with approximately 42dB and 39dB, respectively.

<table>
<thead>
<tr>
<th>Image</th>
<th>Our method (dB)</th>
<th>Wang’s method (dB) [145]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>41.46</td>
<td>38.2</td>
</tr>
<tr>
<td>Peppers</td>
<td>39.24</td>
<td>38.7</td>
</tr>
<tr>
<td>Goldhill</td>
<td>42.22</td>
<td>39.8</td>
</tr>
<tr>
<td>San Diego</td>
<td>38.81</td>
<td>Not Available (NA)</td>
</tr>
<tr>
<td>Bridge</td>
<td>39.43</td>
<td>NA</td>
</tr>
<tr>
<td>Baboon</td>
<td>40.22</td>
<td>NA</td>
</tr>
</tbody>
</table>

3.5.2 Robustness

The robustness of our proposed WBCT watermarking method has been tested against different attacks including non-geometrical and geometrical attacks. Figures 3.10-3.15 illustrate detailed analysis of watermark robustness of WBCT against JPEG2000 compression, Gaussian noise, salt and pepper noise, contrast stretching, circular shifting and scaling, respectively. Furthermore, we compare these results with two conventional wavelet approaches based on Wang and Lin [145] and Tsai and Lin [146]. The experiments on WBCT against different attacks are summarised in Table 3.2, and comparative results between WBCT and the two wavelet transforms are shown in Table 3.3. The normalised correlation value below a threshold of approximately 0.23 means it has failed to detect the embedded watermark.

For JPEG and JPEG2000 compression attacks, different quality factors (QF) were used on watermarked images. Tables 3.2 and 3.3 show the effectiveness of our algorithm even when $QF = 10$, whereas the other two algorithms only provided results at $QF = 20$ and $QF = 30$. For the case of JPEG2000 shown in Figure 3.10, our method achieved results even at $QF=10$. This has not been reported by other existing watermarking algorithms at this low compression value. Figure 3.11 illustrates the images distorted under different Gaussian white noise. Similarly, salt and pepper noise addition for different amount of
noise densities is shown in Figure 3.12. Figure 3.13 shows the images under contrast stretching attack for different degrees of strength. Tables 3.2 and 3.3 also highlight the results achieved for mean filtering, histogram equalisation, median filter, Gaussian filtering, and sharpening. From the results, our method outperformed the other two algorithms in all cases, with median filtering the only exception in Table 3.3.

For geometrical attacks, we investigate the robustness of our proposed WBCT watermarking algorithm against different amounts of pixels due to circular shifting and scaling, as shown in Figures 3.14 and 3.15, respectively. Table 3.3 summarises the different non-geometrical and geometrical attacks used in comparing our method with the other two methods. In Table 3.3, “shifting A” indicates circular shifting, and “shifting B” indicates a deletion of lines followed by duplication of the adjacent lines. Figure 3.16 illustrates the watermark detected from watermarked images ‘Goldhill’ after attacked with JPEG Compression $QF = 20$ (Figure 3.16(a)), 400 pixels random shifted (Figure 3.16(b)), 20 degree of contrast stretching (Figure 3.16(c)) and Gaussian white noise with
variance = 0.01 (Figure 3.16(d)), respectively. Overall, our method achieved relatively better performance than Wang [145] and Tsai [146] except for the rotation and scaling attacks in Table 3.3.

**Figure 3.10:** Performance of our method after JPEG2000 compression

**Figure 3.11:** Performance of our method after Gaussian white noise
Chapter 3. Novel Robust Watermarking Algorithm Based on WBCT

Figure 3.12: Performance of our method after Salt & Pepper noise

Figure 3.13: Performance of our method after contrast stretching
Chapter 3. Novel Robust Watermarking Algorithm Based on WBCT

Figure 3.14: Performance of our method after pixel circular shifting

Figure 3.15: Performance of our method after scaling
TABLE 3.2: Performance of our method under different attacks.

<table>
<thead>
<tr>
<th>Image</th>
<th>JPEG (QF = 10)</th>
<th>Mean Filter (3 x 3)</th>
<th>Mean Filter (5 x 5)</th>
<th>Histogram Equalization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>0.28</td>
<td>0.54</td>
<td>0.29</td>
<td>0.47</td>
</tr>
<tr>
<td>Peppers</td>
<td>0.26</td>
<td>0.60</td>
<td>0.23</td>
<td>0.45</td>
</tr>
<tr>
<td>Goldhill</td>
<td>0.34</td>
<td>0.60</td>
<td>0.32</td>
<td>0.55</td>
</tr>
<tr>
<td>San Diego</td>
<td>0.52</td>
<td>0.59</td>
<td>0.29</td>
<td>0.60</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.48</td>
<td>0.63</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>Baboon</td>
<td>0.45</td>
<td>0.65</td>
<td>0.30</td>
<td>0.77</td>
</tr>
</tbody>
</table>

3.6 Summary

In this chapter, a novel robust image watermark embedding and detection algorithm in wavelet-based contourlet transform domain was presented. Through experiments, most of the energy relations between parent and children non-redundant contourlet coefficients maintained 75% of invariance before and after JPEG compression QF = 10, although the image was distorted significantly. Therefore, performance improvement was obtained by means of embedding a watermark via exploiting the modulation of the energy relations. By comparing with Wang and Lin[145] and Tsai and Lin’s [146] methods, the experimental results based on 16 set of attacks using 6 images showed that our non-redundant contourlet method was more robust to attacks such as JPEG, JPEG2000 compression, pixel shifting, histogram equalisation, median filter, Gaussian filtering, and sharpening. The results showed that our non-redundant contourlet method was highly robust to different kinds of attacks including non-geometrical and geometrical attacks. These include JPEG2000 compression (as low as QF=10), 400 pixels circular shifting, and contrast stretching, (as low as 20%). However, for the median filtering, scale and rotation attacks, our results are lower than other two methods that need to be further improved.
Table 3.3: Comparison of our method’s performance with two other Methods

<table>
<thead>
<tr>
<th>Attacks</th>
<th>Image</th>
<th>Wang’s Method</th>
<th>Tsai’s Method</th>
<th>Our Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG (QF = 30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.15</td>
<td>0.45</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.23</td>
<td>0.44</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.34</td>
<td>0.37</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>JPEG (QF = 25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>NA</td>
<td>0.37</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>0.29</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NA</td>
<td>0.23</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Median Filter (4 x 4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.23</td>
<td>0.38</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.24</td>
<td>0.33</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>0.36</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Median Filter (5 x 5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>NA</td>
<td>0.43</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>0.32</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NA</td>
<td>0.41</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>Shifting A (9 pixels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.26</td>
<td>0.27</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.29</td>
<td>0.35</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.29</td>
<td>0.36</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Shifting B (9 pixels)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.25</td>
<td>0.29</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.26</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.28</td>
<td>0.31</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Multiple Watermarking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.11</td>
<td>0.24</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.18</td>
<td>0.29</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.22</td>
<td>0.25</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>Scale &amp; Rotation (1°)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.24</td>
<td>0.25</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.15</td>
<td>0.25</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.17</td>
<td>0.26</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Scale &amp; Rotation (−0.75°)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.24</td>
<td>0.30</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.38</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>0.30</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>Gaussian Filtering</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.64</td>
<td>0.89</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.56</td>
<td>0.91</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.74</td>
<td>0.92</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>Sharpening</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.46</td>
<td>0.87</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.39</td>
<td>0.63</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.62</td>
<td>0.89</td>
<td>0.91</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.16: Attacked watermarked image 'Goldhill' - (a) JPEG Compression QF=20, (b) 400 pixel random shifted, (c) 20 degree of contrast stretching, (d) Gaussian white noise, variance=0.01.
With the tremendous growth and use of digital cameras and video devices, the need to verify the collected digital content for law enforcement applications such as crime scene investigations and traffic violations, becomes paramount if they are to be used as evidence in courts. Semi-fragile watermarking has become increasingly important within the past few years as it can be used to verify the content of images by accurately localising the tampered area and tolerating some non-malicious manipulations. There have been a number of different transforms used for semi-fragile image watermarking, as reviewed in Chapter 2.

In this chapter, we present two novel transforms for semi-fragile watermarking, using the Slant transform (SLT) as a block-based algorithm and the wavelet-based contourlet transform (WBCT) as a non-block based algorithm. The proposed SLT is compared with existing DCT and PST semi-fragile watermarking schemes. Experimental results indicate that the SLT is more accurate for copy and paste attacks with non-malicious manipulations, such as additive Gaussian noise. In Chapter 3, WBCT has already been successfully demonstrated its capability for robust watermarking. For semi-fragile watermarking, watermarking embedding is performed again by modulating the parent-child relationship in the contourlet domain. Experimental results using the same test images have demonstrated that our proposed WBCT method achieves good performances in localising the tampered regions, even when the image has been subjected to non-malicious manipulations such as JPEG/JPEG2000 compressions, Gaussian noise, Gaussian filtering, and contrast stretching. The average false positive rate is found to
be approximately 1% while maintaining an average false negative rate below 6.5%. We will also analyse and compare the difference between these two schemes.

### 4.1 Novel SLT Based Semi-fragile Watermarking Algorithm

This section provides an introduction to the slant transform, and discusses the details of the embedding, detection and authentication processes associated with watermarking. Experimental results are also presented in this section.

#### 4.1.1 The Slant Transform (SLT)

The Slant Transform has been applied to image coding in the past [147] and was recently adopted for robust image watermarking [148]. The SLT can be considered as a fast computational algorithm that provides a significant bandwidth reduction, resulting in a lower mean-square error for moderate size image blocks [147]. In addition, for textured images, the quality of the Slant Transformed images is higher than images coded by using other transforms such as DCT and Hadamard [149]. Moreover, as a similar image processing application to Walsh-Hadamard transform, Slant transform can be identified as a sub-optimum for energy compaction. This is essential for digital watermarking as the robust information hiding can be ensured by capturing the spread of middle to higher frequency bands [148]. Furthermore, Slant transform is simpler, faster and especially suitable for highly textured images [148]. Hence, the Slant Transform is proposed for semi-fragile watermarking and authentication of images in this section.

The authentication as the method to corroborate the genuineness of an object is mainly focusing on examining whether the image has been tampered or not, the location(s) of tampered region(s) and to what extent it has been changed. Furthermore, the SLT can also be used for compressing the original image [149], providing a mean to self-recovering the tampered regions by embedding the compressed cover image into the LSBs of the watermarked image [115]. The forward and inverse of SLT [147–149] can be expressed as follows:

\[
[V] = [S_N][U][S_N]^T \tag{4.1}
\]

\[
[U] = [S_N]^T[V][S_N] \tag{4.2}
\]

where [U] represents the original image of size \(N \times N\), [V] represents the transformed components and \([S_N]\) is the \(N \times N\) unitary Slant matrix given by
Chapter 4. *Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT* 47

\[ [S_N] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & a_N & b_N & 0 & 1 & 0 & 0 \\ 0 & 1 & I_{(N/2)-2} & 0 & 0 & 1 & 0 & 0 \\ -b_N & a_N & 0 & 0 & b_N & a_N & 0 & 0 \\ 0 & 0 & I_{(N/2)-2} & 0 & 0 & -I_{(N/2)-2} \end{bmatrix} \begin{bmatrix} S_{N/2} \\ 0 \\ 0 \end{bmatrix} \] (4.3)

where \( S_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \) as base case, \( I_{(N/2)-2} \) is the identity matrix of dimension \((N/2) - 2\), and \( a_{2N} = \left( \frac{3N^2}{4N^2-1} \right)^{1/2}, b_{2N} = \left( \frac{N^2-1}{4N^2-1} \right)^{1/2} \) are constants.

### 4.1.2 SLT Watermark Embedding Process

A novel semi-fragile Slant Transform digital watermarking method is adopted based on previous work relating to PST [13] and a self-restoration method [115]. The embedding process using the Slant Transform is illustrated in Figure 4.1. The first 7 bits of the original image are extracted and divided into 8 x 8-pixel blocks, SLT method is then applied to each block. The watermark embedding algorithm is then utilised, which is illustrated in the pseudo-code below. The watermarks for each block are then randomly generated by input a key as a seed. The obtained watermarks are embedded into the mid-band of each 8 x 8-pixel block. After watermark embedding, frequency coefficients of each block of the watermarked image are converted back by using the inverse Slant Transform. Consequently, the first 7 bits of watermarked image is obtained.

\[
\text{If } w = 1 \text{ And } x > \tau, \text{ Then } y = x, \text{ Else } y = \alpha. \\
\text{If } w = 0 \text{ And } x < -\tau, \text{ Then } y = x, \text{ Else } y = -\alpha. 
\]

where \( w \) is the watermark bit, \( x \) is the SLT coefficient of the host, \( y \) is the modified SLT coefficient. Through empirical experiments, we set \( \tau = 3.5 \) is the threshold which controls the perceptual quality of the watermarked image, and \( \alpha \) is a constant that randomly selected between 2.5 and \( \tau \).

The original image is also divided into 8 x 8-pixel blocks and undergoes the same Slant Transform; compression for each sub-block is then achieved by discarding the high frequency coefficients. Accordingly, 64 bits information for each block is acquired after compression and then encrypted by utilizing a key. Obtained blocks are then shuffled, e.g. the value of block 1 moves to block 50, the value of block 35 moves to block 10.
Chapter 4. Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT 48

Therefore, the LSBs of the watermarked image are then gained. Finally, the combination of the first 7 bits and LSBs of the watermarked images are formed into final watermarked image.

![Proposed SLT watermark embedding process.](image)

**Figure 4.1:** Proposed SLT watermark embedding process.

### 4.1.3 SLT Watermark Detection, Authentication and Restoration

The proposed semi-fragile Slant Transform for image authentication and restoration method is shown in Figure 4.2. Similar to the embedding process, the first 7 bits of the test image are extracted and divided into $8 \times 8$-pixel blocks by applying SLT and then apply the detection algorithm to the first 7 bits. Meanwhile, the LSBs are extracted from the test image and only the LSBs of the detected regions are quantized back for recovery according to authentication result. Consequently, authenticated and recovered images can be output.

The watermark bits can be detected by extracting the watermarked coefficients $y$. If $y$ is larger than 0, the watermark bit value is 1; if $y$ is smaller than 0, the watermark bit value is 0. The retrieved watermark needs to be compared with the original watermark. After the watermark bits from the entire block have been retrieved, the comparison between the watermark bits can be accomplished by using the correlation coefficient $\rho$, computed as follows:

$$
\rho = \frac{\sum \sum (w' - \bar{w'}) \cdot (w - \bar{w})}{\sqrt{\sum \sum (w' - \bar{w'})^2 \sum \sum (w - \bar{w})^2}} \quad (4.4)
$$

where $w$ is the original and $w'$ is the retrieved watermarks corresponding to the block. The correlation coefficient $\rho$ can be compared with a pre-determined threshold value $\lambda = 0.5$, which based on empirical observation of 1000 watermarked images. If $\rho < \lambda$, which indicates that the block has been tampered with as part of authentication process.
This is followed by restoration of the tampered regions based on the decompression and extraction of the LSBs for the watermarked image.

![Figure 4.2: Our proposed SLT watermark detection, authentication and restoration process.](image)

### 4.1.4 Experimental Results

A number of experiments have been carried out to evaluate the performance of the proposed SLT watermarking scheme. The proposed watermarking scheme is compared with two other watermarking schemes: the PST-based [13] and the DCT-based [114]. For a fair comparison, the embedding strength of the watermark in each scheme is adjusted such that the peak signal-to-noise ratio (PSNR) of the watermarked images is approximately 33 dB, which is considered acceptable subjectively. The performance of the watermarking schemes is measured in terms of the false positive rate \( P_{FP} \), the false negative rate \( P_{FN} \) and the average detection rate \( P_{avg} \), defined as:

\[
P_{FP} = \frac{\text{Number of pixels in the untampered region detected as tampered}}{\text{Total number of pixels in the untampered region}}
\]

\[
P_{FN} = \frac{\text{Number of pixels in the tampered region detected as untampered}}{\text{Total number of pixels in the tampered region}}
\]

\[
P_{avg} = \left( 1 - \frac{P_{FP} + P_{FN}}{1 + N_a} \right) \times 100.
\]

where \( N_a \) is the number of area(s) have been tampered with. A number of standard test images are used in the experiments and the results for 6 images, each of size 512 × 512 are obtained.

**JPEG Compression Attack**

Table 4.1 shows that SLT, DCT and PST are compared by applying JPEG compression
Chapter 4. Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT

attack to 6 different grayscale images (512 × 512) in order to determine the $P_F$. As can be seen from the table below, SLT, DCT and PST have similar false positive rates when $QF = 85$. After experiencing 75% JPEG compression attack, the $P_{FPR}$ of SLT is still considerably low with average rate of 1.2%, whereas PST and DCT have the higher average false positive rates of 86% and 31.4%, respectively. Although the $P_{FPR}$ of all three methods have increased when $QF = 65$, SLT still has the lowest increased rate of 30.8% comparing with the average value of $P_{FPR}$ of PST and DCT, of 92.2% and 88.2% respectively. The reason for the relatively better results using the Slant Transform was that the embedding locations concentrated mainly on the middle frequency band, which is considered to be more robust, whereas DCT and PST mainly concentrated more on high frequencies. Overall, the results indicate that the SLT watermarking method achieves lower detection errors than PST and DCT based on the JPEG compression attack.

**Table 4.1:** Comparative performance in term of $P_{FPR}$ (%) the watermarking schemes against JPEG compression with varying quality factor

<table>
<thead>
<tr>
<th>Test Image</th>
<th>$QF = 85$</th>
<th>$QF = 75$</th>
<th>$QF = 65$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SLT</td>
<td>PST</td>
<td>DCT</td>
</tr>
<tr>
<td>Lena</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Baboon</td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.4</td>
<td>1.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Trucks</td>
<td>0.2</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>Ship</td>
<td>0.2</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>San Diego</td>
<td>0.0</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.2</td>
<td>0.7</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Copy and Paste Attack**

The copy and paste attack is utilized to compare the performance in term of detection rates of SLT, DCT and PST for six grayscale test images (512 × 512) as given in Table 4.2. Three different tampering rates of 10%, 20% and 30% will be applied to each test image to analyse the overall detection rate of the three transform methods. The tamper tests are performed with 100 random locations on each image. Consequently, 5400 test images are obtained based on this experimental setup. Table 4.2 shows the comparative performance of the three watermarking schemes against copy and paste attack with different amount of tampering. However, the results show that PST is the most sensitive method as it has the highest overall detection rate after experiencing all three tamper tests (10%, 20% and 30%) for all images. Figure 4.3(a-c), shows the original, watermarked, tampered, authenticated and restored images for the image **Trucks**, respectively.
Table 4.2: Comparative performance in term of $P_{avg}$ (%) the watermarking schemes against copy and paste attack

<table>
<thead>
<tr>
<th>Test Image</th>
<th>10% tamper</th>
<th></th>
<th>20% tamper</th>
<th></th>
<th>30% tamper</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SLT</td>
<td>PST</td>
<td>DCT</td>
<td>SLT</td>
<td>PST</td>
<td>DCT</td>
</tr>
<tr>
<td>Lena</td>
<td>96.0</td>
<td>97.6</td>
<td>95.5</td>
<td>97.9</td>
<td>98.7</td>
<td>97.1</td>
</tr>
<tr>
<td>Baboon</td>
<td>96.7</td>
<td>97.3</td>
<td>96.3</td>
<td>98.1</td>
<td>98.8</td>
<td>96.5</td>
</tr>
<tr>
<td>Bridge</td>
<td>96.3</td>
<td>97.5</td>
<td>95.4</td>
<td>97.9</td>
<td>98.6</td>
<td>97.0</td>
</tr>
<tr>
<td>Trucks</td>
<td>95.7</td>
<td>97.6</td>
<td>95.1</td>
<td>97.6</td>
<td>98.7</td>
<td>96.7</td>
</tr>
<tr>
<td>Ship</td>
<td>96.1</td>
<td>97.6</td>
<td>95.2</td>
<td>97.7</td>
<td>98.8</td>
<td>96.2</td>
</tr>
<tr>
<td>San Diego</td>
<td>96.6</td>
<td>97.6</td>
<td>95.4</td>
<td>98.1</td>
<td>98.8</td>
<td>96.6</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>96.2</td>
<td>97.5</td>
<td>95.5</td>
<td>97.9</td>
<td>98.7</td>
<td>96.7</td>
</tr>
</tbody>
</table>

Figure 4.3: Demonstration of the image Trucks in SLT semi-fragile watermarking scheme

**Combined JPEG Compression and Copy and Paste Attack**

In Table 4.3, the six watermarked images (512×512) are compressed with three different JPEG compression rates QF of 85, 75 and 65. The experimental setup is similar to the previous copy and paste attack with 100 random locations for tampered areas. Overall, the PST achieves a relatively higher detection rate than DCT and SLT after experiencing QF of 85. However, for detection, it is as worse as at 54.1% with $QF = 75$. In comparison, SLT has the highest $P_{avg}$ of 91.9% at $QF = 75$ and 65. From the analysis, SLT achieves a more accurate detection result than PST and DCT. As a whole, the result indicates that the best $P_{avg}$ among the three methods is SLT, which has 91.9% detection rate with $QF = 75$. However, all the attacked images could not be recovered by any of the three transform schemes after applying JPEG compression attack. This is due to
the fact that the restoration technique is based on LSB embedding in the spatial domain of the watermarked image which is fragile. As such, it can be easily removed by JPEG compression.

<table>
<thead>
<tr>
<th>Test Image</th>
<th>QF = 85</th>
<th>QF = 75</th>
<th>QF = 65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>92.1</td>
<td>94.9</td>
<td>91.6</td>
</tr>
<tr>
<td>Baboon</td>
<td>92.1</td>
<td>94.7</td>
<td>92.1</td>
</tr>
<tr>
<td>Bridge</td>
<td>91.9</td>
<td>94.0</td>
<td>91.6</td>
</tr>
<tr>
<td>Trucks</td>
<td>92.0</td>
<td>94.5</td>
<td>91.2</td>
</tr>
<tr>
<td>Ship</td>
<td>91.5</td>
<td>93.7</td>
<td>91.9</td>
</tr>
<tr>
<td>San Diego</td>
<td>93.1</td>
<td>94.5</td>
<td>92.2</td>
</tr>
<tr>
<td>Average</td>
<td>92.1</td>
<td>94.4</td>
<td>91.8</td>
</tr>
</tbody>
</table>

4.1.5 Further Experimental Results

According to the results obtained in Section 4.1.4, we found that our SLT based scheme could be further compared with the DCT and PST based semi-fragile watermarking schemes using the following additional test criteria:

- The LSB image restoration method has been excluded when compared with the PST and DCT, as it will fail to work if any JPEG Compression has been applied to the test image.

- The $P_{FPR}$ and $P_{FNR}$ have been calculated separately instead of $P_{avg}$.

- In addition to six common test images, three simulated law enforcement images have also been included for analysis. These simulated images are ‘Gun’, ‘Car1’ and ‘Car2’, as shown in Figure 4.4.

- For cross-comparison purposes across the difference schemes, eight watermarks (with different watermark sequences) are randomly embedded in the mid-frequency coefficients each $8 \times 8$ block in the respective transform domain. In Figure 4.5, the shaded areas are considered as random embedding locations for all the SLT, PST and DCT based schemes.

- Additive Gaussian noise has been added into test images for further comparison as it is also considered to be a common and unintentional operation.
Similar to the first set of experiments, we compare the performance of the watermarking schemes against the copy and paste attack. In the attack, a number of blocks within the watermarked image are replaced with blocks randomly selected from the same test image. To achieve a better statistical confidence level, the experiment is repeated 100 times using different watermark keys and tampered blocks, and the average results are then calculated. Figure 4.6, shows the watermarked, tampered, and authenticated images for image ‘Carl’.

The performance of the watermarking schemes is also measured in terms of the false positive rate ($P_{FPR}$), and false negative rate ($P_{FNR}$). Table 4.4 compares the performance of the watermarking schemes against the copy and paste attack, where 20% of the blocks have been tampered. It can be observed that the three watermarking schemes perform similarly against the copy and paste attack, with the $P_{FNR}$ of SLT performing approximately lower than the others, by 0.06% approximately. For $P_F$, the performance of SLT and PST are same at 0%, whereas DCT is 0.15% higher than others on average.

The performance of the watermarking schemes against the copy and paste attack is also compared against the presence of JPEG compression and additive Gaussian noise. The
results are presented in Tables 4.5-4.8. It should be noted that a slightly moderate JPEG compression and additive Gaussian noise are considered to be legitimate manipulations and commonly considered as part of the system processing. Hence, the semi-fragile watermarking schemes are expected to be robust against these manipulations.

For JPEG compression with quality factor $QF = 85$ in Table 4.5, the $P_{FPR}$ of SLT and PST based watermarking schemes perform similarly at 0.03% and 0.01%, in comparison with the results obtained when no JPEG compression is applied. In contrast to the DCT based scheme, the $P_{FPR}$ is ten time higher than the SLT based scheme, at 0.3% on average. Especially, the three law enforcement images are highest at approximately 0.6% on average based on the DCT scheme. The $P_{FNR}$ of all watermarking schemes achieved on average at 10% on average, which is also similar to the results obtained when no JPEG compression is applied. Hence, these results demonstrate that our semi-fragile watermarking scheme could detect the tampered areas correctly with acceptable errors, and is robust to JPEG compression non-malicious manipulation.
However, with increased compression (QF=75) as shown in Table 4.6, there is a clear difference in the performance of the watermarking schemes. From the results, it can be seen that the $P_{FPR}$ of PST and the DCT based schemes are at approximately 0.19% and 1.5%, respectively, which performed much better than the proposed SLT based scheme, approximately at 16.22%. The $P_{FNR}$ of SLT is also gradually increased by 0.31%, compared to QF=85. Moreover, there is a difference in the performance of the proposed watermarking schemes for different images on $P_{FPR}$. For images with high texture (Baboon and San-Diego), then $P_{FPR}$ is less than 10%, which is better in comparison to other images with less texture (Lena and Gun) with exceeded 20%.

For additive Gaussian noise, as shown in Tables 4.7 and 4.8, the $P_{FPR}$ of our proposed SLT based scheme increases on average from 8.59% to 13.35% with noise variance increasing from 0.003 to 0.005. Our proposed SLT based scheme also achieves much lower false positive rates than the DCT and PST based schemes, at 11.01% and 15.71% for noise variance is at 0.003, and 18.43% and 23.88% with noise variance is 0.005, respectively. The three law enforcement images for our SLT based scheme archived the lowest false positive rates at 5.75% 5.16% and 6.35%, with noise variance at 0.003 as shown in Table 4.7. Therefore, it is clear from the results obtained that the proposed SLT based scheme provides an improvement of robustness to additive Gaussian noise as compared to other two schemes.

**Table 4.5: New comparative performance of the watermarking schemes against copy and paste attack (20%) and JPEG compression (QF=85).**

<table>
<thead>
<tr>
<th>Test Image</th>
<th>SLT</th>
<th>PST</th>
<th>DCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_{FPR}$</td>
<td>$P_{FNR}$</td>
<td>$P_{FPR}$</td>
</tr>
<tr>
<td>Lena</td>
<td>0</td>
<td>8.58</td>
<td>0</td>
</tr>
<tr>
<td>Baboon</td>
<td>0.01</td>
<td>9.63</td>
<td>0</td>
</tr>
<tr>
<td>Ship</td>
<td>0</td>
<td>9.47</td>
<td>0</td>
</tr>
<tr>
<td>Trucks</td>
<td>0</td>
<td>9.89</td>
<td>0</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.05</td>
<td>9.56</td>
<td>0.03</td>
</tr>
<tr>
<td>San Diego</td>
<td>0</td>
<td>9.55</td>
<td>0</td>
</tr>
<tr>
<td>Gun</td>
<td>0.03</td>
<td>9.1</td>
<td>0.01</td>
</tr>
<tr>
<td>Car1</td>
<td>0.01</td>
<td>9.78</td>
<td>0</td>
</tr>
<tr>
<td>Car2</td>
<td>0.17</td>
<td>9.11</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.03</td>
<td>9.41</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Chapter 4. Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT 56

4.2 Novel WBCT Based Semi-fragile Watermarking Algorithm

In this section, we discuss our proposed WBCT semi-fragile watermarking algorithm by analysing the patterns of the parent and children relationships and how they can be affected by the different manipulations in the WBCT domain. We also discuss the details of the watermark embedding, detection, and authentication processes. Experimental results are also presented in this section.

Table 4.6: New comparative performance of the watermarking schemes against copy and paste attack (20%) and JPEG compression (QF=75).

<table>
<thead>
<tr>
<th>Test Image</th>
<th>SLT</th>
<th>PST</th>
<th>DCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_{FPR}$</td>
<td>$P_{FNR}$</td>
<td>$P_{FPR}$</td>
</tr>
<tr>
<td>Lena</td>
<td>21.07</td>
<td>9.58</td>
<td>0</td>
</tr>
<tr>
<td>Baboon</td>
<td>9.69</td>
<td>9.84</td>
<td>0.01</td>
</tr>
<tr>
<td>Ship</td>
<td>20.21</td>
<td>9.32</td>
<td>0.01</td>
</tr>
<tr>
<td>Trucks</td>
<td>15.04</td>
<td>9.22</td>
<td>0</td>
</tr>
<tr>
<td>Bridge</td>
<td>12.54</td>
<td>9.8</td>
<td>0.32</td>
</tr>
<tr>
<td>San Diego</td>
<td>8.94</td>
<td>9.99</td>
<td>0</td>
</tr>
<tr>
<td>Gun</td>
<td>24.84</td>
<td>9.85</td>
<td>0.26</td>
</tr>
<tr>
<td>Car1</td>
<td>15.89</td>
<td>10.44</td>
<td>0.03</td>
</tr>
<tr>
<td>Car2</td>
<td>17.79</td>
<td>9.48</td>
<td>1.11</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>16.22</td>
<td>9.72</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table 4.7: New comparative performance of the watermarking schemes against copy and paste attack (20%) and Additive Gaussian noise (variance=0.003).

<table>
<thead>
<tr>
<th>Test Image</th>
<th>SLT</th>
<th>PST</th>
<th>DCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_{FPR}$</td>
<td>$P_{FNR}$</td>
<td>$P_{FPR}$</td>
</tr>
<tr>
<td>Lena</td>
<td>10.21</td>
<td>9.48</td>
<td>11.72</td>
</tr>
<tr>
<td>Baboon</td>
<td>10.39</td>
<td>9.4</td>
<td>11.47</td>
</tr>
<tr>
<td>Trucks</td>
<td>10.53</td>
<td>8.79</td>
<td>12.17</td>
</tr>
<tr>
<td>Bridge</td>
<td>10.19</td>
<td>10.46</td>
<td>11.6</td>
</tr>
<tr>
<td>San Diego</td>
<td>10.26</td>
<td>9.93</td>
<td>11.45</td>
</tr>
<tr>
<td>Gun</td>
<td>5.75</td>
<td>9.24</td>
<td>9.64</td>
</tr>
<tr>
<td>Car1</td>
<td>5.16</td>
<td>10.17</td>
<td>8.97</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>8.49</td>
<td>9.61</td>
<td>11.01</td>
</tr>
</tbody>
</table>
Table 4.8: New comparative performance of the watermarking schemes against copy and paste attack (20%) and Additive Gaussian noise (variance=0.005).

<table>
<thead>
<tr>
<th>Test Image</th>
<th>SLT</th>
<th>PST</th>
<th>DCT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_{FPR}$</td>
<td>$P_{FNR}$</td>
<td>$P_{FPR}$</td>
</tr>
<tr>
<td>Lena</td>
<td>13.52</td>
<td>8.83</td>
<td>16.94</td>
</tr>
<tr>
<td>Baboon</td>
<td>13.49</td>
<td>9.52</td>
<td>15.80</td>
</tr>
<tr>
<td>Trucks</td>
<td>13.46</td>
<td>10.18</td>
<td>16.87</td>
</tr>
<tr>
<td>Bridge</td>
<td>13.59</td>
<td>10.22</td>
<td>16.90</td>
</tr>
<tr>
<td>San Diego</td>
<td>13.50</td>
<td>10.26</td>
<td>16.64</td>
</tr>
<tr>
<td>Car2</td>
<td>17.04</td>
<td>9.28</td>
<td>22.3</td>
</tr>
<tr>
<td>Average</td>
<td>13.35</td>
<td>9.71</td>
<td>18.43</td>
</tr>
</tbody>
</table>

4.2.1 Parent and Children Coefficient Relations before and after Non-malicious Manipulations

In Section 3.2, we investigated the characteristics of the energy relations of the original images between the parent and the children coefficients before and after JPEG compression. The results showed that these invariant relations still maintained above 75% when the QF=10. A three level WBCT was applied to the test images, the robustness characteristics of the parent and children relationship were then utilised for our proposed robust watermarking scheme. In this section, we further analyse the invariant relations of WBCT for different non-malicious manipulations to determine if it is also be feasible for semi-fragile watermarking. Note that we analyse these parent and children invariant relationships after two level WBCT applied, due to more watermarks are required for our proposed WBCT-based semi-fragile watermarking. For our experiment, we analyse the following non-malicious manipulations:

- Mild compression, JPEG and JPEG 2000, up to 50%
- $3 \times 3$ Gaussian filtering
- Additive Gaussian noise (PSNR above 35db)
- Contrast stretching (1%)

The parent and children coefficient relations before and after JPEG compression are first analysed in detail. The original image and its JPEG compressed images are initially decomposed by a two level WBCT. The average percentage of invariant energy relations between parent and children coefficients in the non-redundant contourlet domain before
and after JPEG compression is then calculated, similar to the discussion that mentioned in Section 3.2.

Figure 4.7 illustrates the average of the percentages of the numbers of invariant relations in 12 different subbands, as shown in Figure 3.3(b), by using six images. This figure shows the percentages of the numbers of invariant relations to the total numbers of relations against the quality factors (QF) of JPEG compression. The average percentage of invariant relations is defined as follows:

$$P_{average} = \frac{\sum_{i=1}^{12} P_i}{12}$$  \hspace{1cm} (4.8)

where $P_i (i = 1, 2, \cdots, 12)$ represent the percentages of the numbers of invariant relations to the total numbers of relations in $i$ subbands. Further analysis are performed for JPEG2000, Gaussian noise, Gaussian filtering and contrast stretching. The JPEG2000 results are shown in Figure 4.8 for QF=90 to QF=50. Different Gaussian noise, Gaussian filtering and contrast stretching results are summarised in Table 4.9. Six standard test images are used in this experiment to determine the invariant energy relationship. For JPEG and JPEG2000 compression, we observe from Figures 4.7 and 4.8 that the percentages of invariant relations gradually decrease with increasing compression rates. For example, in Figure 4.7, for QF=90, the highest invariant relation exceeds 92% while the lowest achieves approximately 85%. For JPEG2000, all the invariant relations are above 98% as shown in Figure 4.8. Even though all the test images have been compressed significantly at QF=50, the average invariant relations remained between 72% and 86% as shown in Figure 4.7.

![Figure 4.7: The average percentages of invariant relations after JPEG compression](image-url)
Overall, the JPEG2000 compression results achieve better performance as compared to JPEG compression results. The reason for this is due to both JPEG2000 and WBCT are wavelet-based methods. In Table 4.9, the results for contrast stretching also demonstrate an excellent invariance with the results averaging 98%, followed by Gaussian filtering averaging 90%. Of all non-malicious manipulations, Gaussian noise achieves the lowest performance at approximately 79% of invariant relation. From Figures 4.7, 4.8 and Table 4.9, we observe that highly textured images, such as San Diego can achieve relatively better results than the other five images. Overall, for all images after JPEG/JPEG2000 at QF=90 to QF=50, Gaussian noise (PSNR above 35db), Gaussian filtering and contrast stretching (1%), over 70% invariant relations can be achieved by exploiting and adopting the modulation of their energy relationship.

Table 4.9: The average percentages of invariant relations after three types manipulations

<table>
<thead>
<tr>
<th>Test image</th>
<th>Gaussian noise</th>
<th>Gaussian filtering</th>
<th>Contrast stretching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>72.26</td>
<td>90.06</td>
<td>97.18</td>
</tr>
<tr>
<td>Boats</td>
<td>78.50</td>
<td>89.84</td>
<td>97.56</td>
</tr>
<tr>
<td>Trucks</td>
<td>82.98</td>
<td>90.56</td>
<td>98.89</td>
</tr>
<tr>
<td>San Diego</td>
<td>88.56</td>
<td>90.21</td>
<td>99.24</td>
</tr>
<tr>
<td>Peppers</td>
<td>75.03</td>
<td>87.99</td>
<td>97.79</td>
</tr>
<tr>
<td>Goldhill</td>
<td>77.48</td>
<td>90.51</td>
<td>96.53</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>79.14</strong></td>
<td><strong>89.86</strong></td>
<td><strong>97.87</strong></td>
</tr>
</tbody>
</table>
4.2.2 Coefficient Differences Analysis

In this section, we analyse the patterns of the parent and children relationships and how they can be affected by the different manipulations in the WBCT domain. The 'Goldhill', as one of commonly used test image, is used for our detailed analysis. In Figure 4.9, six scatter plots illustrate the differences between the absolute value of parent-coefficients and the averages of its children-coefficients before manipulations and the differences after manipulations. The diagonal line represents the $y=x$ axis. From Figure 4.9(a) to (e), each scatter plot illustrates the difference before and after non-malicious manipulations with cluster points varying along the diagonal line. In Figure 4.9(f), the differences after malicious manipulation (copy and paste) are changed significantly and clusters are formed and spread around the origin.

From our analysis, we conclude that the characteristic of parent-children relationship is robust against a significant amount of non-malicious manipulations and still fragile to content modification. This demonstrates that the parent-children relationship is indeed feasible for incorporation into our proposed semi-fragile watermarking algorithm.

4.2.3 WBCT Watermark Embedding Process

From our experiments described above, we can summarise that any alterations of the image can be detected from any of the 12 subbands. Since different subbands represent different directional information, we can reasonably expect that embedding a watermark into four different subbands can achieve a good trade-off between image quality and false negative rate. The detail of selecting four subbands will be explained in Section 4.2.6.

In our proposed method, the size of the cover image is $512 \times 512$ and the watermark is $128 \times 128$. The watermark is a pseudo-random binary $(1, 0)$ sequence. The block diagram of the proposed embedding process is shown in Figure 4.10. To begin with, the original image is decomposed into 12 sets of parent and children coefficients by applying WBCT. Afterwards, the parent-children relationships of four subbands are extracted. According to these relationships, the watermark bits, which are encrypted a random binary sequence obtained by a key, are embedded by modulating the corresponding parent coefficients, as follows:
Figure 4.9: Scatter plots showing the distributions of the coefficient differences before and after manipulations of image Goldhill (a) JPEG (QF=70) (b) JPEG2000 (QF=70) (c) Gaussian noise (\(\sigma=0.0003\)) (d) Gaussian filtering (e) Contrast stretching (f) copy and paste (64 x 128)
Chapter 4. Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT

62

\[ P' = \begin{cases} 
P, & ((|P| - |C_{avg}|) \geq T) \land (w = 1)) \\
P, & ((|P| - |C_{avg}|) < T) \land (w = 0)) \\
P + |C_{avg}| + T - P) \times K_1, & ((|P| - |C_{avg}|) < T) \land (P \geq 0) \land (w = 1)) \\
P - |C_{avg}| + T - |P|) \times K_1, & ((|P| - |C_{avg}|) < T) \land (P < 0) \land (w = 1)) \\
P + (P - |C_{avg}| - T) \times K_2, & ((|P| - |C_{avg}|) < T) \land (P \geq 0) \land (w = 0)) \\
P - (P - |C_{avg}| - T) \times K_2, & ((|P| - |C_{avg}|) < T) \land (P < 0) \land (w = 0)) 
\] (4.9)

where \( P \) is denoted as a parent coefficient in the image, \( C_{avg} \) is the average of four children coefficients, and \( w \) is the watermark bit. The threshold \( T \) controls the perceptual quality and robustness of the watermarked image, where \( T = 10 \). The parameters \( K_1 \) and \( K_2 \) are both constants. Finally, the watermarked image is reconstructed by applying the inverse WBCT transform.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure410.png}
\caption{WBCT semi-fragile watermark embedding process}
\end{figure}

4.2.4 WBCT Watermark Detection Process

The detection and authentication process is shown in Figure 4.11. The WBCT is first performed on the test image, which is decomposed into 12 sets of parent and children coefficients. A key is used for extracting four subbands of the parent-children relationships from the 12 sets. The watermark bits \( w' \) are then extracted from these relationships, using the following detection algorithm:

\[ w' = \begin{cases} 
1, & ((|P'| - |C'_{avg}|) \geq (T + M)) \\
0, & ((|P'| - |C'_{avg}|) \leq (T - M)) \\
-1, & ((T - M) < (|P'| - |C'_{avg}|) < (T + M)) 
\end{cases} \] (4.10)
where \( w' = -1 \) represents the area that has not been tampered with, \( T = 10 \) and \( M \) is an error tolerance margin value to decrease the false positive rates caused by non-malicious manipulations. However, higher values of \( M \) result in increasing false negative rates, while lower values of \( M \) result in increasing false positive rate. Thus, a value for \( M \) to adjust the trade-off between the false positive and false negative rates needs to be determined.

![Diagram](image)

**Figure 4.11:** WBCT semi-fragile watermark detection and image authentication process

Figures 4.12-4.14 show the histograms of differences before and after non-malicious manipulations in the first embedding subband of the image ‘Goldhill’. Note that in order to illustrate and compare the histograms clearly, Figures 4.12-4.14 have been scaled. The differences are defined as the original \( |P| - |C_{\text{avg}}| \) minus the manipulated \( |P'| - |C'_{\text{avg}}| \). These distributions exhibit a sharp peak at zero amplitude and tail off rapidly on both sides of the peak. This implies that the differences distribute sparsely, as the majority of differences are close to zero, which further prove that most of the parent-children relationships before and after non-malicious manipulations maintain their invariance. Similar distributions are also observed from other test images during the experiments. We found that the differences between the distributions vary from 5 to -5. Therefore, we set the error tolerance margin value \( M = 5 \) in order to minimise the errors.

![Histograms](image)

**Figure 4.12:** The histogram of differences before and after JPEG compression (a) QF=90 (b) QF=70 (c) QF=50
Chapter 4. Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT

4.2.5 WBCT Watermarked Image Authentication Process

For the authentication process as shown in Figure 4.11, a difference image is obtained by comparing the original watermark with the extracted watermark. The authentication algorithm is shown as follows:

\[
\text{diff} = \begin{cases} 
1, & (w \neq w') \\
0, & (w = w' \vee w' = -1)
\end{cases}
\]  \hspace{1cm} (4.11)

This difference image is used for locating the tampered regions. The difference image is divided into four parts, and each part represents the difference image of each subband. In order to obtain more directional information, the four parts are fused into one difference image through XOR operation, as pseudo-code below:

\begin{verbatim}
if Part1 = Part2 = Part3 = Part4 = 0 then
    Difference = 0
else
    Difference = 1
end if
\end{verbatim}
Examples for the fusion results are shown in Figure 4.15. The white spots represent the detected tampered region from four subbands which are then fused into one difference image. It can be clearly seen from the fused image that the white spots are now much more prominent.

Finally, for the authenticated image, we apply the morphological operators to improve the detection performance. Most of the false positive errors distribute sparsely as a result of artefacts caused by signal processing operations such as JPEG compression, whereas false negative errors from the copy and paste attack distribute more densely relatively. Morphological operators are commonly used as a nonlinear technique in image processing [150] to reduce false positive and false negative rates. Erosion (denoted as \( A \ominus B \)) and dilation (denoted as \( A \oplus B \)) are two fundamental morphological operators. For many applications, the two operators are commonly combined to form opening and closing transformation. The opening of \( A \) by \( B \), denoted \( A \circ B \), is defined as \( A \circ B = (A \ominus B) \oplus B \). The closing of \( A \) by \( B \), denoted \( A \bullet B \), is defined as \( A \bullet B = (A \oplus B) \ominus B \).

In our proposed method, we first use the open and then close operations that could first decrease false negative rate and then decrease the false positive rate for the authenticated image. Experiments showing the improvements of detection performance are presented in Section 4.2.6.

### 4.2.6 Experimental Results

The PSNR is also used to evaluate the perceptual distortion of these test images before and after watermark embedding and the results are illustrated in Figures 4.16. The x-axis represents the number of subbands randomly selected for embedding the watermark bits. Figure 4.16 shows the PSNR values decrease with increasing number of subbands embedded. However, there is some slight fluctuation and the reason for that could be due to different subbands having different influence on the image quality. The PSNR
may decrease if a more effective subband has been randomly selected. We observe that by using four subbands embedded an acceptable image quality above 30dB is achieved. Figure 4.17 shows that the $P_{FNR}$ decrease with increasing number of subbands embedded when copy and paste attack with 64×128 pixels applied. Using four subbands also obtain lower $P_{FNR}$ than using only one, two or three subbands. Therefore, we decide to embed watermark bits into four random subbands for our algorithm as a trade-off between imperceptibility and $P_{FNR}$.

**FIGURE 4.16:** PSNR with different number of subband embedded randomly

**FIGURE 4.17:** $P_{FNR}$ with different number of subband embedded randomly

To evaluate the performance of our semi-fragile watermarking scheme, similar to Section 4.1.5, nine test images of size 512×512 are also used for our experiments. These images
include common test images such as Lena, Boats, Trucks, San Diego, Peppers, and Goldhill, as well as three simulated law enforcement images, Gun, Car1 and Car2 (as shown in Figure 4.4). The PSNR of these watermarked images is approximately 33 dB. Figure 4.18(a)-(c), shows the watermarked, tampered and authenticated images for the ‘Car2’ photograph, respectively.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{car2_images.png}
\caption{The watermarked, tampered, and authenticated images for image ‘Car2’}
\end{figure}

In order to analyse the false positive and false negative rates, we investigate the following manipulations:

- JPEG compression only from QF=90 to 50 (Figure 4.19)
- JPEG2000 compression only from QF=90 to 50 (Figure 4.20)
- 3 x 3 Gaussian filtering only (Table 4.10)
- Additive Gaussian noise (PSNR above 35db) only (Table 4.10)
- Contrast stretch (1%) only (Table 4.10)
- JPEG compression QF=90, 70, 50 with copy and paste attack (Table 4.11)
- JPEG2000 compression QF=90, 70, 50 with copy and paste attack (Table 4.12)
- 3 x 3 Gaussian filtering with copy and paste attack (Table 4.13)
- Additive Gaussian noise (PSNR above 35db) with copy and paste attack (Table 4.13)
- Contrast stretch (1%) with copy and paste attack (Table 4.13)

Figures 4.19 and 4.20 illustrate the detection performance for JPEG and JPEG2000 at different quality factors of compression. The $P_{FPR}$ increase gradually as the quality
factor decreases. In the case of high compression at QF=50, the $P_{FP_R}$ are relatively low; less than 20% for JPEG compression and 6% for JPEG2000. The results clearly indicate that the detection performance for JPEG2000 compression is much better than JPEG at the same quality factor. The performances of our algorithm against additive Gaussian noise, filtering and contrast stretching are given in Table 4.10. From the results, we can observe that our proposed algorithm is robust against different signal processing operations, which are considered to be non-malicious manipulations.

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|}
\hline
Test image & Gaussian noise & Gaussian filtering & Contrast stretching \\
\hline
Lena & 8.03 & 1.20 & 0.95 \\
Boats & 7.86 & 1.17 & 1.44 \\
Trucks & 4.22 & 0.51 & 3.49 \\
San Diego & 2.69 & 0.90 & 0.98 \\
Peppers & 8.33 & 0.93 & 0.32 \\
Goldhill & 7.28 & 1.22 & 0.27 \\
Gun & 8.40 & 1.47 & 0.88 \\
Car1 & 5.22 & 1.90 & 0.32 \\
Car2 & 6.49 & 1.90 & 0.10 \\
\hline
Average & 6.50 & 1.24 & 0.97 \\
\hline
\end{tabular}
\caption{Performance of false positive rate after Gaussian noise, Gaussian filtering and contrast stretching}
\end{table}

The performance of the proposed watermarking algorithm against the copy and paste
attack with $64 \times 128$ pixels is also compared in the presence of non-malicious manipulations. Tables 4.11 and 4.12 illustrate the watermarked images that are JPEG and JPEG2000 compressed with QF=90, 70 and 50, after copy and paste modifications have been made. The detection performance after copy and paste attacks with additive Gaussian noise, Gaussian filtering and contrast stretching are given in Table 4.13. In Table 4.11, the results indicate that our method can detect the tampered regions accurately. On average, $P_{FN}$ is approximately 1%, while $P_{FP}$ is below 4%. Test images 'Trucks' and 'San Diego' and image 'Car2' indicate a better performance with approximately 2% for both $P_{FP}$ and $P_{FN}$. In terms of JPEG2000 compression with the copy and paste attack, the results given in Table 4.12 indicate better performances than JPEG compression. In Tables 4.11 and 4.12, the two outdoor images, 'Car1' and 'Car2', achieve much better performance than the indoor image 'Gun' in terms of $P_{FP}$. When the false positive rates are below 6%, $P_{FN}$ are approximately 0.5%. In particular, the 'Goldhill' test image performs better when $P_{FN}$ are very close to 0% and the $P_{FP}$ is approximately 4%. The $P_{FP}$ and $P_{FN}$ from Tables 4.11 to 4.13 indicate that our proposed WBCT based semi-fragile watermarking scheme is able to authenticate and localise the tampered regions accurately, as well as being sufficiently robust against some legitimate attacks.

Samples of authenticated images are shown in Figures 4.21 to 4.23. The modification results shown in Figure 4.21(b) have two areas tampered with JPEG compression (QF=70). Figure 4.22(b) shows that four areas have been tampered JPEG compression

![Figure 4.20: Performance of false positive rate after JPEG2000 compression](image-url)
Chapter 4. Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT

(QF=70) and Figure 4.23(b) with one area tampered with contrast stretching. These results also indicate that the artifacts caused by non-malicious manipulations distribute sparsely, which can be removed by morphological operations.

**Table 4.11: Performance after copy and paste with JPEG compression**

<table>
<thead>
<tr>
<th>Test Image</th>
<th>QF90</th>
<th>QF70</th>
<th>QF50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_{FP}$</td>
<td>$P_{FN}$</td>
<td>$P_{FP}$</td>
</tr>
<tr>
<td>Lena</td>
<td>4.47</td>
<td>0.69</td>
<td>3.69</td>
</tr>
<tr>
<td>Boats</td>
<td>5.49</td>
<td>0.61</td>
<td>5.60</td>
</tr>
<tr>
<td>Trucks</td>
<td>1.81</td>
<td>0.73</td>
<td>0.83</td>
</tr>
<tr>
<td>San Diego</td>
<td>2.29</td>
<td>0.89</td>
<td>1.73</td>
</tr>
<tr>
<td>Peppers</td>
<td>4.27</td>
<td>0.08</td>
<td>3.16</td>
</tr>
<tr>
<td>Goldhill</td>
<td>4.63</td>
<td>0.55</td>
<td>6.29</td>
</tr>
<tr>
<td>Gun</td>
<td>6.11</td>
<td>0.36</td>
<td>4.35</td>
</tr>
<tr>
<td>Car1</td>
<td>2.48</td>
<td>0.72</td>
<td>1.25</td>
</tr>
<tr>
<td>Car2</td>
<td>1.23</td>
<td>1.17</td>
<td>6.68</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>3.64</td>
<td>0.64</td>
<td>3.73</td>
</tr>
</tbody>
</table>

**Table 4.12: Performance after copy and paste with JPEG000 compression**

<table>
<thead>
<tr>
<th>Test Image</th>
<th>QF90</th>
<th>QF70</th>
<th>QF50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_{FP}$</td>
<td>$P_{FN}$</td>
<td>$P_{FP}$</td>
</tr>
<tr>
<td>Lena</td>
<td>3.49</td>
<td>0.61</td>
<td>5.05</td>
</tr>
<tr>
<td>Boats</td>
<td>5.40</td>
<td>0.53</td>
<td>5.49</td>
</tr>
<tr>
<td>Trucks</td>
<td>1.81</td>
<td>0.64</td>
<td>1.81</td>
</tr>
<tr>
<td>San Diego</td>
<td>3.05</td>
<td>0.53</td>
<td>2.03</td>
</tr>
<tr>
<td>Peppers</td>
<td>4.64</td>
<td>0</td>
<td>8.74</td>
</tr>
<tr>
<td>Goldhill</td>
<td>4.63</td>
<td>0.55</td>
<td>4.92</td>
</tr>
<tr>
<td>Gun</td>
<td>6.20</td>
<td>0.34</td>
<td>7.08</td>
</tr>
<tr>
<td>Car1</td>
<td>2.48</td>
<td>0.60</td>
<td>2.48</td>
</tr>
<tr>
<td>Car2</td>
<td>1.33</td>
<td>0.96</td>
<td>1.27</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>3.67</td>
<td>0.53</td>
<td>4.32</td>
</tr>
</tbody>
</table>

4.3 Comparison Between SLT and WBCT Based Schemes

In Sections 4.1 and 4.2, we discussed our proposed semi-fragile watermarking schemes based on SLT and WBCT image processing techniques. In this section, we will analyse and compare the difference between these two schemes, as summarised in Table 4.14.

In Table 4.14, we can see that in SLT-based scheme, the image is divided into $8 \times 8$-pixel blocks, then the watermarks are embedded into mid-frequency of each block that is applied with DCT. On the other hand, in WBCT-based scheme, the watermarks are randomly embedded into HL2, LH2 and HH2 sub-bands after applying 2 level of WBCT to
Table 4.13: Performance after copy and paste with three signal processing

<table>
<thead>
<tr>
<th>Test Image</th>
<th>Gaussian noise</th>
<th>Gaussian filtering</th>
<th>Contrast stretching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$P_{FPR}$</td>
<td>$P_{FNR}$</td>
<td>$P_{FPR}$</td>
</tr>
<tr>
<td>Lena</td>
<td>3.83</td>
<td>0.36</td>
<td>3.37</td>
</tr>
<tr>
<td>Boats</td>
<td>9.5</td>
<td>0.49</td>
<td>6.37</td>
</tr>
<tr>
<td>Trucks</td>
<td>7.15</td>
<td>0.31</td>
<td>11.06</td>
</tr>
<tr>
<td>San Diego</td>
<td>4.5</td>
<td>0.70</td>
<td>4.91</td>
</tr>
<tr>
<td>Peppers</td>
<td>1.42</td>
<td>0.87</td>
<td>5.30</td>
</tr>
<tr>
<td>Goldhill</td>
<td>3.63</td>
<td>0.25</td>
<td>4.27</td>
</tr>
<tr>
<td>Gun</td>
<td>2.39</td>
<td>1.23</td>
<td>2.49</td>
</tr>
<tr>
<td>Car1</td>
<td>7.52</td>
<td>0.40</td>
<td>6.63</td>
</tr>
<tr>
<td>Car2</td>
<td>4.49</td>
<td>1.20</td>
<td>1.03</td>
</tr>
<tr>
<td>Average</td>
<td>4.94</td>
<td>0.65</td>
<td>5.05</td>
</tr>
</tbody>
</table>

Table 4.14: Comparison Between SLT and WBCT Semi-fragile

<table>
<thead>
<tr>
<th></th>
<th>SLT</th>
<th>WBCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Embedding</td>
<td>Yes, 8 x 8</td>
<td>No, 2 level of WBCT</td>
</tr>
<tr>
<td>Content Adaptive</td>
<td>No</td>
<td>Yes, parent-children relationships</td>
</tr>
<tr>
<td>Size of Image</td>
<td>512 x 512</td>
<td>512 x 512</td>
</tr>
<tr>
<td>No. of Watermarks</td>
<td>32768</td>
<td>16384</td>
</tr>
<tr>
<td>PSNR</td>
<td>Around 33dB</td>
<td>Around 33dB</td>
</tr>
<tr>
<td>Block Authentication</td>
<td>Yes, block independent</td>
<td>No</td>
</tr>
<tr>
<td>Error Tolerance Margin</td>
<td>$\tau = 0.5$</td>
<td>$M = 5$</td>
</tr>
<tr>
<td>Reduce Errors Process</td>
<td>No</td>
<td>Yes, morphological operators</td>
</tr>
<tr>
<td>Experimental Results</td>
<td>Reasonable</td>
<td>Good</td>
</tr>
</tbody>
</table>

the image. As a result, comparing to SLT-based scheme, the WBCT-based watermarked image could avoid the issue of blockness distortion, which is the commonly drawback issue for block-based watermarking schemes. Moreover, the WBCT domain scheme is more preferable for use with a wide range of images, as a result of the unique parent and child relationship for each image. This characteristic of parent-child relationships can be utilised for semi-fragile watermark embedding, extraction and authentication processes, and is adaptive for a wide range of images, each with varying details. Both schemes tested with 512 x 512 images and obtained acceptable PSNR values approximately 33dB, in spite of embedding 32768 (8 bits each block, 4096 blocks) watermarks for SLT-based scheme and 16384 (128 x 128) watermarks for WBCT-based scheme. In addition, they also used error tolerance margin, that $\tau = 0.5$ for SLT-based and $M = 5$ for WBCT-based, which could tolerated some non-malicious manipulations, particularly, JPEG compression. However, the experimental results have been further improved with the help of morphological operators in WBCT-based scheme.
Chapter 4. Novel Semi-fragile Watermarking Algorithms Based on SLT and WBCT

4.4 Summary

In this chapter, we have discussed two proposed semi-fragile watermarking schemes, namely, SLT-based and WBCT-based. For the SLT semi-fragile watermarking scheme, the watermark was embedded into the middle frequency SLT coefficients of non-overlapping blocks of the test images. The performance of the SLT based semi-fragile scheme was compared with the DCT and PST based schemes. The comparative studies showed that the SLT-domain watermarking scheme performed better against the copy and paste attack and additive Gaussian noise. However, the PST and DCT-domain watermarking schemes performed better than the SLT-domain watermarking against JPEG compression. For the WBCT semi-fragile watermarking scheme, watermarking bits were embedded by modulating the parent-children relationship in the contourlet domain. The experimental results demonstrated that our proposed WBCT watermarking scheme
achieved good performances in detecting different kinds of manipulations with $P_{FNR}$ at approximately 1%, whilst maintaining a $P_{FPR}$ below 6.5%. Overall, the use of the parent-children relationship of WBCT allowed our algorithm to detect and localise the manipulated areas accurately when certain degrees of non-malicious manipulations were applied. Furthermore, we also analysed and compared the difference between these two schemes.
Figure 4.23: ‘Trucks’ image (a) watermarked image (b) copy-and-pasted image with Contrast stretching (c) authenticated image (d) authenticated image with morphological operations
Chapter 5

Proposed Image Forensics Technique for Semi-fragile Watermarking

As mentioned in Chapters 2 and 4, semi-fragile watermarking has become increasingly important to verify the content of images and localise the tampered areas, while tolerating some non-malicious manipulations. In the literature, the majority of semi-fragile algorithms have applied a predetermined threshold to tolerate errors caused by JPEG compression. However, this predetermined threshold is typically fixed and cannot be easily adapted to different amounts of errors caused by unknown JPEG compression at different quality factors (QFs). In this chapter, we analyse the relationship between QF and threshold, and propose the use of generalised Benford’s Law as an image forensics technique for semi-fragile watermarking. Hence, our scheme can adaptively adjust the threshold for images based on the estimated QF, improving accuracy rates in authenticating and localising the tampered regions for semi-fragile watermarking. We also investigate the improved method on our proposed SLT and WBCT-based semi-fragile watermarking schemes, as discussed in Chapter 4.
5.1 Three Common Approaches Using Pre-determined Threshold in Semi-fragile Watermarking

Semi-fragile watermarking scheme has been used to authenticate and localise malicious tampering of image content, while permitting some non-malicious or unintentional manipulations. These manipulations can include some mild signal processing operations such as those caused by transmission and storage of JPEG images. However, watermarked images could be compressed by unknown JPEG QFs. As a result, in order to authenticate the images, these algorithms have to set a pre-determined threshold that could allow them to tolerate different QF values when extracting the watermarks.

The art of determining the threshold values for semi-fragile watermarking schemes has been extensively documented by several researchers [19], [20], [21], [24]. In this section, we review three common approaches. The first approach uses a threshold for authenticating each block of the image [19],[21]. In this scheme, if correlation coefficients (between the extracted watermark and its corresponding original watermark) is smaller than a specified threshold, this block is then classified as a tampered block, as given in Equation 5.1.

\[
    cr(w, w') < \tau, \max(\tau) - \tau = TM
\]  

(5.1)

where \(\max(\tau) = 1\) is the maximum threshold value with \(w = w'\), and \(TM\) is the JPEG compression tolerance margin. We discuss this approach in more detail in Section 5.2. The second approach uses a threshold, which has been pre-determined during the watermark embedding process [20], [21]. An example is illustrated in Figure 5.1, where the watermarks \(w\) are embedded into each side of threshold \(\tau\) according to the watermark value (e.g., 0 or 1), by shifting or substituting the corresponding coefficient. The value of \(T\) and \(-T\) controls the perceptual quality of the watermarked image. Threshold \(\tau\) is determined empirically to detect the watermark while extracting the watermarks \(w'\). \(TM\) is the JPEG compression tolerance margin. If \(w' > \tau\) then \(w' = 1\), otherwise \(w' = 0\).

The third approach uses a threshold for comparison with the results from the Tamper Assessment Function (TAF) during the authentication of images [24]. The extracted watermarks \(w'\) and their corresponding original watermarks \(w\) are calculated by using \(TAF\), as in Equation 5.2.

\[
    TAF(w, w') = \frac{1}{N_w} \sum_{i=1}^{N_w} w(i) \oplus w'(i)
\]  

(5.2)
Figure 5.1: The pre-determined threshold during the watermark embedding process.

where \( N_w \) is the length of the watermark, and \( \oplus \) is the exclusive-OR (XOR) operator. The \( TAF \) value is compared with a threshold \( \tau \), where \( 0 \leq \tau \leq 1 \). If \( TAF(w, w') > \tau \), then the watermarked image is considered as a tampered image, otherwise it is not. The tolerance margin can also be denoted as \( TM = 1 - \tau \). The thresholds \( \tau \) mentioned previously are pre-determined which will result in some fixed tolerance margins. A significant amount of research has been dedicated to improving the watermark embedding algorithms by analysing the characteristics of JPEG coefficients of the compressed watermarked image [22], [23], [24]. Alternatively, error correction coding (ECC) has been used for improving watermark detection and authentication rates [20]. However, as far as we know, the relationship between QF and threshold has not been discussed in the literature. If the QF could be estimated, then appropriate thresholds could be adapted for each test image, before initialising the watermark extraction and authentication process.

5.2 Issue of Threshold in a Simple DCT Semi-fragile Watermarking Algorithm

In this section, the feasibility of our proposed method is investigated in detail. By analysing the first approach previously reviewed in [19],[21], a simple semi-fragile watermarking algorithm based on the discrete cosine transform (DCT) and the importance of threshold is also described.

5.2.1 The DCT Watermark Embedding Process

As shown in Figure 5.2, the original image is divided into non-overlapping sub-blocks of \( 8 \times 8 \) pixels and DCT is applied to each block. The watermark embedding process is
achieved by modifying the randomly selected mid-frequency of the DCT coefficients in each block as follows:

\[
cof' = \begin{cases} 
cof, & (cof \geq T \land w = 1) \lor (cof \leq -T \land w = -1) \\
\alpha, & (cof < T \land w = 1) \\
-\alpha, & (cof > T \land w = -1)
\end{cases}
\]  

(5.3)

where \(cof\) is the original DCT coefficient, \(cof'\) is the modified DCT coefficient, \(w\) is the watermark bits generated via a pseudo-random sequence (1 and -1) using a secret key, \(T > 0\) determines the perceptual quality of the watermarked image and \(\alpha \in \left[\frac{T}{2}, T\right]\) is the watermark strength factor. The inverse DCT is then applied to each block to obtain the watermarked image.

5.2.2 The DCT Watermark Detection and Authentication Process

In Figure 5.3, the test image is first divided into non-overlapping sub-blocks of 8x8 pixels, and DCT is then applied to each block. The watermark detection algorithm shown in Equation 5.4 is then applied.

\[
w' = \begin{cases} 
1, & (cof' \geq 0) \\
-1, & (cof' < 0)
\end{cases}
\]  

(5.4)

where \(w'\) is the extracted watermark bits and \(cof'\) is the DCT coefficient of the test image. The extracted watermark bits \(w'\) from each block are compared with its corresponding original watermark bits \(w\) to obtain the correlation coefficient \(cr\) as shown in Equation 5.5.

\[
cr(w, w') = \frac{\sum(w' - \bar{w}') \cdot (w - \bar{w})}{\sqrt{\sum(w' - \bar{w}')^2 \sum(w - \bar{w})^2}}
\]  

(5.5)

The correlation coefficient of each block is then compared with a pre-determined threshold \(-1 \leq \tau \leq 1\) as given in Equation 5.6.
Chapter 5. Proposed Image Forensics for Semi-fragile Watermarking

\[
\text{Block} = \begin{cases} 
\text{Not tampered}, & cr(w, w') \geq \tau \\
\text{tampered}, & cr(w, w') < \tau 
\end{cases} \tag{5.6}
\]

5.2.3 The Importance of the Threshold

As we mentioned in 4, the magnitude of the threshold affects the false positive rate \(P_{FPR}\), which is the percentage of un-tampered blocks detected as being tampered and the false negative rate \(P_{FNR}\), which is the percentage of tampered blocks detected as being un-tampered. Figure 5.4 shows that the \(P_{FNR}\) decreases if the threshold is in close proximity to 1. This also leads to an increase in the \(P_{FPR}\). However, if the threshold is set to be of a close proximity to -1, then the \(P_{FNR}\) increases and the \(P_{FPR}\) will decrease. This results in a dilemma in determining a suitable threshold. For the proposed semi-fragile watermarking scheme, the threshold is set at 0.5, which provides a good trade-off between \(P_{FPR}\) and \(P_{FNR}\).

![Figure 5.3: An illustration of the DCT watermark detection and authentication processes.](image)

![Figure 5.4: The relationship among threshold \(P_{FPR}\) and \(P_{FNR}\).](image)

Figures 5.5 and 5.6 illustrate the overall relationship between the threshold, false positive and false negative detection rates. The watermarked image *Lena* has been tampered
with a rectangular block and JPEG compressed at $QF = 75$. Figure 5.5(a) shows the pre-determined threshold $T = 0.5$ used for authentication. The authenticated image shows that the proposed semi-fragile watermarking scheme can localise the tampered region with reasonable accuracy, but with some false positive detection errors. In Figures 5.5(b) and 5.5(c), the lower and upper thresholds $T = 0.3$ and $T = 0.7$ were used for comparison, respectively. Figure 5.5(b) shows that the false positive rate has decreased whilst the false negative rate has increased in the authenticated image. Figure 5.5(c) shows the image has a lower false negative rate but with a higher false positive rate. From this comparison, $T = 0.5$ was chosen for JPEG compression at $QF = 75$. However, if $QF = 95$, then $T = 0.5$ may not be adequate as shown in Figure 5.6(a). The false negative rate is higher than Figure 5.6(b) with $T = 0.9$. Therefore, it would be advantageous to be able to estimate the QF of JPEG compression, so that an adaptive threshold can be applied for increasing the authentication accuracy.

In this chapter, we discuss our proposed method [151] using the generalised Benford’s Law, as an image forensics technique to estimate the QF for semi-fragile watermarked images. The background of Benford’s Law, generalised Benford’s Law and their relationship with the watermarked image, JPEG compressed watermarked image are discussed in the next section.

![Figure 5.5: Different thresholds for $QF = 75$](image)

5.3 Generalized Benford’s Law vs. Watermarked Images

As discussed in Section 2.3.2, Benford’s Law has recently attracted a significant amount of research interests in image processing and image forensics. As the 1st digits of DCT coefficients of a digital image obey the Benford’s Law, Fu et al. [134] proposed a generalised Benford’s Law, used for estimating the QF of the JPEG compressed image for image forensics. In this section, the feasibility of generalised Benford’s Law for use in semi-fragile watermarking was first investigated. In our experiment, we selected 1338
uncompressed grayscale images from the Uncompressed Image Database (UCID) [152] for analysis to ensure that there was no compression performed on the images previously. Throughout this section we adhere to the same terminology as used in [134], where JPEG coefficients refer to the $8 \times 8$ block-DCT coefficients after the quantisation.

Figure 5.7 illustrates the comparison between the probability distribution of Benford’s Law, mean distribution of the $1^{st}$ digit of block-DCT coefficients of 1338 images and the watermarked images. The average PSNR between the original images and watermarked images was approximately $35.71 \text{dB}$, which is considered to be of acceptable image quality. Figure 5.7 shows that the distribution of the $1^{st}$ digits of the block-DCT coefficients for the uncompressed images obeys the Benford’s Law closely. This was also observed by Fu et al. [134] in their analysis. In terms of the watermarked images, the mean distribution also follows the Benford’s Law. The mean standard deviations of the 1338 uncompressed images and their watermarked images are considered to be relatively small, as shown in Table 5.1. The average divergence [134] for watermarked images is also found to be small at 0.0115. This indicates a good fit between the Benford’s Law and watermarked images. The average divergence is given in Equation 5.7.

$$x^2 = \sum_{i=1}^{9} \frac{(p_i' - p_i)}{p_i}$$

(5.7)

where $p_i'$ is the $1^{st}$ digit probability of the DCT coefficients of the watermarked images and $p_i$ is the $1^{st}$ digit probability from Benford’s Law in Equation 2.6. Hence, the results indicate that the probability distribution of the $1^{st}$ digits of the block-DCT coefficients of the watermarked images follow the Benford’s Law. Figure 5.8(a) illustrates an example of $8 \times 8$ DCT coefficients. The $1^{st}$ digits of the AC coefficients are then extracted as shown in Figure 5.8(b).
Proposed Image Forensics for Semi-fragile Watermarking

0.35

0.3

0.25

0.2

0.15

0.1

0.05

0

1 2 3 4 5 6 7 8 9

First digit

Probability

0.35

0.3

0.25

0.2

0.15

0.1

0.05

0

1 2 3 4 5 6 7 8 9

Table 5.1: Mean standard deviations of 1338 images

<table>
<thead>
<tr>
<th>First digit</th>
<th>Original images</th>
<th>Watermarked images</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0139</td>
<td>0.0145</td>
</tr>
<tr>
<td>2</td>
<td>0.0084</td>
<td>0.0078</td>
</tr>
<tr>
<td>3</td>
<td>0.0067</td>
<td>0.0068</td>
</tr>
<tr>
<td>4</td>
<td>0.0050</td>
<td>0.0049</td>
</tr>
<tr>
<td>5</td>
<td>0.0037</td>
<td>0.0030</td>
</tr>
<tr>
<td>6</td>
<td>0.0032</td>
<td>0.0023</td>
</tr>
<tr>
<td>7</td>
<td>0.0028</td>
<td>0.0021</td>
</tr>
<tr>
<td>8</td>
<td>0.0028</td>
<td>0.0023</td>
</tr>
<tr>
<td>9</td>
<td>0.0022</td>
<td>0.0021</td>
</tr>
</tbody>
</table>

Figures 5.9 to 5.11 illustrate the comparisons between the probability distribution of Benford's Law, generalized Benford's Law and the mean distributions of the 1st digits of block JPEG coefficients of the watermarked images compressed at $QF = 100, 75, 50$, respectively. Table 5.2 summarises the mean standard deviations obtained for the 1338 original and watermarked images, JPEG compressed at the three QF rates, which are considered to be relatively small. Furthermore, as shown in Table 5.3, the $\chi^2$ divergences indicate a good fitting between the generalized Benford's Law and watermarked images compressed with different QFs. The results indicate that the probability distributions of the 1st digits of JPEG coefficients of the watermarked images, in Figures 5.9 to 5.11,
obey the generalised Benford's Law model proposed by Fu et al. [134], as given in Equation 2.7. Hence, we could employ their model to estimate the unknown QF of test images to determine the threshold for authentication. The improved authentication process is described in Section 5.4.

Figure 5.9: 1st digit of JPEG coefficients (QF = 100)
Chapter 5. *Proposed Image Forensics for Semi-fragile Watermarking*  

**Figure 5.10:** 1st digit of JPEG coefficients ($QF = 75$)

**Figure 5.11:** 1st digit of JPEG coefficients ($QF = 50$)
Table 5.2: Mean standard deviations of 1338 JPEG compressed images

<table>
<thead>
<tr>
<th>1st digits</th>
<th>Original images</th>
<th>Watermarked images</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QF100</td>
<td>QF75</td>
</tr>
<tr>
<td>1</td>
<td>0.0828</td>
<td>0.0327</td>
</tr>
<tr>
<td>2</td>
<td>0.0165</td>
<td>0.0067</td>
</tr>
<tr>
<td>3</td>
<td>0.0169</td>
<td>0.0066</td>
</tr>
<tr>
<td>4</td>
<td>0.0163</td>
<td>0.0058</td>
</tr>
<tr>
<td>5</td>
<td>0.0142</td>
<td>0.0049</td>
</tr>
<tr>
<td>6</td>
<td>0.0123</td>
<td>0.0043</td>
</tr>
<tr>
<td>7</td>
<td>0.0107</td>
<td>0.0037</td>
</tr>
<tr>
<td>8</td>
<td>0.0094</td>
<td>0.0032</td>
</tr>
<tr>
<td>9</td>
<td>0.0084</td>
<td>0.0027</td>
</tr>
</tbody>
</table>

Table 5.3: Average $x^2$ of 1338 compressed watermarked images

<table>
<thead>
<tr>
<th>QF</th>
<th>Model Parameters</th>
<th>$x^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>q</td>
</tr>
<tr>
<td>100</td>
<td>1.456</td>
<td>1.47</td>
</tr>
<tr>
<td>70</td>
<td>1.412</td>
<td>1.732</td>
</tr>
<tr>
<td>50</td>
<td>1.579</td>
<td>1.882</td>
</tr>
</tbody>
</table>

5.4 The Improved Authentication Method

In this section, we explain the improved authentication process which uses the generalised Benford’s Law model. In Figure 5.12, the test image is divided into non-overlapping blocks of 8 x 8 pixels and DCT is applied to each block. The watermark detection process then extracts the watermark bits using a secret key. The same test image is also used for detecting the QF in the quality factor estimation process. This process works by firstly classifying the test image as compressed (.jpg) or uncompressed (.bmp) by checking the extension name of the image. If the test image has been compressed, it will then be recompressed with the largest QF, starting from $QF = 100$ to $QF = 50$, in decreasing steps of 5. We decrease in steps of 5 as this gives us the most frequently used quality factors for JPEG compressed images (i.e. 95, 90, 85 etc.).

For each compressed test image, the probability distribution of the 1st digits of JPEG coefficients is obtained. Each set of values is then analysed by employing the generalized Benford’s Law and using the best curve-fitting to plot the data. In order to obtain the goodness of fit, we calculate the Sum of Squared Error (SSE) of the recompressed images. We can detect the QF of the test image by iteratively calculating the SSE for all QFs (starting at $QF = 100$, and decreasing in steps of 5), and as soon as $SSE < 10^{-6}$, we have reached the estimated QF for the test image. As per the pseudo-code given below, the threshold $10^{-6}$ has been set to allow us to detect the QF of the test image.
This threshold value was reported in [134], and has been verified by the results in our experiment.

\[
\text{if } SSE < 10^{-6} \text{ then } \\
\quad \text{QF has been detected} \\
\quad \text{break} \\
\text{end if}
\]

Figure 5.13 illustrates the results of estimating the QF for a test image that has previously been compressed with \( QF = 70 \). Three curves have been drawn in order to fit the three probability distribution data sets: generalized Benford's Law for \( QF = 70 \), the test image recompressed with \( QF = 70 \), and separately recompressed at \( QF = 90 \). The distribution of \( QF = 90 \) shows the worst fit and is shown to fluctuate considerably, while the distribution of \( QF = 70 \) is a generally decreasing curve, which also follows the trend of generalized Benford's Law. These results indicate that if the test image has been double compressed without the same quality factor, the probability distribution will not obey the generalised Benford's Law.

Once the QF is estimated, the threshold \( T \) can be adapted according to different estimated QFs, based on the following conditions in Equation 5.8 that could reduce the error detection rates. Finally, the correlation coefficient between original watermarks and extracted watermarks for each block is compared using the attuned threshold \( T \) as part of the authentication process, in order to determine whether any blocks have been tampered with.

\[
T = \begin{cases} 
0.9, & QF \geq 90 \\
0.7, & 90 > QF > 75 \\
0.5, & QF \leq 75 
\end{cases}
\] (5.8)
Chapter 5. Proposed Image Forensics for Semi-fragile Watermarking

5.5 Experimental Results (for a Simple DCT Semi-fragile Watermarking Scheme)

The watermarked images are generated by a simple DCT domain based semi-fragile watermarking algorithm (as discussed in Section 5.2.1) using the 1338 test images from UCID [152]. In order to achieve a fair comparison, different embedding parameters are randomised for each image such as the watermark location, watermark string and watermark bits. Ten types of test images, with and without attacks are considered for our analysis, as shown in Figure 5.14. Each set illustrated in Figure 5.14 is performed individually for the 1338 watermarked images.

Table 5.4 summaries the results obtained for the test images that have been JPEG compressed only. To evaluate the accuracy of the quality factor estimation process, each test image has been blind compressed from $QF = 100$ to $QF = 50$ in decreasing steps of 5. For each JPEG compression, the quality factor estimation process was used to determine the QF. The mean estimated QFs for all 1338 test images and each correctly identified detection accuracy rate $P_{de}$ for each JPEG compression quality factor are shown in Table 5.4, based on equation 5.9.
where $\vartheta$ is the number of images with their QF correctly detected and $\beta$ is the number of images tested. The only exceptions for lower correct detection rates, $P_{de}$, were obtained for $QF = 50$, $QF = 60$, and $QF = 100$. In the case of $QF = 50$, $P_{de}$ was very low at approximately 18.2%, indicating that the process was probably detecting QFs close to $QF = 55$. For $QF = 60$, and $QF = 100$, the detection rates were slightly better at 38.6% and 65.7%, respectively. For comparison, both the mean estimated QF value and correct detection rate were used for each result to estimate the actual QF for the images. The QFs were then grouped into three different ranges: $QF > 90$, $90 > QF > 75$ and $QF \leq 75$. The grouping into three QF ranges did not have an overall effect on the authentication process. Results obtained for $P_{de2}$, which show the correct detection accuracy rates in these QF ranges, also were on average at 99%. Two further experiments were performed with the test image: no modification, and copy and paste attack (5%). All of the detected QFs achieved for both experiments were approximately 99%, and fit well in the upper range of $QF \geq 90$. 

\[
P_{de} = \frac{\vartheta}{\beta} \times 100\% \quad (5.9)
\]
Table 5.5 summaries the results obtained for test images that have been attacked via copy and paste and then JPEG compressed. Each watermarked image has been tampered randomly in different regions by applying a copy and paste attack to 5% of the watermarked image (9830 pixels in 384512 pixels image), and also compressed with different QF values. In order to further investigate our proposed method, we undertake the analysis of increasing the copy and paste attack area in the watermarked image. Tables 5.6 and 5.7 illustrate the average QF estimation rates based on 1338 watermarked images, each attacked via copy and paste (20% and 30%), and then JPEG compression. Tables 5.5 to 5.7 show that each watermarked image was exposed to three different regions of tampering: 5%, 20%, and 30%. Each tampering iteration was performed by selecting random blocks from the watermarked images. Each tampered watermarked image is then blind JPEG compressed from $QF = 100$ to $QF = 50$ in decreasing steps of 5. Consequently, 44154 test images were obtained during these copy and paste attack and JPEG compression simulations. The results showed that the quality factor estimation process was highly accurate even under these attacks.

From Table 5.5, the lowest correct detection rates were obtained for $QF = 50$, $QF = 60$, and $QF = 100$ with a tamper region of 5%. On the other hand, in Tables 5.6 and 5.7, the correct detection rates for $QF = 100$ were increased to 94% and 100%, when the tampered regions were 20% and 30%, respectively. However, the correct detection rates $P_{de}$ for $QF = 50$ and $QF = 60$ were still maintaining the lowest in Tables 5.6 and 5.7. Nevertheless, from Tables 5.5 to 5.7, all the results of $P_{de2}$ showed the correct detection rates in three QF ranges ($QF \geq 90$, $90 > QF > 75$ and $QF \leq 75$), with an overall average of 99%. As such, the threshold can be adapted into the three QF ranges according to the estimated QF of each test image as described in Section 5.4.

As mentioned in Chapter 4, semi-fragile watermarking techniques can permit some non-malicious or unintentional manipulations. Aside from JPEG compression as one of the main unintentional manipulations, other non-malicious manipulation can include image enhancement techniques such as median filtering, average filtering, Gaussian lowpass filtering, and histogram equalisation. Hence, we perform the experiments to obtain 58872 test images by firstly tampering 20% of the watermarked image, before applying these image enhancement techniques along with blind JPEG compression from $QF = 100$ to $QF = 50$ in decreasing steps of 5. The results are shown in Tables 5.8 to 5.11. Table 5.8 indicates that the average performance of the correct detection rates $P_{de}$ is increased when the median filtering is added to the test images. When $QF = 95$ to $QF = 65$, the correct detection rates $P_{de}$ reach 100%, which implies that all of the corresponding QFs for each test images are correctly detected.
Chapter 5. Proposed Image Forensics for Semi-fragile Watermarking

In Table 5.9, the test images are subjected to average filtering. The results illustrate that $P_{de}$ have been decreased relatively small, while the results of $QF = 95$ to $QF = 65$ still maintained at over 96%. Table 5.10 shows that the highest correct detection rates (100%) are achieved when Gaussian lowpass filtering is applied to the test images with $QF = 100$ to $QF = 80$. However, the $P_{de}$ decreased approximately 8% from $QF = 75$ to $QF = 50$, and $P_{de2}$ decreased to the lowest at 96.6%. In Table 5.11, we evaluate the results of the test images after histogram equalisation. From Tables 5.4 to 5.11, we found that $P_{de}$ for $QF = 60$ increased to 87.5% (the highest $P_{de}$ for $QF = 60$), whereas $P_{de}$ for $QF = 50$ decreased to the lowest at 4%. Tables 5.8 to 5.11 show the correct detection rates $P_{de2}$ were highly accurate with an overall average of 99.5%, which can also be adapted to adjust the threshold into three ranges.

**Table 5.4: JPEG compression only**

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>$P_{de}$</th>
<th>T</th>
<th>$P_{de2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>98.2</td>
<td>65.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>94.9</td>
<td>97.3%</td>
<td>0.9</td>
<td>98.8%</td>
</tr>
<tr>
<td>90</td>
<td>90.1</td>
<td>98.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>84.2</td>
<td>91.4%</td>
<td>0.7</td>
<td>99.1%</td>
</tr>
<tr>
<td>80</td>
<td>79.8</td>
<td>97.3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>75.4</td>
<td>97.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>69.8</td>
<td>98.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>64.4</td>
<td>93.7%</td>
<td>0.5</td>
<td>99.4%</td>
</tr>
<tr>
<td>60</td>
<td>62.4</td>
<td>38.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>55.2</td>
<td>94.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>54.3</td>
<td>18.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.5: Combined (5%) copy and paste attack and JPEG compression**

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>$P_{de}$</th>
<th>T</th>
<th>$P_{de2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>98.6</td>
<td>72%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>95.0</td>
<td>100%</td>
<td>0.9</td>
<td>99.1%</td>
</tr>
<tr>
<td>90</td>
<td>90.1</td>
<td>98.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>84.8</td>
<td>97.9%</td>
<td>0.7</td>
<td>99.3%</td>
</tr>
<tr>
<td>80</td>
<td>79.9</td>
<td>99.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>75.2</td>
<td>99.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>69.9</td>
<td>99.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>64.5</td>
<td>98.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>61.5</td>
<td>63.9%</td>
<td>0.5</td>
<td>99.2%</td>
</tr>
<tr>
<td>55</td>
<td>54.9</td>
<td>96.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>53.3</td>
<td>20.4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 5. Proposed Image Forensics for Semi-fragile Watermarking

Table 5.6: Combined (20%) copy and paste attack and JPEG compression

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>$P_{de}$</th>
<th>$T$</th>
<th>$P_{de2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>99.7</td>
<td>94%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>95.0</td>
<td>100%</td>
<td>0.9</td>
<td>100%</td>
</tr>
<tr>
<td>90</td>
<td>90.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>85.0</td>
<td>100%</td>
<td>0.7</td>
<td>99.9%</td>
</tr>
<tr>
<td>80</td>
<td>80.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>75.5</td>
<td>98%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>70.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>65.2</td>
<td>96%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>61.6</td>
<td>68%</td>
<td>0.5</td>
<td>99.9%</td>
</tr>
<tr>
<td>55</td>
<td>55.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>54.9</td>
<td>20%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7: Combined (30%) copy and paste attack and JPEG compression

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>$P_{de}$</th>
<th>$T$</th>
<th>$P_{de2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>95.00</td>
<td>100%</td>
<td>0.9</td>
<td>99.9%</td>
</tr>
<tr>
<td>90</td>
<td>90.2</td>
<td>99.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>84.8</td>
<td>98%</td>
<td>0.7</td>
<td>99.9%</td>
</tr>
<tr>
<td>80</td>
<td>79.1</td>
<td>99.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>75.2</td>
<td>99.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>69.4</td>
<td>99.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>65.3</td>
<td>99.5%</td>
<td>0.5</td>
<td>99.9%</td>
</tr>
<tr>
<td>60</td>
<td>62.1</td>
<td>58%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>55.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>55.2</td>
<td>14%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.6 Improved Results (for Both SLT and WBCT Semi-fragile Watermarking Schemes)

In this section, we further analyse the use of Benford's Law to improve the accuracy rates in authenticating and localising the tampered regions for our proposed SLT and WBCT based semi-fragile watermarking schemes.

5.6.1 Benford's Law for SLT-based Semi-fragile Watermarking

In Section 4.1.3, we discussed the SLT based semi-fragile watermark detection and authentication process. The original watermarks and the retrieved watermarks corresponding to the block from test image were compared by using the correlation coefficient $\rho$. Then, the correlation coefficient $\rho$ could be compared with a pre-determined threshold value $\lambda$. If $\rho < \lambda$, the authentication process indicates that the block has been tampered...
Table 5.8: Combined (20%) copy and paste attack and Median filtering (3 × 3) and JPEG compression

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>$P_{de}$</th>
<th>$T$</th>
<th>$P_{de2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>99.9</td>
<td>98%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>95.1</td>
<td>100%</td>
<td>0.9</td>
<td>100%</td>
</tr>
<tr>
<td>90</td>
<td>90.3</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>85.0</td>
<td>100%</td>
<td>0.7</td>
<td>100%</td>
</tr>
<tr>
<td>80</td>
<td>80.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>75.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>70.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>65.0</td>
<td>100%</td>
<td>0.5</td>
<td>98.0%</td>
</tr>
<tr>
<td>60</td>
<td>62.4</td>
<td>52%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>56.6</td>
<td>92%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>54.0</td>
<td>30%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.9: Combined (20%) copy and paste attack and Average filtering (3 × 3) and JPEG compression

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>$P_{de}$</th>
<th>$T$</th>
<th>$P_{de2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>99.8</td>
<td>96%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>94.9</td>
<td>98%</td>
<td>0.9</td>
<td>100%</td>
</tr>
<tr>
<td>90</td>
<td>90.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>85.0</td>
<td>100%</td>
<td>0.7</td>
<td>100%</td>
</tr>
<tr>
<td>80</td>
<td>80.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>74.5</td>
<td>98%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>67.2</td>
<td>96%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>58.5</td>
<td>90%</td>
<td>0.5</td>
<td>100%</td>
</tr>
<tr>
<td>60</td>
<td>59.7</td>
<td>54%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>48.4</td>
<td>88%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>44.4</td>
<td>58%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

with. In this case, the $\lambda$ has been set as 0.5 in order to get acceptable error rates. However, both $P_{PR}$ and $P_{NR}$ could be reduced by using our improved authentication process based on the generalized Benford's Law.

The test image is first estimated to obtain the QF by using the generalized Benford's Law. Then, the threshold $\lambda$ can be adapted according to different estimated QFs, based on the following conditions in Equation 5.10. The adaptive three thresholds $\lambda$ were estimated 0.95, 0.75 or 0.5 through experiments. Finally, the correlation coefficient $\rho$ of each block was compared with this adaptive threshold $\lambda$ to determine whether any blocks have been tampered with.
TABLE 5.10: Combined (20%) copy and paste attack and Gaussian lowpass filtering \((3 \times 3)\) and JPEG compression

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>(P_{de})</th>
<th>(T)</th>
<th>(P_{de2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>95.0</td>
<td>100%</td>
<td>0.9</td>
<td>100%</td>
</tr>
<tr>
<td>90</td>
<td>90.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>85.0</td>
<td>100%</td>
<td>0.7</td>
<td>100%</td>
</tr>
<tr>
<td>80</td>
<td>80.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>72.9</td>
<td>89.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>68.4</td>
<td>85.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>59.7</td>
<td>91.8%</td>
<td>0.5</td>
<td>96.6%</td>
</tr>
<tr>
<td>60</td>
<td>61.3</td>
<td>46.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>57.0</td>
<td>91.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>57.9</td>
<td>32.7%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE 5.11: Combined (20%) copy and paste attack and Histogram equalization and JPEG compression

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>(P_{de})</th>
<th>(T)</th>
<th>(P_{de2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>99.7</td>
<td>93.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>94.9</td>
<td>97.9%</td>
<td>0.9</td>
<td>100%</td>
</tr>
<tr>
<td>90</td>
<td>90.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>84.9</td>
<td>97.9%</td>
<td>0.7</td>
<td>98.9%</td>
</tr>
<tr>
<td>80</td>
<td>79.7</td>
<td>97.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>75.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>70.0</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>65.0</td>
<td>100%</td>
<td>0.5</td>
<td>100%</td>
</tr>
<tr>
<td>60</td>
<td>60.6</td>
<td>87.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>55.1</td>
<td>97.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>54.8</td>
<td>4.1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\lambda = \begin{cases} 
0.95, & QF \geq 95 \\
0.75, & 95 > QF > 80 \\
0.5, & QF \leq 80 
\end{cases} \tag{5.10}
\]

In Table 5.12, the results based on nine images show the improved performance of the SLT semi-fragile watermarking scheme against copy and paste attack comparing with their original results obtained from Section 4.1.5. The adaptive threshold has been set to 0.95 if the test image has been compressed with \(QF \geq 95\) or uncompressed. From Table 5.12, \(P_{FNR}\) has decreased significantly, from approximately 9.55% to 1%. Especially, images 'baboon', 'San Diego' and 'Gun', have been decreased 9% compared with their original results. On average, \(P_{FNR}\) has approximately decreased 8.55%.
In addition, Table 5.13 shows an improved performance of the SLT semi-fragile watermarking scheme against copy and paste attack and JPEG compression (QF=85) by using the generalized Benford's Law. The adaptive threshold has been set to 0.75 as the estimated QF is between 95 and 80. Similarly, $P_{FNR}$ also has also decreased significantly, from 9.41% to 2.16% on average. However, when applying JPEG compression to the watermarked images, $P_{FNR}$ increases to approximately 2.16% comparing with Table 5.12. However, with the improved authentication process based on the generalized Benford's Law, $P_{FNR}$ still decreases approximately by 7.25%. On the other hand, $P_{FPR}$ has risen slightly to 0.04%, with approximately only 0.01% increased on average, as indicated in Table 5.13.

**Table 5.12:** The improved performance of the SLT semi-fragile watermarking scheme against copy and paste attack (20%) by using Benford's Law.

<table>
<thead>
<tr>
<th>Test Image</th>
<th>Original $P_{FPR}$</th>
<th>Improved $P_{FPR}$</th>
<th>Errors Decreased/Increased $P_{FPR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>0</td>
<td>9.06</td>
<td>0</td>
</tr>
<tr>
<td>Baboon</td>
<td>0</td>
<td>9.41</td>
<td>-9.04</td>
</tr>
<tr>
<td>Ship</td>
<td>0</td>
<td>9.99</td>
<td>-7.6</td>
</tr>
<tr>
<td>Trucks</td>
<td>0</td>
<td>9.47</td>
<td>-8.98</td>
</tr>
<tr>
<td>Bridge</td>
<td>0</td>
<td>9.38</td>
<td>-8.94</td>
</tr>
<tr>
<td>San Diego</td>
<td>0</td>
<td>9.65</td>
<td>-9.15</td>
</tr>
<tr>
<td>Gun</td>
<td>0</td>
<td>10.05</td>
<td>-9.61</td>
</tr>
<tr>
<td>Car1</td>
<td>0</td>
<td>9.54</td>
<td>-7.79</td>
</tr>
<tr>
<td>Car2</td>
<td>0</td>
<td>9.37</td>
<td>-8.87</td>
</tr>
<tr>
<td>Average</td>
<td>0</td>
<td>9.55</td>
<td>-8.55</td>
</tr>
</tbody>
</table>

**Table 5.13:** The improved performance of the SLT semi-fragile watermarking scheme against copy and paste attack (20%) and JPEG compression (QF=85) by using Benford's Law.

<table>
<thead>
<tr>
<th>Test Image</th>
<th>Original $P_{FPR}$</th>
<th>Improved $P_{FPR}$</th>
<th>Errors Decreased/Increased $P_{FPR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>0</td>
<td>8.58</td>
<td>-6.22</td>
</tr>
<tr>
<td>Baboon</td>
<td>0.01</td>
<td>9.63</td>
<td>-7.4</td>
</tr>
<tr>
<td>Ship</td>
<td>0</td>
<td>9.47</td>
<td>-7.63</td>
</tr>
<tr>
<td>Trucks</td>
<td>0</td>
<td>9.89</td>
<td>-7.63</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.05</td>
<td>9.56</td>
<td>-7.41</td>
</tr>
<tr>
<td>San Diego</td>
<td>0</td>
<td>9.55</td>
<td>-7.3</td>
</tr>
<tr>
<td>Gun</td>
<td>0.03</td>
<td>9.1</td>
<td>-6.68</td>
</tr>
<tr>
<td>Car1</td>
<td>0.01</td>
<td>9.78</td>
<td>-7.78</td>
</tr>
<tr>
<td>Car2</td>
<td>0.17</td>
<td>9.11</td>
<td>-7.16</td>
</tr>
<tr>
<td>Average</td>
<td>0.03</td>
<td>9.41</td>
<td>-7.25</td>
</tr>
</tbody>
</table>
5.6.2 Benford's Law for WBCT-based Semi-fragile Watermarking

Similarly, in this section, we show that the WBCT based semi-fragile watermark detection process can also be improved by using the generalized Benford's Law. As discussed in Section 4.2.4, the $M$, the error tolerance margin value, is pre-determined to be 5 and is used to control $P_{FP}$ and $P_{FN}$. In the new improved watermark detection process, the test image is first used to estimate the QF value through the generalized Benford's Law. The threshold $M$ was then set adaptively according to the estimated QF, as given in Equation 5.11. The watermark bits $w'$ were extracted from the parent and children relationships using the adaptive threshold $M$, as discussed in Section 4.2.4.

\[
M = \begin{cases} 
3, & QF \geq 90 \\
4, & 90 > QF \geq 70 \\
5, & QF < 70 
\end{cases} \quad (5.11)
\]

In Equation 5.11, the adaptive thresholds 3, 4, and 5 were found through our experiments, which could reduce the detection rates according to different QFs. Tables 5.14 and 5.15 show the experiential results of improved performance of the WBCT-based semi-fragile watermarking scheme against copy and paste attack with JPEG compression (QF=90 and 70) based on the same nine images. The results are compared with their original results obtained from Section 4.2.6. As shown in Table 5.14, the adaptive threshold $M$ is set to 3, when the estimated $QF \geq 90$. $P_{FP}$ has reduced 1.35%. For image 'Lena', $P_{FP}$ increased to 1.34% from 4.47%. When the test image has been detected with JPEG compression (QF=70), the adaptive threshold $M$ is set as 4, with $P_{FP}$ decreased to 1.50% on average, as indicated in Table 5.15. Particularly, for image 'Car2', $P_{FP}$ has decreased by 5.96% from 6.68% to 0.72%. However, $P_{FN}$ in Tables 5.14 and 5.15 have increased 0.32% and 0.22% on average, respectively.
Chapter 5. Proposed Image Forensics for Semi-fragile Watermarking

Table 5.14: The improved performance of the WBCT semi-fragile watermarking scheme against copy and paste attack (20%) and JPEG compression (QF=90) by using Benford’s Law.

<table>
<thead>
<tr>
<th>Test Image</th>
<th>Original $P_{FPR}$</th>
<th>Original $P_{FNR}$</th>
<th>Improved $P_{FPR}$</th>
<th>Improved $P_{FNR}$</th>
<th>Errors Decreased/Increased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>4.47</td>
<td>0.69</td>
<td>1.34</td>
<td>1.1</td>
<td>-3.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.41</td>
</tr>
<tr>
<td>Boats</td>
<td>5.49</td>
<td>0.61</td>
<td>4.13</td>
<td>1.08</td>
<td>-1.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.47</td>
</tr>
<tr>
<td>Trucks</td>
<td>1.81</td>
<td>0.73</td>
<td>1.33</td>
<td>0.98</td>
<td>-0.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.25</td>
</tr>
<tr>
<td>San Diego</td>
<td>2.29</td>
<td>0.89</td>
<td>2.1</td>
<td>1.08</td>
<td>-0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.19</td>
</tr>
<tr>
<td>Peppers</td>
<td>4.27</td>
<td>0.08</td>
<td>4.91</td>
<td>0.34</td>
<td>+0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.26</td>
</tr>
<tr>
<td>Goldhill</td>
<td>4.63</td>
<td>0.55</td>
<td>1.99</td>
<td>0.83</td>
<td>-2.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.28</td>
</tr>
<tr>
<td>Gun</td>
<td>6.11</td>
<td>0.36</td>
<td>3.27</td>
<td>0.49</td>
<td>-2.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.13</td>
</tr>
<tr>
<td>Car1</td>
<td>2.48</td>
<td>0.72</td>
<td>0.84</td>
<td>1.19</td>
<td>-1.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.47</td>
</tr>
<tr>
<td>Car2</td>
<td>1.23</td>
<td>1.17</td>
<td>0.74</td>
<td>1.58</td>
<td>-0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.41</td>
</tr>
<tr>
<td>Average</td>
<td>3.64</td>
<td>0.64</td>
<td>2.29</td>
<td>0.96</td>
<td>-1.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.32</td>
</tr>
</tbody>
</table>

Table 5.15: The improved performance of the WBCT semi-fragile watermarking scheme against copy and paste attack (20%) and JPEG compression (QF=70) by using Benford’s Law.

<table>
<thead>
<tr>
<th>Test Image</th>
<th>Original $P_{FPR}$</th>
<th>Original $P_{FNR}$</th>
<th>Improved $P_{FPR}$</th>
<th>Improved $P_{FNR}$</th>
<th>Errors Decreased/Increased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lena</td>
<td>3.69</td>
<td>0.74</td>
<td>1.53</td>
<td>1.08</td>
<td>-2.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.34</td>
</tr>
<tr>
<td>Boats</td>
<td>5.6</td>
<td>0.5</td>
<td>1.62</td>
<td>0.8</td>
<td>-3.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.3</td>
</tr>
<tr>
<td>Trucks</td>
<td>0.83</td>
<td>1.19</td>
<td>1.81</td>
<td>1.22</td>
<td>+0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.03</td>
</tr>
<tr>
<td>San Diego</td>
<td>1.73</td>
<td>1.77</td>
<td>2.51</td>
<td>2.17</td>
<td>+0.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.4</td>
</tr>
<tr>
<td>Peppers</td>
<td>3.16</td>
<td>0.54</td>
<td>4.35</td>
<td>0.7</td>
<td>+1.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.16</td>
</tr>
<tr>
<td>Goldhill</td>
<td>6.29</td>
<td>0.53</td>
<td>3.74</td>
<td>0.38</td>
<td>-2.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.15</td>
</tr>
<tr>
<td>Gun</td>
<td>4.35</td>
<td>0.42</td>
<td>2.89</td>
<td>0.59</td>
<td>-1.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.17</td>
</tr>
<tr>
<td>Car1</td>
<td>1.25</td>
<td>1.15</td>
<td>0.94</td>
<td>1.44</td>
<td>-0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.29</td>
</tr>
<tr>
<td>Car2</td>
<td>6.68</td>
<td>1.31</td>
<td>0.72</td>
<td>1.76</td>
<td>-5.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.45</td>
</tr>
<tr>
<td>Average</td>
<td>3.73</td>
<td>0.91</td>
<td>2.23</td>
<td>1.13</td>
<td>-1.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.22</td>
</tr>
</tbody>
</table>

5.7 Summary

In this chapter, we discussed the relationship between QF and threshold, and proposed a framework incorporating the generalised Benford’s Law as an image forensics technique to accurately detect unknown JPEG compression levels in semi-fragile watermarked images. We discussed the limitations of using predetermined thresholds in semi-fragile watermarking algorithm. In our improved semi-fragile watermarking method, the test image was first analysed to detect its previously unknown quality factor for JPEG compression by using the generalised Benford’s Law model, before proceeding with the semi-fragile authentication process. The results showed that QFs could be accurately detected for most unknown JPEG compressions. In particular, the average QF detection rate was as high as 96% for watermarked images compressed with QFs between
95 and 65, and 99% when the image was subjected to tampering of 5% pixels of the image and compressed with QFs from 95 to 65. The threshold was adapted into three specific ranges according to the estimated QF of each test image. We also applied the generalised Benford’s Law to improve our proposed SLT and WBCT based semi-fragile watermarking schemes, which discussed in Chapter 4. The improved results showed that $P_{FNR}$ has deceased approximately by 8% for SLT-based scheme, and approximately by 1.4% reducing for $P_{FPR}$ on average in WBCT-based semi-fragile watermarking scheme.
Chapter 6

Conclusion and Future work

6.1 Conclusion

In this thesis, we introduced four novel robust and semi-fragile transform based image watermarking related schemes. These include wavelet-based contourlet transform (WBCT) for both robust and semi-fragile watermarking, slant transform (SLT) for semi-fragile watermarking as well as using generalised Benfords Law to adaptively adjust the appropriate threshold for improving semi-fragile watermarking technique.

One of the important applications of digital watermarking technology is copyright protection and ownership identification for digital images. To achieve this goal, robust watermarking has been rapidly developed in the past fifteen years. Robust watermarking is designed to survive various manipulations, such as JPEG compression, additive noise, filtering and geometric distortions. In contrast to conventional methods operating in the wavelet domain, we proposed a novel robust watermarking algorithm using non-redundant contourlet transform that exploited the energy relations between "parent" and "children" coefficients. In our proposed scheme, we embedded the watermarks by modulating these energy relations. The modulation was performed by modifying the parent coefficients in relation with children coefficients. Through experiments, we found that most of the energy relations between "parent" and "children" non-redundant contourlet coefficients maintained 75% of invariance before and after JPEG compression with QF=10, even when the image was distorted significantly. Therefore, performance improvement was obtained by means of embedding a watermark exploiting the modulation of the energy relations. By comparing with two other wavelet methods, the experimental results showed that our method was more robust to attacks such as JPEG and JPEG2000 compression, pixel shifting, histogram equalisation, Gaussian filtering.
and sharpening. However, for the median filtering, scale and rotation attacks, our results are lower than other two methods that need to be further improved.

In our proposed two semi-fragile watermarking schemes for image authentication, the block based SLT algorithm, and the non-block based wavelet-based contourlet transform were analysed and compared in detail. For the SLT scheme, the watermark was embedded into the middle frequency SLT coefficients of non-overlapping blocks of the test images. The comparative studies showed that the SLT-domain watermarking scheme performed better against the copy and paste attack and additive Gaussian noise by comparing with the DCT and PST based schemes. However, the PST and DCT-domain watermarking schemes performed better than the SLT-domain watermarking against JPEG compression. For instance, in the new comparative performance of the watermarking schemes against copy & paste attack (20\%) and JPEG compression (\text{QF}=75), $P_{FFR}$ was achieved at 16.22\% for SLT-based scheme, whereas PST and DCT achieved on average at 0.19\% and 1.5\%, respectively.

For the WBCT based scheme, we analysed the parent and children coefficients relationship before and after non-malicious manipulations, as well as copy and pasted attack. We found that most of the parent and children relationships before and after non-malicious manipulations maintained their invariance, whereas the coefficients relations for copy and paste attack exhibited a sharp peak at zero amplitude and tail off rapidly on both sides of the peak. This implies that the differences distribute sparsely, as the majority of differences were close to zero. Therefore, the watermark bits were embedded by modulating the parent and children relationship in the contourlet domain for our proposed semi-fragile watermarking scheme. The experimental results demonstrated that our proposed WBCT watermarking scheme achieved good performances in detecting different kinds of manipulations such as JPEG/JPEG2000 compression, Gaussian noise, Gaussian filtering and contrast stretching with $P_{FNR}$ at approximately 1\%, whilst maintaining a $P_{FFR}$ below 6.5\%. Overall, the use of the parent and children relationship of WBCT allowed our algorithm to detect and localise the manipulated areas accurately when certain degrees of non-malicious manipulations were applied. When compared with the SLT-domain semi-fragile watermarking scheme, the WBCT domain scheme was found to be successful for a wide range of images, as a result of the unique parent and child relationship for each image. This characteristic of parent-child relationships can be utilised for semi-fragile watermark embedding, extraction and authentication processes.

Semi-fragile watermarking has become increasingly important to verify the content of images and localise the tampered areas, while tolerating some non-malicious manipulations, such as JPEG compression. In the literature, the majority of semi-fragile algorithms have applied a predetermined threshold to tolerate errors caused by JPEG
compression. However, this predetermined threshold is typically fixed and cannot be easily adapted to different amounts of errors caused by unknown JPEG compression at different quality factors (QFs). We reviewed three typical methods of employing predetermined thresholds in semi-fragile watermarking algorithms and the limitations of using predetermined thresholds were also highlighted. We analysed the relationship between QF and threshold, and proposed an improved method by utilizing the generalised Benford’s Law model for semi-fragile watermarking scheme. The test image was first analysed to detect its previously unknown quality factor for JPEG compression, before proceeding with the semi-fragile authentication process. The results showed that QFs could be accurately detected for most unknown JPEG compressions. In particular, the average QF detection rate was as high as 96% for watermarked images compressed with QFs between 95-65, and 99% when the image was subjected to tampering of 5% pixels of the image and compressed with QFs between 95-65. The threshold was adapted into three specific ranges according to the estimated QF of each test image. We also applied the generalised Benford’s Law to improve our proposed SLT and WBCT based semi-fragile watermarking schemes. The improved results showed that the $P_{FNR}$ has been deceased around 8% for SLT-based scheme, and about 1.4% decrease for $P_{FPR}$ on average in WBCT-based semi-fragile watermarking scheme. In addition, we applied different image enhancement techniques, such as median filtering, average filtering, Gaussian low-pass filtering, and histogram equalisation, to these test images, and the results showed that the QF correct detection rates were above 90%. Hence, our proposed image forensics method can be used to adaptively adjust the threshold for images based on the estimated QF, improving accuracy rates in authenticating and localising the tampered regions for semi-fragile watermarking.

6.2 Future Work

As future work for our proposed robust watermarking method, we plan to extend our proposed algorithms further by improving the watermark robustness against different forms of mild signal processing attacks. The embedding method could be further improved by analyzing the parent and children WBCT coefficients before and after different attacks, such as, Gaussian white additive noise, different image filtering operations and some geometrical modifications. In addition, the watermark detection results could also be improved by having a more adaptive watermark detection threshold according to different forms of attacks. Currently, many researchers are focusing on the analysis of the characteristics of the image to find the invariant areas for embedding the watermarks. This trend could lead to a better design and development of robust watermarking scheme against geometrical attacks such as print-and-scan process. Another potential
area is the use of pattern recognition and neural network techniques to determine the type of attacks from the retrieved watermark patterns which could further enhance the development of robust image watermarking algorithms.

For semi-fragile watermarking, we plan to improve our proposed algorithms further with increasing accuracies in localisation and authentication, against different forms of mild signal processing attacks. Moreover, a security issue may occur when applying semi-fragile watermarking, as the watermark is embedded into each image with same key, which can be extracted easily by an attacker. Although much research is aimed at solving this issue, it is computationally intensive. It would be advantageous to develop a new and secure semi-fragile watermarking algorithm that could reduce the computational requirements of security. Both of SLT and WBCT based schemes need to be further analysed to find the optimal watermark embedding, detection and authentication algorithms, and adaptive correlated parameters, such as number of watermarks, watermark locations, watermark strength and threshold. Furthermore, the self-recovery and restoration of tampered regions requires further investigation, particularly in the use of more advanced restoration techniques such as irregular sampling and iterative techniques. Recently, we have proposed a novel fast self-restoration scheme resisting to JPEG compression for semi-fragile watermarking, which has been accepted by 10th International Workshop on Digital-forensics and Watermarking (IWDW11). In the watermark embedding process, we embed ten watermarks (six for authentication and four for self-restoration) into each 8 × 8-pixel block of the original image. We then utilise four (4 × 4-pixel) sub-blocks mean pixel values (extracted watermarks) to restore its corresponding (8 × 8-pixel) blocks first four DCT coefficients for image content recovering. We also plan to merge the ten watermarks for authentication and self-restoration together to reduce the number of watermarks for each block, which could further improve the imperceptibility of both watermarked and restored images.

Based on the Benford’s Law, we plan to further analyse and estimate other signal processing operations caused by transmission such as Gaussian noise, median filtering, Gaussian filtering and print-scan processes in semi-fragile and robust watermarking of images. We are also planning to utilise other image forensics techniques that could provide an improvement in semi-fragile watermarking schemes. In our recent research, we proposed two image forensics techniques based on analysing DWT coefficients of a image by using the Benford’s Law. Firstly, we proposed a scheme to analyse DWT coefficients and JPEG2000 compressed images using the Benford’s Law for image forensic applications. The uncompressed DWT coefficients were found to obey the Benford’s Law based on 1338 test images. We also analysed the compressed DWT coefficients with different compression rates for JPEG2000 images. The results indicated that the compressed DWT coefficients still obeyed the Benford’s Law with some slight difference between them.
For instance, the mean divergence for JPEG2000 compression rate at 0.1 was 0.0108, which was much higher than uncompressed DWT coefficients. Hence, from these initial results, we can estimate a presence of JPEG2000 compression and could further analyse to estimate the unknown JPEG2000 compression rates in the image. Secondly, we proposed an image forensics technique of extracting images with glare feature. We found via experiments that 142 images have this irregularity from 1338 images, and the proportion of glare feature (unbalance light) in an image could affect the magnitude of digit 5 in 1st digits probabilities. In order to further analyse irregular Benford's Law curve, the intensity distribution histogram of test images was also calculated for comparison. We found a good correlation between the 1st digits probabilities and intensity distribution of its gray level. Furthermore, we simulated histogram equalization attack on the test images, and results showed our proposed method based on 1st digits probabilities are not affected by the histogram equalization attack, and the glare featured images could be detected more effective than analysing graylevels distribution of the images. Presently, We are planning to improve our two image forensics schemes. The coded stream of JPEG2000 can be analysed to further improve the accuracy of detecting unknown JPEG2000 compression rates, as well as using the proposed method to accurately estimate double compression in JPEG2000 images. For detecting glare featured images, more tests will be performed on images under different natural light conditions, and further research with Fourier Transform instead of DWT for comparative analysis will also be considered. Finally, we plan to further analyse the characteristic of parent and children coefficients relationship of WBCT that could be utilised for image forensics.
Bibliography


Appendix 1

Three book chapters, one international journal and six international conference papers have been accepted and published, as listed below:


16th International Conference on Digital Signal Processing (DSP2009), 5-7 July 2009, Santorini, Greece


- Duan, G., Ho, A.T.S., Zhao, X, “A Novel Non-Redundant Contourlet Transform for Robust Image Watermarking Against Non-Geometrical and Geometrical Attacks,” IET 5th International Conference on Visual Information Engineering (VIE08), Xi’an, China, 29 July - 1 August 2008

Appendix 2

In the Appendix, we attached our two conference papers that did not discussed in the thesis, and a recent book chapter.
Abstract—Whilst it is sometimes essential that a scene is well lit before image capture, too much light can cause exposure or glare-based problems. Typically, glare is introduced to images when the camera is pointed towards the light source, and results in a visible distortion in the image. In this paper, we analyse and identify images that contain the ‘glare’ property using the empirical Benford’s Law. The experiment is performed on 338 images, and extracts discrete wavelet High High (HH), High Low (HL) and Low High (LH) sub bands as raw data. The significant digit from each coefficient of all sub bands is then calculated. We then analyse the probability of occurrence of large digits against smaller digits to detect anomalies. All images containing these anomalies are further analysed for the identification of additional salient features. This analysis is performed in accordance with the Benford’s Law plot and helps in probability intensity histogram and divergence. Our results indicate that 142 images have irregular Benford’s Law curves. For most images, the irregularity occurs at the 5th digit. After visual examination, we have found the unbalanced light and high level of brightness in these images. To measure the intensity of light in an image, we compute the probability histogram of gray levels. These results also correlate with the irregular peak identified from the Benford’s Law curves. In addition, the divergence is then computed, which shows the deviation between the actual Benford’s Law curve and the Benford’s Law graph of an image. Our proposed technique is novel and has a potential to be an image forensic tool for quick image analysis.

I. INTRODUCTION

The field of digital image forensics is striving hard to restore images lost in digital content. Images, which are becoming a part of today’s life, are growing vulnerable to digital forgery [1]. The readily available image manipulation software, such as Photoshop and GIMP are primarily responsible for this. A statistical law called, Benford’s Law [2]–[5], has been used previously in accounting forensics to detect fraudulent data by Nigrini [5]. Similarly, another study in the field of psychology noticed a peak at the digit 5 when people are asked to choose from a tampered set of data [6]. In recent years, it has attracted the attention of image processing experts.

In 2005, Acebo et al [7], showed how light intensity in an image can be used to determine if an image is genuine or computer generated. Unlike [7], we have applied DWT before calculating 1st digit probabilities and analysed images in the frequency domain, which is better to separate edge details from low frequencies. Similarly, Fu et al [8] have applied Benford’s Law on DCT coefficients in order to detect unknown JPEG compression. In our previous work [9], we have analysed DWT coefficients using Benford’s Law and audited the processing history applied to JPEG2000 images. We have noticed a sharp peak at digit 5 in Benford’s Law curve for some images, which became the basis for this paper.

Therefore, we propose a novel use of the Benford’s Law to identify unbalanced lighting in an image with the help of DWT. In this paper, we will analyse the irregularity in the Benford’s Law curve that appears for images with a certain feature. This feature is ‘glare’, when appeared in or near a field of view induces unbalanced light that makes an image comparatively brighter in various parts. In effect, the image loses its visual quality. To determine if an image possesses unbalanced lighting, we will apply a single level of DWT to an image and compute its 1st digit probabilities of DWT coefficients. Images with irregular Benford’s Law curves are then identified and analysed further. The divergence, adopted from Fu et al [8] and Acebo et al [7], which shows how much the Benford’s Law curve of an image deviates from the actual Benford’s Law, is also calculated. Furthermore, the intensity histogram, used for measuring the strength of gray levels in various sections of an image, will compare with its 1st digit probabilities.

The rest of the paper is organised as follows: Section II, will briefly describe Benford’s Law, Discrete Wavelet Transform and intensity distribution for gray level images. Section III illustrates our proposed glare image detection method, followed by our experimental results and analysis in Section IV. This is followed by conclusion and future work in Section V.

II. BACKGROUND

A. Benford’s Law

Benford’s Law, also known as the 1st digits law and significant digits law, was introduced by Frank Benford in 1938 [3]. Then, Hill [4] expressed Benford’s Law as a logarithmic distribution, for analysis of the probability distribution of the 1st digit (1 – 9) of the number from natural data in
tatistics. The distribution for Benford’s Law can be expressed by Equation 1.

\[ p(x) = \log_{10} \left( 1 + \frac{1}{x} \right), \quad x = 1, 2, \ldots, 9 \quad (1) \]

Where \( x \) is the 1st digit of the number and \( p(x) \) is the probability distribution of \( x \). A typical probability distribution of Benford’s Law is shown in Figure 1. Any peaks or irregularities in the curve mean that the data has been tampered with or is unnatural [5], [6], [7].

\[ p_{r_k} = \frac{n_k}{M N}, \quad k = 0, 1, 2, \ldots, L - 1 \quad (2) \]

Where \( M \) and \( N \) are the total number of pixel in an image, \( n_k \) is the number of pixels that have intensity \( r_k \), and \( L \) is the possible intensity levels in the image (256 for 8 bit image).

## III. PROPOSED METHOD

In this section, we explain the proposed glare image detection process which uses the Benford’s Law model. As shown in Figure 3, the test image is first decomposed into LL, LH, HL, and HH sub-bands through single level DWT (2, 3 and 4 levels of DWT can also be applied, but could decrease the detection rates). In our experiment, we used the wavelet with Daubechies 9/7 filter, which is standard for lossy JPEG2000 compression. Secondly, the DWT coefficients of LH, HL and HH sub-bands are extracted (LL sub-band is not used as it contains highest amount of energies and is more sensitive to low frequencies [11]), and then the probability distribution of the 1st digits of these DWT coefficients is obtained by using the Benford’s Law model, as shown in Equation 1. Finally, in order to detect glare, these 1st digit probabilities undergo an image retrieval process, as per the pseudocode below.

\[
\text{if } p(x) < p(x + 1) \text{ then}
\]

This 1st digit’s probability does not obey Benford’s Law

This image could have glare

end if

In addition, with the purpose of further investigate the relationship between the 1st digit probability and intensity distribution, the intensities of the image are calculated using Equation 2, which estimate the strength of gray levels present in an image in the spatial domain. Moreover, the divergence is also calculated, which measures how much deviation of the 1st digit probability graph from the actual Benford’s Law curve. The divergence adopted from Acebo et al [7] and Fu et al [8], is shown in Equation 3.

\[
x^2 = \sum_{i=1}^{9} \left( \frac{p'_i - p_i}{p_i} \right)^2, \quad i = 1, 2, \ldots, 9 \quad (3)
\]

Where \( p'_i \) is the actual 1st digit probability of the DWT coefficients and \( p_i \) is the Benford’s Law from Equation 1.
Our proposed glare image detection method has detected 42 (128, 123 and 121 if 2, 3 and 4 levels of DWT applied, respectively) out of 1338 images from UCID [10], which had irregular Benford's Law curves. Figures 4(b) and 6(b) show how two irregular Benford's Law curves of images 'Statue 1' (Figure 4(a)) and 'Street' (Figure 6(a)), which compared with the actual Benford's Law curve. By comparing with the regular Benford's Law curve of image 'Statue 2' (Figure 5(a)) visually, the unbalance light caused by glare can also be found from Figures 4(a) and 6(a), which possesses extra amount of brightness in some parts. Moreover, the intensity distribution histograms of a normal image 'Statue 2', in Figure 5(c), is somewhat smoother than the glare images 'Statue 1' and 'Street', in Figures 4(c) and 6(c).

In image 'Statue 1', due to its glare affect, the top boundary of the image can not be found visually, which also correlates with its 1st digit probabilities, which do not obey Benford's Law. Especially, its irregular peak at digit 5 in Figure 4(b), which is nearly five times higher. Similarly, the top boundary of image 'Street' is also difficult to find visually and the amount of its unbalanced light is much less than the image 'Statue 1'. Furthermore, in Figure 6(b), the irregular peak at digit 5 from its 1st digit probabilities is only twice as high. In addition, the intensity distribution histogram in figure 6(c), is smoother than image 'Statue 1' in Figure 4(c). Therefore, we conclude that the amount of unbalanced light in an image could affect the magnitude of the digit 5 in the probabilities. In contrast, the distribution of gray levels in the intensity histogram in Figure 5(c) for image 'Statue 2', is spread across a wider range of gray levels smoothly. In Figure 5(b), the image obeys Benford's Law and there appears to be no significant brightness in the image. The divergence between these 1st digit probabilities and Benford's Law is shown in Table I.

<table>
<thead>
<tr>
<th>Table I</th>
<th>Divergence from Benford's Law curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>Divergence</td>
</tr>
<tr>
<td>Statue 1</td>
<td>1.7590</td>
</tr>
<tr>
<td>Statue 2</td>
<td>0.0024</td>
</tr>
<tr>
<td>Street</td>
<td>0.0919</td>
</tr>
</tbody>
</table>

From Table I, we observe that the divergence of image Statue 2' is showing the best fit to Benford's Law, at 0.0024, in comparison, the worst fit at 1.759 is image 'Statue 1' and followed by image 'Street' at 0.0919. Furthermore, by deducing common artefacts that are present in both sets of images that follow or do not follow Benford's Law, the only feature left is glare (extra brightness). Hence, we can conclude that the glare feature in an image results in an irregularity in its 1st digit probabilities.

In addition, we also simulated an attack by appending artificial glare into the images. The attack is implemented by using "lens flare filter" function with the brightness increased 135% in Adobe Photoshop CS5 image editing software. From this experiment, the results showed that both regular and irregular 1st digit probabilities are not influenced by the artificial glare, and therefore our method could also identify whether there is natural or artificial glare present in an image.

V. Conclusion and Future Work

In this paper, we presented a method of extracting images with glare (such as unbalanced lighting) out of the bulk of images in DWT domain using Benford's Law. Any irregularities in an image could be detected as peaks via the Benford's Law curves. The peaks were mainly located at digit 5. We found via experiments that 142 images have this irregularity from 1338 images, and the amount of glare feature (unbalanced light) in an image could affect the magnitude of digit 5 in 1st digit probabilities. In order to further analyse the irregular Benford's Law curve, the intensity distribution histogram of test images was also calculated for comparison. We found a good correlation between the 1st digit probabilities and intensity distribution of its gray level. The divergence was also calculated between the 1st digit probability curve and Benford's Law, such as image 'Statue 2', at 0.0024. Our method could also identify whether there is natural or artificial glare present in an image. In future work, more tests will be performed on images under different natural light conditions. Further research with Fourier Transform instead of DWT for comparative analysis will also be considered.

Acknowledgment

Ghulam Qadir would like to acknowledge the financial support of COMSATS Institute of Information Technology, Islamabad, Pakistan for the sponsorship of his PhD scholarship.

References

Fig. 4. (a) Statue 1, (b) 1st digit probabilities & Benford's Law, (c) Probability Histogram

Fig. 5. (a) Statue 2, (b) 1st digit probabilities & Benford's Law, (c) Probability Histogram

Fig. 6. (a) Street, (b) 1st digit probabilities & Benford's Law, (c) Probability Histogram
Estimating JPEG2000 Compression for Image Forensics Using the Benford’s Law

Ghulam Qadir, Xi Zhao and Anthony TS Ho
Dept. of Computing, Faculty of Engineering and Physical Sciences,
University of Surrey,
Guildford, Surrey, GU2 7XH, UK

ABSTRACT

With the tremendous growth and usage of digital images nowadays, the integrity and authenticity of digital content is becoming increasingly important, and a growing concern to many government and commercial sectors. Image Forensics, based on a passive statistical analysis of the image data only, is an alternative approach to the active embedding of data associated with Digital Watermarking.

Benford’s Law was first introduced to analyse the probability distribution of the 1st digit (1-9) numbers of natural data, and has since been applied to Accounting Forensics for detecting fraudulent income tax returns [9]. More recently, Benford’s Law has been further applied to image processing and image forensics. For example, Fu et al. [5] proposed a Generalised Benford’s Law technique for estimating the Quality Factor (QF) of JPEG compressed images. In our previous work, we proposed a framework incorporating the Generalised Benford’s Law to accurately detect unknown JPEG compression rates of watermarked images in semi-fragile watermarking schemes. JPEG2000 (a relatively new image compression standard) offers higher compression rates and better image quality as compared to JPEG compression. In this paper, we propose the novel use of Benford’s Law for estimating JPEG2000 compression for image forensics applications. By analysing the DWT coefficients and JPEG2000 compression on 1338 test images, the initial results indicate that the 1st digit probability of DWT coefficients follow the Benford’s Law. The unknown JPEG2000 compression rates of the image can also be derived, and proved with the help of a divergence factor, which shows the deviation between the probabilities and Benford’s Law.

Based on 1338 test images, the mean divergence for DWT coefficients is approximately 0.0016, which is lower than DCT coefficients at 0.0034. However, the mean divergence for JPEG2000 images compression rate at 0.1 is 0.0108, which is much higher than uncompressed DWT coefficients. This result clearly indicates a presence of compression in the image. Moreover, we compare the results of 1st digit probability and divergence among JPEG2000 compression rates at 0.1, 0.3, 0.5 and 0.9. The initial results show that the expected difference among them could be used for further analysis to estimate the unknown JPEG2000 compression rates.

Keywords: Image Forensics, JPEG2000, Benford’s Law, DWT, DCT, JASPER

1. INTRODUCTION

As digital imaging devices such as digital cameras, camcorders and scanners have become very popular and widely available in the market place; the use of digital images has grown considerably in our daily life. This, coupled with the ease-of-use and effectiveness of advanced image manipulation software, has made altered images ubiquitous in the digital world, raising a concern regarding the integrity of images. In order to restore trust in an image, the field of image forensics has been developed to analyse images based solely on the image data itself. The primary focus of image forensics is to detect and authenticate any kind of manipulation in a digital image. Image forensics can be viewed as an alternate approach to digital watermarking, where secret information (watermarks) are embedded into an image to protect its authenticity. The advantage of Image forensics, however, is that it works only with the data in hand, and does not require embedding any additional information. As described in [1, 2] image forensics also isolates anomalies that might exist due to non-malicious processing (such as a change in file format) or intentional, malicious modifications
(such as cloning or creating composites), as well as identifying the difference between natural and unnatural images [3]. A natural image possesses its original characteristics, such as shape, contrast and size. On the other hand, this image becomes unnatural if any of these characteristics are changed.

Fridrich et al [2] classified image forensic techniques according to six different categories: source classification; device identification; images linking to source device [4]; processing history recovery; integrity detection; and anomaly investigation. Processing history recovery relates to the part recovery of the processing chain associated with an image [2]. This area of image forensics focuses on detecting non-malicious alterations in an image such as lossy compression (JPEG, JPEG2000), resizing, and colour/contrast adjustments. Fu et al. [5] proposed an image forensic technique to detect the Quality Factor (QF) of unknown JPEG compressed images. They found that DCT coefficients of an image obey the Benford's Law distribution closely, and that the 1st digit probability distributions do not follow the Generalized Benford's Law if the image has been compressed twice with different QFs. Hence, the actual QF can be accurately estimated by analysing its JPEG coefficients according to the Generalised Benford's Law [6-8]. Benford's Law is a statistical model of probabilities [8], used originally in accounting forensics [9] to detect financial frauds. We also adapted this approach to determine adaptive thresholds that could improve the authentication accuracy in semi-fragile watermarking [10].

Zhang et al. [11] proposed a double compression detection technique for JPEG2000 compressed images. Double compression occurs when an image is saved twice in same format with different or similar compression. In their scheme, they applied the Discrete Wavelet Transform (DWT) to a JPEG2000 compressed image, to extract the High/Low and Low/High sub-bands of the DWT coefficients. A histogram was then formed by applying the Fast Fourier Transform (FFT) to these extracted coefficients. By analysing the sharp peaks and valleys of this histogram, this test image could be classified according to whether or not it has been subjected to double JPEG2000 compression. There currently exists no literature regarding the analysis of single JPEG2000 compression for image forensics. In this paper, we propose the novel use of Benford's Law for single JPEG2000 compressed images. We will analyse the DWT coefficients of uncompressed and JPEG2000 compressed images based on the Benford's Law. A comparative evaluation of compression rates will also be investigated. In contrast with Fu et al. [5], we analyse the DWT coefficients instead of DCT coefficients before the quantisation step. This is due to the fact that the quantization has no effect in the compression process in a JPEG2000 compression coding system, which will be explained in the next section [12].

The remainder of the paper is organised as follows. Section 2 describes the background of JPEG2000, DWT and the Benford’s Law. In this section, a brief description of the JPEG2000 core coding system is given, along with an explanation of DWT, and the advantages of using DWT over DCT in JPEG2000 compression. It also provides a brief discussion of why Benford’s Law is a useful tool for image forensics. Section 3 discusses the results obtained using the proposed method of utilising Benford’s Law to analyse DWT coefficients of both uncompressed and JPEG2000 compressed images. Finally, Section 4 presents the conclusions and future work.

2. BACKGROUND

2.1 JPEG2000 Compression

In the literature, there are two major classes of image compressions: lossy and lossless [13]. Lossy image compression produces an image with acceptable visual quality but with a significantly smaller file size. One of the most popular lossy compression techniques is the JPEG (Joint Photographic Experts Group) format, which was implemented based on the block Discrete Cosine Transform (DCT) [13]. In JPEG, the compression is achieved by applying the quantization and coding process to the DCT coefficients of an image. The purpose of the quantization step is to remove redundancies of the image data with a Quality Factor (QF), that represents different compression rates. However, the JPEG compressed image can possess blocking artefacts at low quality factors, due to the use of the block DCT. To overcome this, a newer version of JPEG was introduced, JPEG2000 (file extensions jp2 or j2k) [14, 15]. JPEG2000, based on the Discrete Wavelet Transform (DWT) [12], has better quality, with no blocking artefacts; it is more complex in implementation than JPEG. A block diagram of the JPEG2000 core coding system is shown in Figure 1 [15].
As shown in Figure 1, there are three important components: DWT; Quantization; and Embedded Block Coding with Optimized Truncation (EBCOT). Firstly, DWT is applied to the image, followed by quantization. In contrast to JPEG, the quantization is performed by dividing each coefficient with the step size (which does not affect the compression rate in JPEG2000). EBCOT coding is the next step in the process, where the quantized coefficients are formed into bit-planes starting from the Most Significant Bit (MSB) to Least Significant Bit (LSB). These bit-planes then undergo three coding passes: the significance propagation pass; the amplitude refinement pass; and the cleanup pass. Then, the coded stream is achieved by arranging coded bits into quality layers according to desired compression rate. Finally, the code stream is used to reconstruct the JPEG2000 compressed image [12, 14, 15].

In this paper, we used JASPER [16], an open source implementation of JPEG2000 in C, with MATLAB for our experiments. In JASPER, compression parameters are used to apply different compression rates between 0.01 and 1, which represent the percentage size of the original image. A total of 1338 test images, from UCID [17] database are used for our experiment and analysis.

2.2 DWT for JPEG2000

As mentioned above, DWT is used in JPEG2000 for mapping spatial pixels of an image into coefficients in the frequency domain. In contrast with DCT, which divides the image into 8 by 8 blocks, the DWT is applied to the entire image, yielding a much better energy compaction while reducing discontinuities at the same time [12]. As shown in Figure 2, DWT decomposes an image into power of 2 resolution levels by applying a collection of low-pass and high-pass filters onto the image in vertical and horizontal directions. Each resolution level consists of four sub-bands, which are LL, HL, LH, and HH. In JPEG2000, the resolution levels are often between 3 to 8. JPEG2000 compression can be both lossy and lossless, depending upon the type of filter applied in DWT. For instance, Le Gall 5/3 filter is used for lossless compression, and Daubechies 9/7 filter is used for lossy compression [6, 7, 8]. In this paper we will analyse the lossy JPEG2000 images using Benford’s Law in comparison with DCT in [5].
2.3 Benford’s Law

The Benford’s Law, was introduced by Frank Benford in 1938 [7] and was later developed by Hill [8] for analysis of the probability distribution of the first digit (1-9) numbers obtained from natural data in statistics. A typical probability distribution of Benford's Law is shown in Figure 3. By using Benford’s Law, the 1st digit of natural numbers (1-9) can be classified in a specific way, that is smaller digit occurs more often than larger digit. Hill explained the law in terms of statistics, concluding that the nature of probabilities of first digits from 1 to 9 is logarithmic [8]. The distribution for Benford’s Law can be expressed by Equation 1.

\[ p(x) = \log_{10} \left(1 + \frac{1}{x}\right), \quad x = 1, 2, \ldots 9 \]  

(1)

where \( x \) is the first digits of the number and \( p(x) \) is the probability distribution of \( x \).

![Figure 3: Probability distribution of Benford’s Law](image)

3. METHOD AND RESULTS

3.1 DWT coefficients and Benford’s Law

We analyse the DWT coefficients of the uncompressed images to investigate if these coefficients follow the Benford’s Law, and the result will also be a benchmark for the next experiment on JPEG2000 images. The DCT coefficients are non-uniformly distributed; therefore, the Benford’s Law can be successfully applied to the first digits of DCT coefficients [5]. The DWT coefficients have similar property of the DCT, which is illustrated in the following figures. Figure 4 illustrated the image ‘Cameraman’ and its associated probabilities of DCT and DWT coefficients as compared with the Benford’s Law as shown in Figure 5. The ‘Lena’ image is shown in Figure 6 and the associated probabilities of DCT and DWT are shown in Figure 7.
Figure 4: image ‘Cameraman’

Figure 5: 1st digits probabilities of ‘Cameraman’

Figure 6: image ‘Lena’
The results are obtained by applying level 3 DWT of the image, and non-overlapped 8 by 8 block of DCT to uncompressed images, and then calculate the 1st digits probabilities of their corresponded coefficients. Figures 5 and 7 show that the DWT coefficients for the two test images are following the trend of the Benford’s Law. Furthermore, for the lower complexity image ‘Cameraman’ in Figure 5, the trend of DWT coefficients is closer to the Benford’s Law than DCT coefficients. One the other hand, for the higher complexity image ‘Lena’ in Figure 7, both DWT and DCT have similar trends and follow the Benford’s Law. Hence, the results implied that the 1st digits probabilities for DWT coefficients perform better in lower complexity images.

In order to substantiate the results, we conduct the experiment of evaluating the 1st digits probabilities for 1338 uncompressed grayscale images from the Uncompressed Image Database (UCID) [17]. Figure 8 illustrates the comparison between the probability distribution of Benford’s Law, and the mean distribution of 1st digit of uncompressed DWT coefficients of 1338 images. The results also show that the distribution of the 1st digits of the uncompressed DWT coefficients obeys the Benford’s Law closely. In order to evaluate how much deviation of the mean distribution to Benford’s Law, we calculate the average divergence [18], as given in Equation 2.

\[
\chi^2 = \sum_{i=1}^{9} \frac{(p_i' - p_i)^2}{p_i} \quad \text{for } i=1...9
\]

where \(p_i'\) is the actual 1st digits probability of the DWT coefficients and \(p_i\) is the 1st digits probability from Benford’s Law in Equation (1). Based on 1338 test images, the average divergence of mean probability is 0.0016, which is even lower than the divergence of DCT coefficients at 0.0034, observed by Fu et al. [5]. Therefore, the results indicate a good fitting between the probability distribution of Benford’s Law and the uncompressed DWT images.
3.2 Benford's Law and Compressed JPEG2000 Images

From the results analysed from the last section, we conclude that the uncompressed DWT coefficients follow the Benford's Law. We will further analyse the compressed DWT coefficients from JPEG2000 images. The schematic diagram of the conducted experiment is shown in Figure 9. The grayscale original image is first compressed and saved into JPEG2000 format (.jp2) via the JPEG2000 compression software, JASPER. Next, the compressed image is then saved to a different format (.bmp, .tiff). The saved image can be passed to the receiver since the compression rate is unknown. Afterwards, the receiver can read the BMP format image and apply DWT to it. Finally, the receiver calculates the 1st digits Benford's Law of this test image to detect unknown compression rate. In JASPER, compression parameters are used to apply different compression rates between 0.01 and 1, which represent the percentage size of the original image. The bits-per-pixel parameter is set at 8 bits in our experiment.

Based on 1338 images, Figures 10 to 13 show the 1st digits probabilities of compressed DWT coefficients (extract from JPEG2000 compressed image) with different compression rates, 0.1, 0.3, 0.5 and 0.9, which are compared with the Benford's Law, respectively. As we can see from the figures, most of the 1st digital probabilities obey the Benford's Law under different compression rates. However, the trends of compression rates at 0.9 and 0.5 are closer to the Benford's Law than 0.3 and 0.1, respectively. For the divergence evaluation, the mean divergence for JPEG2000 images compression rate at 0.1 is 0.0108, which is approximately 10 times higher than the compression rate at 0.9. These variations in the divergences could be used to detect the compression rate of JPEG2000 images. Hence, for JPEG2000 images, we can conclude that the 1st digits probabilities of compressed DWT coefficients follow the Benford's Law based on the different compression rates. This property could be further explored to accurately estimate unknown JPEG2000 compression in image forensics.
Figure 10: 1st digits probabilities of JPEG2000 compression rate at 0.1

Figure 11: 1st digits probabilities of JPEG2000 compression rate at 0.3
In this paper, we proposed a scheme to analyse DWT coefficients and JPEG2000 compressed images using the Benford’s Law for image forensic applications. The uncompressed DWT coefficients were found to obey the Benford’s Law based on 1338 test images. By using a divergence factor, the mean divergence for DWT coefficients was estimated to be 0.0016, which was lower than the DCT coefficients at 0.0034. These deviations indicated that the DWT coefficients followed the Benford’s Law much closer than the DCT coefficients. In our second experiment, we analysed the compressed DWT coefficients with different compression rates for JPEG2000 images. The results indicated that the compressed DWT coefficients still obeyed the Benford’s Law with some slight difference between them. For example, the mean divergence for JPEG2000 compression rate at 0.1 was 0.0108, which was much higher than uncompressed
DWT coefficients. Hence, from these initial results, we can estimate a presence of JPEG2000 compression and could further analyse to estimate the unknown JPEG2000 compression rates in the image.

For future work, we plan to improve our scheme to accurately estimate unknown compression rates for watermarked images. The coded stream of JPEG2000 can be analysed to further improve the accuracy of detecting unknown JPEG2000 compression rates, as well as using the proposed method to accurately estimate double compression in JPEG2000 images.

ACKNOWLEDGEMENTS

Ghulam Qadir would like to acknowledge the financial support of COMSATS Institute of Information Technology, Islamabad, Pakistan for the sponsorship of his PhD scholarship.

REFERENCES

Image Authentication using Active Watermarking and Passive Forensics Techniques

Xi Zhao, Philip Bateman, and Anthony TS Ho

University of Surrey, UK {x.zhao, p.bateman, a.ho}@surrey.ac.uk

1 Introduction

The primary reason for the requirement of authenticating images stems from the increasing amount of doctored images that are presented as accurate representations of real-life events, but are later discovered to be faked. The history of manipulating images reaches back almost as far as photography itself, and with the ease of use and availability of image editing software, it has become ubiquitous in the digital age. Image authentication schemes attempt to restore trust in the image by accurately validating the data, positively or negatively. Especially for law enforcement scenarios, images captured at the scene, such as for crime scene investigation and traffic enforcement, potentially be used as evidence in the court of law. If an image presented in court as evidence from a crime scene is to be effectively used by the jury, the integrity of the information must not be in question. The role of a scene of crime officer (SoCOs) is to capture, as much as possible, the left-over evidence at the crime scene by taking photographs and collecting any exhibits found. After the collection of evidence, there is no other way of examining the crime scene as a whole, apart from analysing the collected exhibits and photographs taken [1]. In order to maintain the integrity of the images, not only it is essential to verify that the photographic evidence remains unchanged and authentic, but any manipulated regions should also be localised to help identify which parts of the image cannot be trusted. With the tremendous growth and usage of digital cameras and video devices, the requirement to verify the digital content is paramount, especially if it is to be used as evidence in court [2]. Therefore, digital watermarking technique can be utilised for image content authentication applications to verify or authenticate the integrity of the digital media content.

Digital watermarking is the process of embedding relevant information (such as a logo, fingerprint and serial number), into a media. This technique can be applied to different media types such as video, audio and image content. An example of digital visible watermark is the translucent logos that are often seen embedded at the corner of videos or images, in an attempt to prevent copyright infringement. However, these visible watermarks can be targeted and removed rather simply by cropping the media,
or overwriting the logos. Subsequently, the field of digital watermarking is primarily focused on invisible watermarks, which are imperceptible and operate by tweaking the physical data of the media [3, 4]. There are three different classifications associated with digital watermarking, depending on the applications: robust, fragile and semi-fragile. Robust watermarking is primarily designed to provide copyright protection and proof of ownership for digital images. The most important property of robust watermarking is its ability to tolerate certain signal processing operations that usually occur during the lifetime of a media object, as well as preventing any more deliberate attacks.

Fragile and semi-fragile digital watermarking techniques are often utilised for image content authentication. Fragile watermarking schemes are designed to detect any possible manipulations that affect the watermarked image pixel values [5, 6]. In comparison, while fragile watermarking is aptly named because of its sensitivity to any form of attack, semi-fragile watermarking is more robust against attack, and can be used to verify tampered content within images for both malicious and non-malicious manipulations [7–9]. In addition, semi-fragile schemes make it possible to verify the content of the original image, as well as permitting alterations caused by non-malicious (unintentional) modifications such as system processes. Moreover, semi-fragile watermarking is more focused on detecting intentional attacks than validating the originality of the image [10, 11]. During the image transmission, the mild signal processing errors caused by signal reconstruction and storage, such as transmission noise or JPEG compression, are permissible. However, the image content tampering such as copy and paste attack will be identified as a malicious attack. Additionally, in the literature, a significant amount of research has been focused on the design of semi-fragile algorithms that could tolerate JPEG compression and other common non-malicious manipulations [12–18]. However, watermarked images could be compressed by unknown JPEG compression rates of various quality factors (QFs). As a result, in order to authenticate the images, these algorithms have to set a pre-determined threshold that could allow them to tolerate different QF values when extracting the watermarks. To determine the threshold more accurately, the generalised Benford’s law can be utilised to estimate the unknown JPEG compression QF; then appropriate thresholds could be adapted for each test image, before initialising the watermark extraction and authentication process. This law has already been successfully used in image forensics technique for JPEG compression evaluation [19]. This adaptive threshold could help to decrease the false alarm and missed detection rates.

In contrast to authenticate the image using active watermarking technique, the image forensics as passive technique has attracted much attention [20–22]. The significant difference is that image forensics seeks to authenticate images based solely on the image data provided in image statistical analysis, meaning it is a passive approach to the problem. As such, no embedded information is loaded into an image, and so the security risks and robustness issues associated with a payload, are avoided. Therefore, image forensics presents itself as an alternative approach to the active insertion of watermarking data to authenticate images. In this chapter, we will review active watermarking techniques, such as fragile and semi-fragile methods as well as
passive image forensics techniques such as camera identification and forgery detection methods for image authentication. Furthermore, we will introduce our three proposed image authentication related methods, which are fragile watermarking scheme in Slant transform (SLT) domain, utilising the generalised Benford’s Law as image forensics technique to improve semi-fragile watermarking technique and the use of the statistical process control (SPC) for camera identification in image forensics research.

The chapter is organized as follows:

- In Section 2, several fragile and semi-fragile watermarking schemes will be reviewed. Our proposed SLT semi-fragile watermarking algorithm is then introduced. The watermark embedding, detection and authentication processes are described in detail as well as the proposed experimental results are analysed and evaluated by comparing with two other transform based scheme, which in Discrete Cosine transform (DCT) and Pinned Sine transform (PST) domain.
- Section 3 discusses three typical methods of employing predetermined thresholds in semi-fragile watermarking algorithms and the limitations of using predetermined thresholds were highlighted from the literature. Then we proposed a framework incorporating the generalised Benford’s Law that could detect unknown JPEG compression QFs in semi-fragile watermarked images to adjust the appropriate threshold with experimental results.
- Section 4 will review image forensics techniques that focus on two main areas, camera identification and image forgery detection and their applications. Then we propose to utilise SPC methods to analyses images captured from different digital camera devices.
- Section 5 gives the conclusion of this chapter and presents some directions for future work of the research.

2 Fragile and Semi-fragile Watermarking

In this section, both fragile and semi-fragile watermarking algorithms for image authentication are reviewed. A detailed discussion on our proposed semi-fragile watermarking schemes in SLT domain to further explain the concept of semi-fragile watermarking is also presented. The results of miss detection rates and false alarm rates are then compared with two existing transforms based on the DCT and PST transforms.

2.1 Literature Review for Fragile and Semi-fragile Watermarking

Fragile Watermarking

As mentioned in Section 1, fragile watermarking schemes should be able to detect any possible manipulations that affect the watermarked image any pixel values. Therefore, it is possible to exploit the inherent weakness of the LSB schemes,
and implement a fragile watermarking scheme in the spatial domain. Fridrich [23] proposed a spatial domain based fragile watermarking scheme that could localise tampered regions of a watermarked image, by adapting Wong's method [24]. The watermark embedding process is shown in Figure 1. The original image is first divided into non-overlapping blocks of equal size 8 by 16. In each block, the seven Most Significant Bits (MSB) of each pixel are extracted, and a cryptographic hash function is applied as illustrated in Figure 2. The logo is also divided into 816 blocks and each block contains information about the original block position, image index, original image dimensions (resolution), camera ID and author ID (PIN). The hashed seven MSBs of each block and its corresponding logo block are subjected to an Exclusive-OR (XOR) operation and then encrypted using a key. Finally, the LSBs of the original image are replaced with the result of the XOR operation and encrypted watermark bits, and the watermarked image is created. In the authentication process, the LSBs of the test image are extracted, and the seven MSBs from each block are hashed as shown in Figure 3. For each block, the LSBs are decrypted with a key, along with its corresponding hashed seven MSBs using the XOR operation. Finally, the authentication process itself is achieved by comparing each block of the image with the corresponding block from the logo. If this set of the block is not the same, the block of the image is flagged as a tampered block.

Fig. 1. Fridrich's fragile watermark embedding algorithm.

```
1 1 0 1 1 0 1
```

Fig. 2. MSBs and LSB of pixel value 221 in 8 bits binary sequence.
Zhang and Wang [25] proposed a statistical scheme of fragile watermarking scheme that embed a folded version of the authentication data derived from five most significant bits (5MSBs) of the original image along with other additional data into the image with acceptable watermarked image quality PSNR as 37.9dB. Their results showed their algorithm could localized the tampered pixels accurately. Then they further improved their method in [25] that could restore the tampered image content after localized the tampered area without any errors [26]. He et al. [27] proposed a conventional self-embedding fragile watermarking scheme based on adjacent-block based statistical detection method (SDM) that could against copy-paste attack and collage attack. Their algorithm could identify the tampered blocks with a probability more than 98% even the tampered area is up to 70% of the host image.

Fragile watermarking scheme can also be applied in transform domain. Li and Shi [5] proposed a fragile watermarking algorithm in Discrete Wavelet Transform (DWT) to achieve the requirements of high security, low distortion, and high accuracy of tamper localization for authenticating JPEG2000 images. Their algorithm could also tolerate vector quantization attack, Holliman-Memon attack, collage attack and transplantation attack. Aslantas et.al [28] proposed intelligent optimization algorithms (IOA) to improve fragile watermarking schemes in discrete cosine transform (DCT) domain. They used IOA which including four genetic algorithm (GA), clonal selection algorithm (CSA), particle swarm optimization (PSO), and Differential Evolution (De) to correct rounding errors caused by transforming an image from the frequency domain to the spatial domain with the objective of improving DCT-based fragile watermarking. The experimental results showed that the CSA produces better PSNR results whereas DE has lower computational time than other algorithms. Yeh and Lee [29] proposed reversible fragile watermarking by utilizing the pyramidal structure method. They select appropriate embedding areas by analysing the pyramid-structure of the image for embed watermark bits in wavelet domain. The experimental results showed that their scheme could successfully localized even when 50% of the watermarked image is tampered as well as detect counterfeiting attack.
Many semi-fragile watermarking techniques have been already proposed by researchers. Lin et al. [31] proposed embedding algorithm that first applied Discrete Cosine Transform (DCT) to 16 by 16 blocks of the cover image, then embed the watermarks in middle to low frequency (except DC coefficient) of each block. Their scheme could identify the tampered area with 75% accuracy under moderate compression and with near 90% accuracy under light compression. Ho et al. [7] proposed a semi-fragile watermarking scheme in Pinned Sine Transform (PST) domain. In their algorithm, the original image is applied by using PST to get the pinned and boundary fields in 8 by 8 blocks. The watermark bits are then inserted into middle to high frequency of each block in the pinned field. The scheme also used a self-restoration method, originally proposed by Fridrich and Goljan [33] to recover the tampered regions. Their scheme could tolerate some common image processing manipulations such as JPEG and wavelet compression, and the detection rate is higher than DCT-based scheme. The algorithm has been further improved by using irregular Sampling instead of the LSB method [15], which aimed to improve the robustness of tampering restoration. Kundur and Hatziankos [18] proposed a DWT based algorithm called telltale tamper-proofing, which made it possible to determine tampered regions in multi-resolutions. Unlike other schemes that use DCT, this method does not require a block division process to detect the tampered regions due to the localisation ability of the wavelet transform. The localization ability of the wavelets in both spatial and frequency domains would potentially indicate a good candidate for semi-fragile watermarking.

Maeno et al. [34] presented two algorithms that focused on signature generation techniques. The first algorithm used random bias to enhance the block based DCT watermarking scheme proposed by Lin and Chang [12]. The second algorithm used nonuniform quantisation on a non-block based semi-fragile watermarking scheme in the wavelet domain. Their experimental results showed their method was fragile to malicious manipulations, but robust to non-malicious manipulations such as JPEG and JPEG2000 compression. Ding et al. [35] also proposed a method by using DWT. In their algorithm, chaos was used to generate a pseudo-random sequence as a watermark, in an effort to improve the overall security. This made an improvement to the more traditional methods of generating a pseudo-random sequence. The subbands \((H_L^1, L_H^2, H_H^2)\) were used for embedding the watermark after applying a 2-level wavelet decomposition of the original image. The normalized cross-correlation (NC) was used to evaluate their algorithm by comparing between the original watermark and the extracted watermark after applying JPEG compression and Additive white Gaussian noise (AWGN) manipulations. Ni et al. [30] proposed a robust lossless data hiding technique that could be employed into semi-fragile watermarking scheme. The different bit-embedding strategies for groups of pixels with different pixel grayscale value distributions and error correction codes are utilized in their scheme. They analyzed their results into two modules, which are lossless and lossy. If the watermarked image has experienced losslessly compression, the watermark bits can be extracted correctly and the image will be classified as authentic and the
original image can be recovered exactly. If this losslessly compressed watermarked image has been further undergone lossy compression, the original image will not be able to be recovered and will be rendered authentic as long as the compression is not so severe that the content has been changed.

2.2 Proposed Slant Transform (SLT) Semi-Fragile Watermarking

This section will discuss our proposed method [36] in detail, which consist of the embedding, detection and authentication processes associated with watermarking.

Slant Transform (SLT)

The Slant Transform has been applied to image coding in the past [37] and was recently adopted for robust image watermarking [38]. The SLT can be considered as a fast computational algorithm provides a significant bandwidth reduction and result in a lower mean-square error for moderate size image blocks [37]. In addition, for textured images, the quality of the Slant Transformed images is higher than images coded by using other transforms such as DCT and Hadamard [39]. Moreover, as a similar image processing application to Walsh-Hadamard transform, Slant transform can be identified as a sub-optimum for energy compaction, which is essential for digital watermarking as the robust information hiding can be ensured by capitalizing the spread of middle to higher frequency bands. Furthermore, Slant transform is simpler, faster and especially suitable for highly textured images [38]. Hence, the Slant Transform is proposed for semi-fragile watermarking and authentication of images in this section. The authentication as the method to corroborate the genuineness of an object is mainly focusing on examining whether the image has been tempered or not, the location(s) of tampered region(s) and to what extent it has been changed can also be identified. Furthermore, the SLT can also be used for compressing the original image [39], providing a means to self-recovering the tampered regions by embedding the compressed cover image into the LSBs of the watermarked image [33]. The forward and inverse of SLT [37–39] can be expressed as follows:

$$[V] = [S_N]^[U][S_N]^T \quad [U] = [S_N]^T[V][S_N]$$

where $[U]$ represents the original image of size $N \times N$, $[V]$ represents the transformed components and $[S_N]$ is the $N \times N$ unitary Slant matrix given by

$$[S_N] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -a_N & b_N & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} S_{N/2} & 0 \\ 0 & S_{N/2} \end{bmatrix}$$

where $I_{(N/2)−2}$ is the identity matrix of dimension $(N/2) − 2$ and
Xi Zhao, Philip Bateman, and Anthony TS Ho

\[ a_{2N} = \left( \frac{3N^2}{4N^2 - 1} \right)^{1/2}, \quad b_{2N} = \left( \frac{N^2 - 1}{4N^2 - 1} \right)^{1/2} \]

are constants.

**Watermark Embedding**

A novel semi-fragile Slant Transform digital watermarking method is adopted based on previous work relating to PST [7] and self-restoration method [33]. The entire embedding process using the Slant Transform is illustrated in Figure 4, which consists of two parts. The first 7 bits of the cover image are extracted and divided into 8 × 8 blocks, SLT method is then applied to each block. The watermark embedding algorithm is then utilised, which is illustrated in the pseudo-code below. The watermarks for each block are then random generated by input a key as a seed. The obtained watermarks are embedded into the midband of each 8 × 8 block. After watermark embedding, frequency coefficients of each block of the watermarked image are converted back by using the inverse Slant Transform. Consequently, the first 7 bits of final watermark image is obtained. The SLT watermark embedding algorithm in pseudo-code form is shown as follows:

```plaintext
If \( w = 1 \) And \( x \geq \tau \), Then \( y = x \), Else \( y = \alpha \).
If \( w = 0 \) And \( x < -\tau \), Then \( y = x \), Else \( y = -\alpha \).
```

where \( w \) is the watermark bit, \( x \) is the SLT coefficient of the host, \( y \) is the modified SLT coefficient, \( \tau \) is the threshold which controls the perceptual quality of the watermarked image and \( \alpha \) is a constant. Similar to part 1, the original image is divided into 8 × 8 sub-blocks and also undergoes the same Slant Transform; compression for each sub-block is then achieved by discarding the high frequency coefficients. Accordingly, 64 bits information for each block is acquired after compression and then encrypted by utilizing a key as a seed. Obtained blocks are then shuffled, e.g. the value of block 1 moves to block 50, the value of block 35 moves to block 10. Therefore, LSBs of the final watermark image are then gained. Finally, the combination of part 1 and part 2 forms the final watermarked image and the key file is generated, which contains information that mentioned previously.

**Watermark Detection, Authentication and Restoration**

The proposed semi-fragile Slant Transform for image authentication and restoration method is shown in Figure 5. Similar to embedding process, the first 7 bits of the test image are extracted and divided into 8 × 8 blocks by applying SLT and then apply the detection algorithm to the first 7 bits, which is explained in the paragraph below. Meanwhile, the LSBs are extracted from the test image and only the LSBs of the detected regions are quantized back for recovery by according to authentication result. Consequently, authenticated and recovered images can be output.

The watermark bits can be detected by extracting the watermarked coefficients \( y \). If \( y \) larger than 0, the watermark bit value is 1; if \( y \) smaller than 0, the watermark
bit value is 0. The retrieved watermark needs to be compared with the watermark that exists in the key file. After the watermark bits from the entire block have been retrieved, the comparison between the watermark bits can be accomplished by using the correlation coefficient $\rho$, computed as follows:

$$\rho = \frac{\sum \sum (w' - \bar{w}) (w - \bar{w})}{\sqrt{\sum \sum (w' - \bar{w}')^2 \sum \sum (w - \bar{w})^2}}$$

where $w$ is the original and $w'$ is the retrieved watermarks corresponding to the block. For error correction, the correlation coefficient $\rho$ can be compared with a predetermined threshold value $\lambda$. If $\rho < \lambda$, which indicates that the block has been tampered as authentication, and which is followed by restoration of the tampered regions based on the decompression and extraction of the LSBs for the watermarked image.

### 2.3 Results & Evaluation

A number of experiments have been carried out to evaluate the performance of the proposed SLT watermarking scheme. The proposed watermarking scheme is compared with two other watermarking schemes: the PST-based [7] and the DCT-based [31]. For a fair comparison, the embedding strength of the watermark in each scheme is adjusted such that the peak signal-to-noise ratio (PSNR) of the watermarked images is around 33 dB, which is subjectively considered as acceptable. The performance of the watermarking schemes is measured in terms of the false positive detection rate ($P_{FP}$), false negative detection rate ($P_{FN}$) and the average detection rate ($P_{avg}$), defined as:
Fig. 5. Our proposed SLT watermark detection, authentication and restoration process.

$$P_{FP} = \frac{\text{Number of pixels in the untampered region as detected as tampered}}{\text{Total number of pixels in the untampered region}}$$

$$P_{FN} = \frac{\text{Number of pixels in the tampered region as detected as untampered}}{\text{Total number of pixels in the tampered region}}$$

and

$$P_{avg} = \left(1 - \frac{P_{FP} + P_{FN}}{1 + N}\right) \times 100. \quad (3)$$

where $N$ is the number of area(s) have been tampered with. A number of standard test images are used in the experiments and the results for 6 images, each of size $512 \times 512$ are reported.

**JPEG Compression Attack**

Table 1 shows that SLT, DCT and PST are compared by applying JPEG compression attack to 6 different grayscale images ($512 \times 512$) in order to determine the false positive rate, i.e. over detected rate. As can be seen from the table below, SLT, DCT and PST have similar error detection rates when $QF = 85$. After experiencing 75% JPEG compression attack, the over detection rate of SLT is still considerably low with average rate of 1.2, whereas PST and DCT have the higher average over detection rates of 86 and 31.4, respectively. Although the over detection rates of all three methods have increased when $QF = 65$, SLT still has the lowest increased rate of 30.8 comparing with the average value of over detection rates of PST and DCT, of 92.2 and 88.2 respectively. The reason for the relatively better results using the Slant Transform was that the embedding locations concentrated mainly in the middle frequency band, which is considered to be more robust, whereas DCT and PST mainly concentrated more on high frequencies. Overall, the results indicate that the SLT watermarking method achieves lower errors than PST and DCT based on the JPEG compression attack.
Table 1. Comparative performance of the watermarking schemes against JPEG compression with varying quality factor.

<table>
<thead>
<tr>
<th>Test Image</th>
<th>QF = 85</th>
<th>QF = 75</th>
<th>QF = 65</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SLT</td>
<td>PST</td>
<td>DCT</td>
</tr>
<tr>
<td>Lena</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Baboon</td>
<td>0.1</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Bridge</td>
<td>0.4</td>
<td>1.2</td>
<td>0.5</td>
</tr>
<tr>
<td>Trucks</td>
<td>0.2</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>Ship</td>
<td>0.2</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td>San Diego</td>
<td>0.0</td>
<td>0.4</td>
<td>0.0</td>
</tr>
<tr>
<td>Average</td>
<td>0.2</td>
<td>0.7</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Copy and Paste Attack

The copy and paste attack is utilized to compare the performance of detection rates SLT, DCT and PST for six grayscale test images (512 x 512) as given in Table 2. Three different tampering rates of 10%, 20% and 30% will be applied to each test image to analyse the overall detection rate of the three transform methods. The tamper tests are performed with 100 random locations on each image. Consequently, 5400 test images are obtained based on this experimental setup. Table 2 shows the comparative performance of the three watermarking schemes against copy and paste attack with different amount of tampering. However, the results show that PST is the most sensitive method as it has the highest overall detection rate after experiencing all three tamper tests (10%, 20% and 30%) of all images. Figure 6(a-e), shows the original, watermarked, tampered, authenticated and restored images for the image Trucks, respectively.

Table 2. Comparative performance of the watermarking schemes against copy-paste attack.

<table>
<thead>
<tr>
<th>Test Image</th>
<th>10% tamper</th>
<th>20% tamper</th>
<th>30% tamper</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SLT</td>
<td>PST</td>
<td>DCT</td>
</tr>
<tr>
<td>Lena</td>
<td>96.0</td>
<td>97.6</td>
<td>95.5</td>
</tr>
<tr>
<td>Baboon</td>
<td>96.7</td>
<td>97.3</td>
<td>96.3</td>
</tr>
<tr>
<td>Bridge</td>
<td>96.3</td>
<td>97.5</td>
<td>95.4</td>
</tr>
<tr>
<td>Trucks</td>
<td>95.7</td>
<td>97.6</td>
<td>95.1</td>
</tr>
<tr>
<td>Ship</td>
<td>96.1</td>
<td>97.6</td>
<td>95.2</td>
</tr>
<tr>
<td>San Diego</td>
<td>96.6</td>
<td>97.6</td>
<td>95.4</td>
</tr>
<tr>
<td>Average</td>
<td>96.2</td>
<td>97.5</td>
<td>95.5</td>
</tr>
</tbody>
</table>
JPEG Compression + Copy and Paste Attack

In Table 3, the six watermarked images ($512 \times 512$) are compressed with three different JPEG compression rates $QF$ of 85, 75 and 65. The experimental setup is similar to the previous copy and paste attack with 100 random locations for tampered areas. Overall, the PST achieves a relatively higher detection rate than DCT and SLT after experiencing $QF$ of 85. However, for detection, it is worse at 54.1% with $QF = 75$. In comparison, SLT has the highest overall detection rate as 91.9% at $QF = 75$ and 65. From the analysis, SLT is showed to achieve a more accurate detection result than PST and DCT. On the whole, the result indicates that the best overall detection rate among the three methods is SLT, which has 91.9% detection rate with $QF = 75$. However, all the attacked images could not be recovered by any of the three transform schemes after applying JPEG compression attack. This is due to the fact that the restoration technique is based on LSB embedding in the spatial domain of the watermarked image which is fragile. As such, it can be easily removed by JPEG compression.

2.4 Summary

In this section, we reviewed a number of different fragile and semi-fragile watermarking schemes. Our proposed SLT semi-fragile watermarking scheme was discussed in detail. The performance of the SLT based semi-fragile scheme was com-
Table 3. Comparative performance of the watermarking schemes against copy-paste attack (20% tampering) followed by JPEG compression with varying quality factor.

<table>
<thead>
<tr>
<th>Test Image</th>
<th>QF = 85</th>
<th>QF = 75</th>
<th>QF = 65</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SLT</td>
<td>PST</td>
<td>DCT</td>
</tr>
<tr>
<td>Lena</td>
<td>92.1</td>
<td>94.9</td>
<td>91.6</td>
</tr>
<tr>
<td>Baboon</td>
<td>92.1</td>
<td>94.7</td>
<td>92.1</td>
</tr>
<tr>
<td>Bridge</td>
<td>91.9</td>
<td>94.0</td>
<td>91.6</td>
</tr>
<tr>
<td>Trucks</td>
<td>92.0</td>
<td>94.5</td>
<td>91.2</td>
</tr>
<tr>
<td>Ship</td>
<td>91.5</td>
<td>93.7</td>
<td>91.9</td>
</tr>
<tr>
<td>San Diego</td>
<td>93.1</td>
<td>94.5</td>
<td>92.2</td>
</tr>
<tr>
<td>Average</td>
<td>92.1</td>
<td>94.4</td>
<td>91.8</td>
</tr>
</tbody>
</table>

pared with the PST and DCT based schemes by using average detection rate calculated from false positive and false negative detection rates. The comparative studies showed that the SLT-domain watermarking scheme performed better against JPEG compression, copy-paste attack, as well as combined JPEG compression and copy-paste attacks than the PST and DCT-domain watermarking schemes.

3 Image Forensics Technique - Benford's Law for Semi-fragile Watermarking

As mentioned in Section 1, semi-fragile watermarking scheme has been used to authenticate and localise malicious tampering of image content, while permitting some non-malicious or unintentional manipulations. These manipulations can include some mild signal processing operations such as those caused by transmission and storage of JPEG images. However, watermarked images could be compressed by unknown JPEG QFs. As a result, in order to authenticate the images, these algorithms have to set a pre-determined threshold that could allow them to tolerate different QF values when extracting the watermarks.

Figures 7 and 8 illustrate the overall relationship between the threshold, false positive and false negative detection rates. The watermarked image Lena has been tampered with a rectangular block and JPEG compressed at $QF = 75$. Figure 7(a) shows the pre-determined threshold $T = 0.5$ used for authentication. The authenticated image shows that the proposed semi-fragile watermarking scheme can localise the tampered region with reasonable accuracy, but with some false positive detection errors. In Figures 7(b) and 7(c), the lower and upper thresholds $T = 0.3$ and $T = 0.7$ were used for comparison, respectively. Figure 7(b) shows that the false positive rate has decreased whilst the false negative rate has increased in the authenticated image. Figure 7(c) shows the image has a lower false negative rate but with a higher false positive rate. From this comparison, $T = 0.5$ was chosen for JPEG compression at $QF = 75$. However, if $QF = 95$, then $T = 0.5$ may not be adequate as shown in Figure 8(a). The false negative rate is higher than Figure 8(b) with $T = 0.9$. There-
fore, it would be advantageous to be able to estimate the QF of JPEG compression, so that an adaptive threshold can be applied for increasing the authentication accuracy. In this section, we discuss our proposed method [40] to utilise the generalised Benford’s Law, as an image forensics technique to estimate the QF for semi-fragile watermarked images. The background of Benford’s Law, generalised Benford’s Law and their relationship with the watermarked image, JPEG compressed watermarked image are also described.

3.1 Benford’s Law and Generalised Benford’s Law

Benford’s Law was introduced by Frank Benford in 1938 [41] and was developed by Hill [42] for analysis of the probability distribution of the first digit (1 – 9) of the number from natural data in statistics. Benford’s Law has also been applied to accounting forensics [43, 44]. The DCT coefficients of a digital image was forward to obey Benford’s Law, it has recently attracted a significant amount of research.
interests in image processing and image forensics [19, 45, 46]. The basic principle of Benford’s Law is given as follows:

\[ p(x) = \log_{10} \left( 1 + \frac{1}{x} \right), x = 1, 2, ... 9 \] (4)

where \( x \) is the first digit of the number and \( p(x) \) is the probability distribution of \( x \). In contrast to digital image watermarking which is an active approach by embedding bits into an image for authentication, image forensics is essentially a passive approach of analysing the image statistically to determine whether it has been tampered with. Fu et al. [19] proposed a generalised Benford’s Law, used for estimating the QF of the JPEG compressed image, as shown in equation 5.

\[ p(x) = N \log_{10} \left( 1 + \frac{1}{s + x^q} \right), x = 1, 2, ... 9 \] (5)

where \( N \) is a normalisation, and \( s \) and \( q \) are model parameters [19]. Their research indicated that the probability distribution of the 1st digit of the JPEG coefficients obey generalised Benford’s Law after the quantisation. Moreover, the probability distributions were not following the generalized Benford’s Law if the image had been compressed twice with different quality factors. Thus, by utilizing this property, the QF of the image can be estimated.

Figures 9 to 11 illustrate the comparisons between the probability distribution of Benford’s Law, generalized Benford’s Law and the mean distributions of the 1st digits of block JPEG coefficients of the watermarked images compressed at \( Q_F = 100, 75, 50 \), respectively. Throughout this section we adhere to the same terminology as used in [19], where JPEG coefficients refers to the 8 x 8 block-DCT coefficients after the quantisation. These results based on 1338 images from [47] indicate a good fitting between generalized Benford’s Law and watermarked images compressed with different QFs. The results indicate that the probability distributions of the 1st digits of JPEG coefficients of the watermarked images, as shown in Figures 9 to 11, obey the generalised Benford’s Law model proposed by Fu et al. [19], in equation 5. Hence, we could employ their model to estimate the unknown QF of test images to adjust the threshold for authentication. The improved authentication process is described in the next section.

3.2 The Improved Authentication Method

In order to improve the detection rate in semi-fragile authentication process, the test image is first used for detecting the QF by the quality factor estimation process. This process works by firstly classifying the test image as compressed or uncompressed by adapting from [19]. If the test image has been compressed, the test image is then recompressed with the largest QF, from \( Q_F = 100 \) to \( Q_F = 50 \), in decreasing steps of 5. We decrease in steps of 5 as this gives us the most frequently used quality factors for JPEG compressed images (i.e. 95, 90, 85 etc.). For each compressed test image, the probability distribution of the 1st digits of JPEG coefficients is obtained. Each set
Fig. 9. 1st digit of JPEG coefficients ($QF = 100$)

Fig. 10. 1st digit of JPEG coefficients ($QF = 75$)
of values are then analysed by employing the generalized Benford's Law equation and using the best curve-fitting to plot the data. In order to obtain the goodness of fit, we calculate the sum of squares due to error (SSE) of the recompressed images. We can detect the QF of the test image by iteratively calculating the SSE for all QFs (starting at $QF = 100$, and decreasing in steps of 5), and as soon as $SSE < 10^{-6}$, we have reached the estimated QF for the test image. The threshold $10^{-6}$, was reported in [19], has been set to allow us to detect the QF of the test image, and has also been verified by the results in our experiment.

Figure 12 illustrates the results of estimating the QF for a test image that has previously been compressed with $QF = 70$. Three curves have been drawn in order to fit the three probability distribution data sets: generalized Benford’s Law for $QF = 70$, the test image recompressed with $QF = 70$, and separately recompressed at $QF = 90$. The distribution of $QF = 90$ shows the worst fit and is considerably fluctuated, while the distribution of $QF = 70$ is a generally decreasing curve, which also follows the trend of generalized Benford’s Law. These results indicate that if the test image has been double compressed without the same quality factor, the probability distribution would not obey the generalised Benford’s Law.

Once the QF is estimated, the threshold $T$ can be adapted according to different estimated QFs, based on the following conditions in Equation 6. Finally, the correlation coefficient between original watermarks and extracted watermarks for each block is compared using the attuned threshold $T$ to authenticate, in order to determine whether any blocks have been tampered with.
3.3 Results & Evaluation

The watermarked images are generated based on a simple DCT domain based semi-fragile watermark embedding scheme by using the 1338 test images from [47]. In order to achieve a fair comparison, different embedding parameters are randomised for each image such as the watermarks location, watermark string and watermark bits. For our analysis, four types of test images with and without attacks are considered as shown in Figure ??.

Table 4 summaries the results obtained for test images that have been JPEG compressed only. To evaluate the accuracy of the quality factor estimation process, each test image has been blind compressed from $QF = 100$ to $QF = 50$ in decreasing steps of 5. For each JPEG compression, the quality factor estimation process was used to determine the QF. The mean estimated QFs for all 1338 test images and each correctly identified detection accuracy rate $P_{de}$ for each JPEG compression quality factor are shown in Table 4, based on equation 7.

$$P_{de} = \frac{\delta}{\beta} \times 100\%$$

where $\delta$ is the number of correctly detected QF and $\beta$ is the number of images tested. The mean estimated QF results indicate the QFs can be estimated with high accuracy. The only exceptions for lower correct detection rates, $P_{de}$, were obtained
for $QF = 50$, $QF = 60$, and $QF = 100$. In the case of $QF = 50$, $P_{de}$ was very low at approximately 18.2%, meaning that the process was probably detecting QFs close to $QF = 55$. For $QF = 60$, and $QF = 100$, the detection rates were slightly better at 38.6% and 65.7%, respectively. For comparison, both the mean estimated QF value and correct detection rate were used for each result to estimate the actual QF for the images. The QFs were then grouped into three different ranges: $QF \geq 90$, $90 < QF < 75$ and $QF < 75$. The grouping into three QF ranges did not have an overall effect on the authentication process. Results obtained for $P_{de2}$ also showed the correct detection accuracy rates in these QF ranges were on average at 99%.

Table 4. QF estimation for watermarked images (JPEG compression only)

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>$P_{de}$</th>
<th>$T$</th>
<th>$P_{de2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>98.16</td>
<td>65.7%</td>
<td>T</td>
<td>$P_{de2}$</td>
</tr>
<tr>
<td>95</td>
<td>94.87</td>
<td>97.3%</td>
<td>0.9</td>
<td>98.8%</td>
</tr>
<tr>
<td>90</td>
<td>90.06</td>
<td>98.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>84.20</td>
<td>91.4%</td>
<td></td>
<td>0.7 99.1%</td>
</tr>
<tr>
<td>80</td>
<td>79.77</td>
<td>97.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>75.33</td>
<td>97.0%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>69.77</td>
<td>98.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>64.42</td>
<td>93.7%</td>
<td></td>
<td>0.5 99.4%</td>
</tr>
<tr>
<td>60</td>
<td>62.42</td>
<td>38.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>55.15</td>
<td>94.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>54.25</td>
<td>18.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5 summaries the results obtained for test images that have been attacked via copy-paste and then JPEG compressed. Each watermarked image has been tampered randomly in different regions by applying a copy-paste attack to 5% of the watermarked image (9830 pixels in 384512 pixels image), and also compressed with different QF values. The results showed that the quality factor estimation process was highly accurate even under these attacks. From Table 5, the lowest correct detection
rates were obtained for $QF = 50$, $QF = 60$, and $QF = 100$. Two other experiments were performed with the test image subjected to only the copy and paste attack and with the test image without any modification. The detected QFs achieved for both experiments were approximately 99, and fit well in the upper range of $QF \geq 90$. Similarly, the results of $P_{det2}$ also showed the correct detection rates in the three ranges were highly accurate with an overall average of 99%. As such, the threshold can be adapted into the three QF ranges according to the estimated QF of each test image as described in Section 3.2.

Table 5. QF estimation for watermarked images (Copy and paste attack + JPEG compression)

<table>
<thead>
<tr>
<th>Actual QF</th>
<th>Estimated QF</th>
<th>$P_{det}$</th>
<th>$T$</th>
<th>$P_{det2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>98.60</td>
<td>72%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>95.00</td>
<td>100%</td>
<td>0.9</td>
<td>99.1%</td>
</tr>
<tr>
<td>90</td>
<td>90.14</td>
<td>98.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>84.83</td>
<td>97.9%</td>
<td>0.7</td>
<td>99.3%</td>
</tr>
<tr>
<td>80</td>
<td>79.95</td>
<td>99.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75</td>
<td>75.22</td>
<td>99.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>69.87</td>
<td>99.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>64.46</td>
<td>98.7%</td>
<td>0.5</td>
<td>99.2%</td>
</tr>
<tr>
<td>60</td>
<td>61.54</td>
<td>63.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>54.93</td>
<td>96.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>53.32</td>
<td>20.4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.4 Summary

In this section, we presented the relationship between QF and threshold, and proposed a framework incorporating the generalised Benford's Law as an image forensics technique to accurately detect unknown JPEG compression levels in semi-fragile watermarked images. We discussed the limitations of using predetermined thresholds in semi-fragile watermarking algorithm. In our improved semi-fragile watermarking method, the test image was first analysed to detect its previously unknown quality factor for JPEG compression by using generalised Benford’s Law model, before proceeding with the semi-fragile authentication process. The results showed that QFs can be accurately detected for most unknown JPEG compressions. In particular, the average QF detection rate was as high as 96% for watermarked images compressed with QFs between 95 — 65, and 99% when the image was subjected to tampering of 5% pixels of the image and compressed with QFs between 95 — 65. The threshold was adapted into three specific ranges according to the estimated QF of each test image.
4 Image Forensics

Recently, an interest has developed in identifying reliable techniques that are capable of accurately proving the authenticity of an image, without the requirement of actively inserting a digital watermark or signature into the data. Whilst the watermarking schemes discussed in section 2 have been shown to be useful for protecting the integrity of the image, there always exists the underlying risk that the watermark data might be forcibly or accidentally removed. When this happens, the image is effectively stripped of its identity, and its integrity is extremely difficult to prove. Forensic techniques aspire to achieve similar objectives but do not rely on the strength of embedded data. Instead, the ambition is to prove the authenticity of an image based solely on the data provided.

The two main areas of focus within the field of image forensics are camera identification and forgery detection. Camera identification is the task of successfully linking suspect images to the source camera that captured the image, in order to provide evidence that the origin of the image is as claimed. For example, a claim might be made by Person A that they captured an image of a compelling real-life event in order to gain acclamation. However, it is possible that Person B makes the same claim, and suggests that it was taken from their camera which happens to be a different make or model. A scrutinised forensic evaluation would attempt to review the properties of both cameras’ image acquisition process, and determine the correct source for the image. This exercise might be relatively trivial if both camera’s are vastly different, but what happens if Person A and Person B both own the same make and model camera? The forensic expert must then locate features in the image acquisition process of both cameras that differ. It should be possible to locate this feature within the data of the image in question, and therefore conclude which device captured the image. Forgery detection, on the other hand, is the practise of ensuring that the content of the image has not been manipulated. One of the most typical forms of content manipulation is splicing, which involves removing content from one image and overwriting it with something similar from another image to form a composite. This type of modification dates back over 150 years; a famous example of which is the Abraham Lincoln portrait [48]. In this example, a portrait of John Calhoun was manipulated such that it appeared as if the portrait was of Abraham Lincoln. In fact, Lincoln never posed for the portrait, and the image was actually constructed by flipping and resizing Lincoln’s head from a head-shot photograph taken by Mathew Brady such that it resembled the same proportions as the Calhoun portrait. Calhoun’s face was then replaced by Lincoln’s face to produce a composite image.

Part of the challenge for image forensics lies in the fact that it is rarely immediately obvious whether or not an image has been manipulated. If a good job has been made of doctoring the image, it will look completely legitimate in plain sight. Therefore a distinction must be made between clean images that have not been altered in any way, and dirty images that are no longer true to their original form. Clean images are typically those that have come directly from the source that created them, without having been subjected to any external post-processing. However, it is often extremely rare to locate an image as clean as this, as most photographers (even at an amateur
level) are likely to enhance their images through image editing software to provide better visual clarity, even though the content itself will remain true. To what extent such enhancements constitute a manipulation remains uncertain at this point. For this chapter we define clean images as those extracted straight from the camera that captured them, and dirty images as any image that has been manipulated in any way, including enhancements. By classifying the images according to these terms, we are effectively suggesting that a clean image accurately represents the exact scene from which the image was captured, and also inherits only the characteristic properties marked into the image data by camera processing.

Figure 15 presents a diagram of the two main areas of research in image forensics. The diagram shows that a given image can either be captured by a digital camera (in which case, the task is to identify anomalies in the camera processes that are also found in the image data), or the image will have been edited by software (in which case, anomalies are found in the image data that reflect manipulations). A suspect image is usually intercepted after either or both of these processes have been instantiated, and it is the job of the forensic specialist to establish the origin of the image.

In this section, we begin by explaining the most significant techniques that have been developed for the camera identification and forgery detection areas. In Section 4.2 we focus purely on camera identification, and present a novel approach to identifying anomalies within image data, before discussing the results of this work in Section 4.3. We then provide a concluding summary in Section 4.4.

4.1 Literature Survey

Camera Identification

One of the earliest reported approaches for digital camera identification characterised the imaging sensor from the device [49]. The imaging sensor is arguably the most important component of the image acquisition process, as it captures the light intensity
of the scene on a pixel-by-pixel basis, and converts it into an electrical signal. From here, the signal will pass through a Colour Filter Array (CFA), which interpolates the colours for each pixel and the image is effectively born. However, it is possible that the imaging sensor operates with an element of noise, caused by hot or dead pixels. Errors such as this can often be seen in the final image, even if the image has been lossy compressed. As the error is likely to be slightly different for several devices, the technique is useful for reliably linking images to the source sensor - and therefore the source camera - that captured the image. However, most modern cameras are able to detect deficiencies in the processing such as this, and often remove the hot or dead pixels altogether. As the scheme relies on the existence of such pixels, it can only be targeted towards cameras that do not correct errors such as these.

In 2006, research by K. S. Choi et al. led to the discovery that the camera lens produces aberrations in images, due to the design and manufacturing process [50]. Lens radial distortion was found to be quite a common property for inexpensive wide-angle lenses, and it causes straight lines to render as curved lines on the camera sensor. A camera lens has various focal lengths and magnifications in different areas, and when the transverse magnification $M_T$ increases with the off-axis image distance $r$, a barrel distortion presents itself, as shown in Fig 16.

By calculating the precise radial distortion for a given device, as well as the relative radial distortion witnessed from a suspect image, it is possible to infer whether or not the image originated from that device. The technique acts as an excellent feature for providing a successful classification, but is likely to be insufficient in isolation. Instead, this feature will need to be used in conjunction with several other similar techniques in order to make a more informed and justified classification.
Arguably the most prominent research in the camera identification area, is that proposed by J. Lukáš et al. [20, 51], and verified by N. Khanna et al. in 2009 [22]. The technique relies on pattern noise, which is a deterministic component that remains consistent for all images that the sensor captures. Pattern noise can be sub-divided into two categories: fixed pattern noise (FPN) and photo-response non-uniformity noise (PRNU). The FPN is an additive noise that is suppressed to varying standards by many camera manufacturers, and is relative to exposure and temperature [20]. For these reasons, it is not reliable for camera identification purposes as it is inconsistent. PRNU, on the other hand, is a multiplicative noise and contains a property referred to as pixel non-uniformity (PNU), which is defined as the sensitivity differences to light at each pixel. The PNU is a direct result of the manufacturing process and is therefore not influenced by exposure and light. Indeed, the PNU noise remains the same for each image that is taken, meaning this component is extremely useful for determining the source camera that captured an image. To complete the classification, a reference pattern for the camera must first be identified. This is achieved by using a denoising filter $F$ and averaging the noise residuals $n^{(k)}$ from multiple images $p^{(k)}$.

$$n^{(k)} = p^{(k)} - F(p^{(k)}).$$

(8)

Selected regions from image $p$ are then checked for the existence of the pattern noise from camera $C$ by calculating the correlation $P_C$ between the noise residual $n = p - F(p)$ with the camera reference pattern $P_C$, as shown in Equation (9).

$$P_C(p) = \text{corr}(n, P_C) = \frac{(n - \bar{n}) \cdot (P_C - \bar{P_C})}{\|n - \bar{n}\| \|P_C - \bar{P_C}\|}.$$  

(9)

where the bar above a symbol denotes the mean value [52].

The pattern noise obtained from a suspect image can now be compared with the pattern noise obtained from the device itself. If the correlation is identical, then there can be little doubt that the image originated from the device, as the chances of two camera’s producing the same pattern noise are extremely remote.
Forgery Detection

Significant progress has also been made in the forgery detection research area for authenticating image content, such as the splicing example discussed at the beginning of this section. In fact, the sensor pattern noise technique introduced by J. Lukás et al. can easily be adapted to authenticate images. As the complete pattern noise exists for every pixel in an image, a manipulated image can be derived when the pattern noise is not present at a particular region of interest. It is important to note, however, that the PRNU noise will not be present in highly saturated areas of clean images, and is also highly suppressed in dark areas, as the noise is multiplicative. Therefore, a region that does not contain the pattern noise should be checked to ensure that neither of these two properties hold true before classifying the image as tampered. Further details of how this can be achieved statistically are discussed in [52].

Whilst much research is concentrated on calculating anomalies in the image acquisition process of digital cameras, and then locating marks of those anomalies in the image data, some researchers have taken a different approach and are considering how "fingerprints" of software manipulation also exist in the image data. The most prolific work from this angle is lead by H. Farid’s research group at Dartmouth college. Specifically, they have reviewed how certain image manipulation operations such as resizing, alter the underlying pattern of pixels in a distinct way [53]. When creating a composite from two or more images, parts of an image are often enlarged (up-sampled), and when this happens, extra pixels are formed. Figure 17 shows what happens when a small 4x4 pixel patch is stretched to produce a 4x7 pixel patch. The numbers contained within the original 4x4 block shown in 17(a), correspond to the brightness at each location. The highlighted rows in 17(b) indicate added information, which is calculated by averaging the values of the immediate neighbours.

---

![Fig. 17. Enlarging a 4x4 pixel patch [53].](image)

(a) A 4x4 pixel patch. (b) Extra pixels added when enlarging.
When images are enlarged in this manner, there exists a perfect correlation between neighbouring pixels, which is a rare property to find in natural images. Therefore, whenever this property is detected by a forensics specialist, they can derive a probability that the image has been manipulated.

In other work, H. Farid's research group also found that composite images can also be identified by studying the light reflected into the subjects eyes. The positioning of white dots (caused by flash photography) indicate the direction of the light when the image was captured [53]. When images are spliced together, these issues are often overlooked. For a clean image, when several people all appear in the scene the correlation of the light direction will match almost exactly. However, when a person has been spliced into the image from another image, the direction of light on the subjects eyes will not match. By studying the light pattern, it is often a fairly trivial process to determine whether the image is genuine or not. Similarly, the author discusses how lighting observations can be applied more generally to images in [21].

In this work, the author explains how the light striking a surface is dependant on the position of the light source. An estimate of the direction of the light source can be derived from an image by reviewing a given object's 2-D surface contour, such as a human jawline and chin. The lighting of this object can ultimately be compared against that of other objects in the photo, and if there exists a mismatch in lighting direction, then the image is likely faked.

As described in this section, there have been significant advances made in the fields of camera identification and forgery detection in recent years. For the remainder of this chapter, we concentrate solely on the camera identification area, and present a novel technique for locating anomalies in image data.

### 4.2 Statistical Process Control

At present, much research for camera identification has been based around identifying anomalies in a camera's image acquisition process, and then hoping to find a "fingerprint" of these properties in the image data. Whilst this research has produced some promising results, it is never easy to generalise the image acquisition process for a wide range of digital cameras, as each process can be quite vastly different from manufacturer to manufacturer. Instead, it is desirable to create a model such that the anomalies for any type of digital camera can be quickly and easily identified. In this section, we discuss how Statistical Process Control (SPC) can be used for such a purpose, and how it fits into the camera identification model as shown in Figure 18.

**Introducing Statistical Process Control**

The theory of SPC was developed in the late 1920's by Dr. Walter Shewhart, a physicist and statistician at the AT&T Bell Laboratories, USA, and was designed in an effort to acknowledge quality control and improvement for the manufacture of goods [54]. Shewhart recognised that products built to a high standard with good quality components, often produced better results in the field. In 1931, Shewhart released a
study of his work that outlined a statistical approach for detecting the degree of control within processes over time [55]. The aim of Shewhart's work was to eliminate unexpected sources of variation that cause the process to operate with less accuracy. These variations were referred to as special-cause, and are caused by irregular events or circumstances that have an obvious impact on the process. Any variation that could be explained, was referred to as common-cause variation. In a perfect world, each measurement taken over time would produce the exact same result. However, in the real world, there are often external influences that affect the performance of processes.

SPC has been successfully applied to many areas of manufacture to maximise the efficiency of production processes to deliver high quality products. It was first applied to automobile manufacture by several Japanese manufacturers, and such was the success of its use on the end product, the Ford Motor Company soon followed [56]. It has since also been applied to industrial applications such as the pulp and paper industry [57–59], and has even been considered for improving healthcare processes [60].

The use of SPC can easily be adapted for use in image processing by substituting the measurements with image data taken from a digital camera. The quality of the complete image acquisition process for the camera can be inferred, and a study of any widely varying images can lead to the discovery of a unique feature of the device that can act as a "fingerprint" for camera identification.

Control Charts

A key tool of SPC for reviewing process variation, are control charts, which are used to graphically display the variation shifts from each measurement. Typically, two control charts are required to expose the data obtained from the process in its entirety: one to display the shifts in the process mean, and one to display the shifts or changes in the amount of process availability [55]. There are several types of control chart, each calculated in different ways, and chosen according to the best fit for the application. Our initial work is focused on individuals charts (commonly referred to as X charts) to display the process mean, and moving range charts (referred to as Rm charts) to display an estimate of the common-cause variability of the process. X and
$R_m$ charts are suited to instances where individual measurements are obtained, and will therefore be useful for representing data collected from multiple images taken by several digital cameras.

Both control charts are comprised of a centreline $CL$ (which is the mean value obtained from all measurements), an upper control limit $UCL$, and a lower control limit $LCL$, as well as the physical data obtained from each measurement $X$. The $UCL$ and $LCL$ are calculated at around $\pm 3$ standard deviations above and below $CL$, respectively, to obtain results with a false-positive margin of approximately $0.27\%$ [55]. If any measurement falls outside of these control limits, then the measurement is considered out-of-control.

Constructing the Control Charts

The construction of the control charts is based completely on the data measurements obtained for $X$. Traditionally, the $R_m$ chart is plotted first, as these charts provide information on the overall process variability. The first step is to calculate the differences between neighbouring values in $X$ to produce $R_m$. The $CL$ is simply the mean of all measurements of $R_m$, denoted as $\bar{R}_m$, and is therefore calculated according to Equation (10).

$$\bar{R}_m = \frac{\sum_{i=1}^{k} R_{mi}}{k}. \quad (10)$$

where $k$ refers to the total number of elements in $R_m$. A table of constants (Table 6) is then used to calculate the $UCL$ and $LCL$ control limits.

Table 6. Constants for Calculating Control Limits [54].

<table>
<thead>
<tr>
<th>Observations in Sample</th>
<th>$d_2$</th>
<th>$A_2$</th>
<th>$D_3$</th>
<th>$D_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.128</td>
<td>1.880</td>
<td>0</td>
<td>3.267</td>
</tr>
<tr>
<td>3</td>
<td>1.693</td>
<td>1.023</td>
<td>0</td>
<td>2.575</td>
</tr>
<tr>
<td>4</td>
<td>2.059</td>
<td>0.729</td>
<td>0</td>
<td>2.282</td>
</tr>
<tr>
<td>5</td>
<td>2.326</td>
<td>0.577</td>
<td>0</td>
<td>2.115</td>
</tr>
<tr>
<td>6</td>
<td>2.534</td>
<td>0.483</td>
<td>0</td>
<td>2.004</td>
</tr>
<tr>
<td>7</td>
<td>2.704</td>
<td>0.419</td>
<td>0.076</td>
<td>1.924</td>
</tr>
<tr>
<td>8</td>
<td>2.847</td>
<td>0.373</td>
<td>0.136</td>
<td>1.864</td>
</tr>
<tr>
<td>9</td>
<td>2.970</td>
<td>0.337</td>
<td>0.184</td>
<td>1.816</td>
</tr>
<tr>
<td>10</td>
<td>3.078</td>
<td>0.308</td>
<td>0.223</td>
<td>1.777</td>
</tr>
<tr>
<td>15</td>
<td>3.472</td>
<td>0.223</td>
<td>0.348</td>
<td>1.652</td>
</tr>
<tr>
<td>20</td>
<td>3.735</td>
<td>0.180</td>
<td>0.414</td>
<td>1.586</td>
</tr>
</tbody>
</table>

The $UCL$ and $LCL$ values are calculated by using Equation (11), where $D_3$ and $D_4$ are obtained when observations in sample $n = 2$. 
When constructing X charts, the CL refers to the mean value of all measurements in X, and is denoted as X. The UCL and LCL values are calculated by adding or subtracting 3 standard deviations from this value, where an estimate of the standard deviation is obtained from Equation (12).

\[
\sigma_X = \frac{R_m}{d_2}.
\]

where \(d_2\) is taken from the table of constants, again when observations in sample \(n = 2\). The UCL and LCL control limits can then be calculated from Equation (13).

\[
\begin{align*}
UCL_X &= \bar{X} + 3\sigma_X \\
LCL_X &= \bar{X} - 3\sigma_X.
\end{align*}
\]

Using Statistical Process Control for Image Forensics

SPC can be used to identify anomalies in the image acquisition process of a digital camera by collecting a series of identical images from the device, and using the mean pixel data (across each colour plane) as the measurements for X. As mentioned previously, the aspiration is that the anomaly investigation process will lead to the uncovering of unique "fingerprints" for camera identification. According to Flickr (a popular image and video hosting site), the current most commonly used cameraphone device on their website is the Apple iPhone.

For our experiments, we therefore use four Apple iPhone 3G devices as primary devices. As cameraphones by definition are not primarily engineered for photography, inexpensive components are typically used, meaning the expectancy of witnessing a poor statistical control is increased. The Apple iPhone devices contain a 2 MegaPixel CMOS sensor and do not process any user settings or a zoom of any kind, meaning the exposure settings, focus, ISO settings, aperture, etc. are all automatically defined, if indeed they exist at all. The results obtained from the Apple iPhone 3G devices are later compared with those obtained from similar devices such as a Sony Ericsson W810i, and two Nokia N97 devices.

When acquiring the image data, it is important to nullify any environmental issues that could affect the data negatively. If the external conditions remain uncontrolled, it is likely that each device produces quite contrasting results, not necessarily because their image acquisition processes are different, but because, for example, the temperature or lighting conditions suddenly change. In our initial work in [61], the
images are acquired from a room that is not subjected to outside lighting as this constantly changes. Instead, the test scene was lit via fluorescent lighting. In addition, the room was air-conditioned to a constant temperature so as to reduce the influence of temperature changes on the image acquisition process. It is worth noting, however, that the SPC model can be applied when these external influences do exist, so long as they are taken into consideration when reviewing the data. For instance, if the temperature increases as each image is taken, this will have an affect on the image acquisition process. As such, if the process becomes less and less controlled over time, then the temperature is a likely cause.

The scene itself comprises a white bowl of colourful confectionaries (Figure 19, where the bright colours maximise the load on the CFA. A location reference point is then set up to determine the position for each device, and a series of 10 images are collected one after the other for each device.

![Example image obtained from test scene.](image)

Fig. 19. Example image obtained from test scene.

### 4.3 Results & Evaluation

In this section, we present the results from our initial implementation of SPC for image forensics, as reported in [61]. We also evaluate the significance of these results, and consider a refined implementation of a similar model based on images captured from a controlled environment. Comparisons are then made between both models, and a critique of how SPC can aid camera identification is provided.

The mean pixel values obtained from 10 images for all four Apple iPhone 3G devices is shown below in Table 7. By taking the 10 values for each device as $X$, control charts can be plotted to display the degree of control about the process mean $\bar{X}$. First, the $R_m$ values are determined by calculating the difference between neighbouring values of $X$. Table 8 shows the $X$ and $R_m$ data for iPhone A.
Table 7. Mean pixel values obtained for all Apple iPhone 3G devices [61].

<table>
<thead>
<tr>
<th>Shot No.</th>
<th>iPhone A</th>
<th>iPhone B</th>
<th>iPhone C</th>
<th>iPhone D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>110.2097</td>
<td>105.4888</td>
<td>104.6182</td>
<td>104.4518</td>
</tr>
<tr>
<td>2</td>
<td>109.6045</td>
<td>106.5222</td>
<td>104.6503</td>
<td>97.4483</td>
</tr>
<tr>
<td>3</td>
<td>109.1334</td>
<td>106.7678</td>
<td>105.4275</td>
<td>97.1251</td>
</tr>
<tr>
<td>4</td>
<td>109.1161</td>
<td>98.0294</td>
<td>105.4614</td>
<td>97.0542</td>
</tr>
<tr>
<td>5</td>
<td>109.2108</td>
<td>98.064</td>
<td>105.676</td>
<td>97.0346</td>
</tr>
<tr>
<td>6</td>
<td>101.3616</td>
<td>97.7303</td>
<td>105.1278</td>
<td>97.0085</td>
</tr>
<tr>
<td>7</td>
<td>101.7246</td>
<td>98.9264</td>
<td>105.4571</td>
<td>97.4346</td>
</tr>
<tr>
<td>8</td>
<td>101.2875</td>
<td>96.9707</td>
<td>105.5588</td>
<td>97.4245</td>
</tr>
<tr>
<td>9</td>
<td>101.2724</td>
<td>98.4508</td>
<td>105.5459</td>
<td>97.0113</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8. $X$ and $R_m$ values obtained for iPhone A.

<table>
<thead>
<tr>
<th>Shot No.</th>
<th>$X$</th>
<th>$R_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>110.2097</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>109.6045</td>
<td>0.605</td>
</tr>
<tr>
<td>3</td>
<td>109.1334</td>
<td>0.471</td>
</tr>
<tr>
<td>4</td>
<td>109.1161</td>
<td>0.017</td>
</tr>
<tr>
<td>5</td>
<td>109.2108</td>
<td>0.095</td>
</tr>
<tr>
<td>6</td>
<td>101.3616</td>
<td>7.849</td>
</tr>
<tr>
<td>7</td>
<td>101.7246</td>
<td>0.363</td>
</tr>
<tr>
<td>8</td>
<td>101.2875</td>
<td>0.437</td>
</tr>
<tr>
<td>9</td>
<td>101.2724</td>
<td>0.015</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>0.420</td>
</tr>
</tbody>
</table>

Using this data, the CL for the $R_m$ control chart is calculated as 1.141. The table of constants from Table 6 is used, where $n = 2$ to obtain $D_3 = 0$ and $D_4 = 3.27$. Subsequently, the $UCL$ and $LCL$ are then calculated as follows:

\[
UCL_{R_m} = D_4 \bar{R}_m = (3.27)(1.141) = 3.729
\]

\[
LCL_{R_m} = D_3 \bar{R}_m = (0.0)(0.294) = 0.0
\]

(14)

Similarly, the CL for the $X$ chart, $\bar{X}$ is calculated as 105.461. The $UCL$ and $LCL$ control limits are calculated according to Equation (15).

\[
UCL_X = \bar{X} + 3\bar{\sigma}_X = 105.461 + (3)(1.011) = 108.497
\]

\[
LCL_X = \bar{X} - 3\bar{\sigma}_X = 105.461 - (3)(1.011) = 102.426.
\]

(15)

The final control charts for iPhone A are shown in Figure 20, where circled nodes indicate that the measurements are out-of-control as they fall outside the control limits.
The $X$ chart (top) shows that every measurement taken from iPhone A is out-of-control. This indicates that the complete image acquisition process is statistically unstable. The $R_m$ chart shows that sample 5 is out-of-control. When mapped to the $X$ chart, this corresponds to the 5th and 6th images. The values of these two images are quite vastly different. The value for image 5 is 109.2108 compared to image 6 which yields the value 101.3616. By reviewing these two images further, it is possible to see a significant change in brightness between these images.

It is now worth evaluating how the results for iPhone A compare with the data obtained from the other iPhone devices. The control charts for iPhone B are shown in Figure 21.
Fig. 21. $X$ and $R_m$ control charts for iPhone B.

Whilst the $X$ chart is undoubtedly far more controlled than the $X$ chart for iPhone A, the same significant drop in values can be seen, this time between image 3 and image 4. Again, when reviewing these two images, a large shift in brightness is observed.

Figure 22 illustrates the $X$ and $R_m$ control charts for iPhone C. For this device, there appears to be no significant shift in measurements, and in fact, the two out-of-control measurements are only marginally outside the control limits. This indicates that this device was far more controlled than the previous two, at least for the 10 observations reviewed.
To complete the exercise for the iPhone devices, the $X$ and $R_m$ control charts are plotted for iPhone D, as shown in Figure 23. The same property of significant shifts in measurement values that was observed for iPhone A and iPhone B, also exists for iPhone D between the first and second images. Again, when reviewing both these images, a change in brightness value can be observed.

For comparison purposes, the same experiment was performed with the Sony Ericsson W810i cameraphone and standalone Samsung NV3 camera. We would expect that the quality of the image acquisition process should be greatly improved for the Samsung NV3, as it is likely to use higher quality components, and more care is likely to have been taken to ensure errors are corrected in the pixel data. The Sony Ericsson W810i on the hand should be more comparable to the iPhone 3G devices,
and if it does not possess the same brightness problem, it may be more controlled - although not as controlled as the Samsung NV3 is expected to be. The control charts for the Sony Ericsson W810i cameraphone are shown in Figure 24.

This device contains no out-of-control measurements, meaning that the image acquisition process is far more controlled compared to that of the iPhone 3G devices. The Samsung NV3 device is even more controlled, with the difference between the highest value measurement (132.04) and the lowest value measurement (131.89) only 0.15.

As we only have access to one Sony Ericsson W810i and one Samsung NV3, we cannot collect enough data to make an informed review of any errors found within
the image data. However, it is clear that the SPC model is reacting to the quality of each device.

Based on these observations it is obvious that there is some aspect of the iPhone image acquisition process that is affecting the brightness. The measurements taken before and after the decrease in data values are actually quite consistent, but the change in brightness is so vast that it renders much of the process out-of-control. It also appears from the SPC experiment as though the change in brightness is time dependant. Whilst iPhone C did not display any signs of this characteristic, it might have showed itself if we took more than 10 measurements. The initial work has therefore highlighted a feature of the image acquisition process for iPhone 3G devices,
that could be analysed in more detail to potentially create a unique "fingerprint" for identifying images captured from these devices.

In our most recent work, we have studied the effect that the lighting conditions have on the image acquisition process. As fluorescent lighting is known to flicker, the overall intensity of light could vary from shot to shot depending from when the image was captured. The environment was therefore adapted such that the fluorescent lighting was replaced by a flicker-free task lamp. The bulb itself emitted a true representation of daylight based on the Spectral Power Distribution ( SPD). It is important to simulate daylight conditions as this ensures the digital camera's are still
processing the images based on real-world lighting. Some lamps will emit light that is faded or distorted, which would not provide useful results for our experiments.

The scene itself was also modified such that we use a light tent to ensure no external light sources filter onto the object. The bowl of confectionaries was also replaced with an X-Rite ColorChecker® chart, as this comprises 24 carefully selected colour squares, that each represent real-world colours (i.e. skin, sky, and landscape tones). The chart is specially designed such that each colour is reflected just as it is in the real-world. The colours within the chart are also defined in terms of their exact RGB reference values which means it will be possible to review exactly how each device is interpreting the colours if necessary. Figure 26 shows a sample image that was captured from this revised environment.

![X-Rite ColorChecker® chart](image)

**Fig. 26.** Sample image taken from the modified test scene.

Finally, the number of measurements taken from each device is increased from 10 to 30 to provide us with a more complete representation of the processing. The calculations involved for constructing the control charts, however, remain the same.

Figure 27 shows the $X$ control chart obtained from iPhone A. Again, the same variation shifts that appeared in the earlier experiment can be noted. At each of these points, the brightness of the images shifted quite significantly. It can also be observed that the fluorescent lighting conditions from the first experiment was not the cause for the error scene in the camera processing, as the difference between the highest and lowest measurements for both experiments is approximately equal with that of this controlled experiment.

Whilst carrying out the experiment, an updated iPhone 3G model was released by Apple called the iPhone 3GS. The updated model carries a 3.2 Megapixel camera, and allows the user to define the focal point of the image. To identify the significance of these improvements, we ran the SPC experiment on the new model. The results of which are expressed as an $X$ chart in Figure 28.
This chart shows that the image acquisition process is far more controlled than that of its predecessor. Each of the 30 measurements taken are under perfect control, and there are no significant shifts in variation as seen for the iPhone 3G devices. A more diligent review of the new 'focus' setting on the iPhone 3GS shows that the exposure of the image is also defined when the focus is set. The exposure then remains the same until the camera is moved, or the conditions of the environment change drastically. This backs up our assumption that the brightness issues witnessed
for the iPhone 3G devices are due to an exposure calculation error, which is why the brightness flicks between dark and bright over time.

The experiment was also performed against two Nokia N97 devices. The camera on the Nokia N97 devices contains a 5 MegaPixel resolution, and allows the user to define white balance settings, ISO settings, and also zoom. To form the most suitable comparison with the results obtained from the iPhone devices, the resolution was set to 2 MegaPixels, and all other settings were disabled where possible, or otherwise set to "Automatic". The $X$ control charts for the two Nokia N97 devices are shown in Figures 29 and 30.

The control charts for the N97 devices show that each measurement - whilst more closely centred around $\bar{X}$ - is again not under complete statistical control. The second N97 device shows even less control than the first. By analysing the out-of-control measurements in greater detail, (or indeed any contrasting measurements) and comparing them with the more controlled images, it is likely that the variation can be explained and a unique "fingerprint" uncovered.

4.4 Summary

In this section, we have introduced image forensics, and outlined the most prolific research in the field - specifically, the latest research for camera identification and forgery detection have been introduced. In addition, we have demonstrated the benefits of using Statistical Process Control for analysing image data on a range of different digital cameras. Based on our initial research in [61], an anomaly in the camera processing elements was identified for iPhone 3G devices, whereby the brightness of the images was fluctuating. By analysing the latest iPhone 3GS model under the
same conditions, we have been able to prove that the newer model does not contain this property, meaning there is promise for the reliable detection of images obtained from iPhone 3G devices, and images captured from the iPhone 3GS.

We have also proved through both experiments, that SPC is useful for modelling the overall control of the image acquisition process for a particular camera. Figure 31 illustrates the degree of variability obtained from the initial experiment in [61] for 6 devices. It is clear from this illustration that the iPhone 3G devices all operate with a similar degree of variation (approximately 21%). The Sony Ericsson W810i is far more controlled, and outputs a degree of variance of approximately 1%. The Samsung NV3 standalone digital camera is even further controlled, and offers a variation of only 0.5%.

5 Conclusion and Future Work

Our future work in the image forensics domain will be concentrated on identifying more benefits of the SPC framework for camera identification. Further control charts can be examined, such as the Exponentially Weighted Moving Average (EWMA) chart, to determine whether there is an even more descriptive tool that can replace the $X$ and $R_m$ control charts. In SPC control charts, the formation of the measurements along CL can be used to derive common-cause and special-cause variation. Therefore, a scrutinised analysis of this content might be useful for isolating unique "fingerprints" for digital cameras. Similarly, Pareto charts and Cause & Effect diagrams have also been proved to be useful for identifying the cause of variation for a range of processes. These techniques could be adapted for use in the image forensics domain for identifying anomalies in the image data.
Fig. 31. Depth of variation for all devices.

References


