Analysis of Attacks on Common Watermarking Techniques

Johnson C. Lee, Student Member, IEEE
Electrical and Computer Engineering Department
University of British Columbia
2356 Main Mall, Vancouver, BC Canada V6T 1Z4

Abstract—Watermarking is a technique that is used in copyright-protection for valuable medias. Watermarking is measured by four quantities: capacity, security, robustness, and imperceptibility. In this study, robustness of five watermarking algorithms will be assessed. The study will also lightly touch on imperceptibility characteristics of the algorithms. The five algorithms are CDMA Spatial Domain Spread Spectrum, Comparison-Based Correlation in DCT Midband, CDMA Wavelet Domain Spread Spectrum, Bruyndonckx, and Cox. The latter two algorithms are fingerprinting algorithms, which is a subset of watermarking algorithms with slightly different design goal. Our attacks, based on easy, unintentional image processing revealed that DCT domain algorithms – Comparison-Based Correlation in DCT Midband and Cox, have the best robustness property to JPEG compression, downscaling, median, and mean filters. However, spatial domain techniques such as Bruyndonckx can achieve better PSNR with less visual impact on the image.

Index Terms— Image processing; Image watermarking; Watermarking algorithms; Watermark robustness

I. INTRODUCTION

Digital watermarking techniques have been developed to protect copyright media properties. Many different watermarking techniques have been presented for different types of multimedia content. Video, for example, is a collection of images and one can take advantage of the interpolation for watermarking between frames. For images, however, there is less capacity to hide the information so the techniques are based on spatial, frequency (DCT), and wavelet domains. In this study, we will examine the affine and JPEG compression resistance of three basic and two advanced watermarking techniques.

II. BACKGROUND

Watermarking is defined as the process of embedding information into object/signal [1]. The information itself, i.e. the watermark, has to be imperceptible. In recent years, watermarking is mainly used for copy-protection or copyright-protection. However, historically watermarking is also used to send sensitive information by hiding it in another object. Such usage of watermarking is also known as steganography. In this paper, we will focus on watermarking for the purpose of copyright-protection.

Copyright-protection is different copy-protection in that copy-protection finds ways of inhibiting the copy process itself whenever the copyright of a media is in question. An example of copy-protection is the Content-Scrambling System (CSS) that is used in DVD copy-protection [2]. It is difficult to achieve copy-protection in open systems as shown by the infamous incident where a Norwegian teenager hacked the CSS system and released a DeCSS program on the Internet. Copyright-protection, on the other hand, is easier to implement and is used when ownership of copyright is in question.

III. PROBLEM DESCRIPTION

The specific application of watermarking for the purpose of copyright-protection is also known as fingerprinting or labelling. This usage of watermarking results in different requirements compared to regular watermarking. In normal watermarking, the quality of an algorithm is measured by four criteria: capacity, security, robustness, and imperceptibility [3]. Capacity measures the amount of data a watermarking algorithm can embed before causing severe visual degradation. Ideally, one would like a high capacity algorithm to embed more information. For fingerprinting application, high capacity algorithms mean more unique copies and can be shown to be more collusion resistant by applying coding theory [4]. Security, on the other hand, deals with who is capable of extracting the watermark from a marked media. Some algorithms allow anyone to extract the embedded information with the knowledge of the watermarking algorithm while others require a key to access the embedded information. Robustness deals with how well the algorithm is able to withstand attacks, both intentional and unintentional. Intentional attacks are usually geometric/affine transformations or addition of noise. Depending on the watermarking algorithm, some geometric transformations or noise can prevent the detection of a watermark; thereby making the media in question unmarked. On the other hand, compression algorithms often unintentionally attack a watermarked media. Compression algorithms’ ultimate goal is to reduce the size of a media without losing too much visual information. However, watermarking algorithms find unnoticeable areas to embed information that compression algorithm deem unimportant to retain during the compression process. The opposing goals of the two common media processing applications results in unintentional attacks on watermarking algorithms. Lastly, a watermarking algorithm’s quality is also measured by its imperceptibility. Most applications do not want the end customer to notice any visual...
differences because of quality standards. For fingerprinting application, imperceptibility is a requirement so the end user does not know about the existence of the watermark.

This study will focus on the robustness of five watermarking algorithms and lightly present the perceptibility issue for each. The robustness tests have been designed mostly for unintentional attacks from regular image processing.

IV. MATERIALS

This study uses existing implementation of the five algorithms found online. The CDMA Spatial Domain Spread, CDMA Wavelet Domain Spread, and Comparison-based Correlation in DCT Midband implementations are in MATLAB and can be found on [5]. The other two algorithms, Bruyndoncks [6] and Cox [7], are IEEE published algorithms with implementations found on [8].

V. METHODS

The attack methodology on the watermarked image is based on the idea that an attacker does not have any access to the original image or the watermark image/signature. The attacks are, therefore, done on the watermarked image using only the watermarked image as input. The individual, i.e. attacker, likely has no idea if the attack worked or not so the results are not known to the attacker.

The first attack is the JPEG compression attack. If the watermarked image is not already in JPEG format, the attacker can simply convert the watermarked image into a JPEG, varying the “quality factor” of JPEG compression to as low as he can before the features he needs on the image deteriorates. Even if the watermarked image is already a JPEG, the attacker can resave as a JPEG using a lower quality factor. This attack is a simple one without the need for complicated image processing software and many image viewers, available online, is able to save JPEG files using different quality factors. Because of how common and easy the JPEG attack is, resistance to JPEG compression is treated as the most important criteria in this assessment of robustness. The goal for each algorithm, therefore, is to find the quality factor point where the watermark can still be extracted.

The next attack is a scaling attacking. Pictures are often taken at a higher resolution and rescale down to fit the user’s needs. Scaling is a lossy form of image processing that is also commonly available. For example, Microsoft Windows offers a paint program that is capable of scaling images. In this attack, the watermarked image will be downscaled and the watermark will be extracted. If the watermarking algorithm requires same size as the original, a resizing process will be done. The percentage size at which a watermarking algorithm can tolerate will be the grading point for that algorithm.

Rotation was originally a factor that was deemed important to measure. However, every algorithm in this experiment failed to tolerate even a single degree of rotation. Therefore, the assessment will ignore the rotation factor. Although it may be possible for one to undo the rotation and attempt to extract the watermark, minor inaccuracies in the undo process will have the same devastating effect on all algorithms. For the reasons shown above, rotation attack will not be discussed in the results section.

Mean and median filters are simple burring functions of image processing software. The resistance of a watermarked algorithm against mean and median filter depends largely on where the watermark information is embedded. High frequency edge embedding will likely suffer from mean and median filters while low frequency intensity embedding will remain relatively resistant to such filter attacks. This study will verify this hypothesis by looking at algorithms that embed in the high frequency versus lower frequency.

In our study, we will use the Lena image, a 512x512 grayscale version, as the embedding media. The Lena image contains a nice mixture of details, flat regions, shadings, and textures. Some algorithms embed information in detail regions while others in flat regions. For these reasons, and that Lena is a commonly used test image, the Lena picture was chosen as our embedding media.

For resizing, also known as downscaling, requires more preprocessing. When an attacker downscales an image to his desired size, there is a loss of information. However, we noticed that most algorithms cannot take in a different sized input image to extract the watermark. Even if the implementation allows a different sized input image, the resulting extraction is always pure noise. Therefore, using MATLAB’s ‘imresize’ function, we can restore the original size of the image as shown in equation (1). All the downscaling results use this resizing to restore the image to the original size.

\[
\text{imresize(Watermarked\_Image, \ Original\_Size\_Watermarked\_Size)}
\]  

For median and mean attacks, no preprocessing is necessary before watermark extraction. MATLAB’s built-in ‘medfilt2(image, [m n]),’ where m and n are the dimensions of the filter, can median filter an image while ‘imfilter(image, fspecial(‘average’, n))’ can mean filter an image using n size filter.
VI. Results

Our study tested five algorithms for their robustness and visual imperceptibility. The five algorithms chosen are CDMA Spatial Domain Spread Spectrum, Bruyndonckx algorithm, Comparison-based Correlation in DCT midband, Cox algorithm, and CDMA Wavelet Domain Spread Spectrum. The CDMA Spatial Domain Spread Spectrum and Bruyndonckx are both spatial domain watermarking techniques. Bruyndonckx is an advanced form of spatial domain watermarking technique with less visual degradation specifically designed for fingerprinting purposes. However, as our study will show, both techniques suffer from the same problems. Similarly, Comparison-based Correlation in DCT midband and Cox algorithm are both watermarking algorithms in the DCT domain. They too suffer from similar robustness problems. Cox algorithm is also a fingerprinting algorithm and is actually a low frequency based algorithm. The mean and median filter issue will be most interesting comparing Cox and spatial domain techniques.

A. Spatial – CDMA Spatial Domain Spread Spectrum

Fig. 1 and Fig. 2 show the before and after images of a CDMA Spatial Domain Spread Spectrum watermarking. The embedded information, an image for this algorithm, is shown in Fig. 3 and the recovered image is shown in Fig. 4.

![Fig. 1. Original Lena](image1)

![Fig. 2. CDMA Spatial Domain Spread Spectrum Watermarked Lena](image2)

![Fig. 3. Original Watermark](image3)

![Fig. 4. Recovered Watermark](image4)

In terms of visual imperceptibility, the CDMA Spatial Domain Spread is a basic algorithm that spreads the information throughout the entire image without determining any characteristics of base image. Therefore, as Figure 1b shows, the watermark is distributed throughout the image in a noise-like manner. Its imperceptibility goal is not well meet but the recovered watermark is clean. The qualitative measurement for imperceptibility is the Peak-Signal-to-Noise-Ratio (PSNR) and for this algorithm, using Lena image, the PSNR is 335.4.

Subjecting the watermarked image to the JPEG attack described in the methods section, the CDMA Spatial Domain technique has a 100% recovery at a JPEG quality factor of 75. The deterioration point is approximately at quality factor of 50 as shown below in Fig. 5. The comparison between techniques will be done at a latter section.

![Fig. 5. CDMA Spatial Watermark, Recovered at Quality Factor 50](image5)

The scaling attack on CDMA Spatial Domain algorithm shows mild tolerance to downscaling. Initial testing scaled the watermarked Lena down to 80% of the original and attempt a recovery. The result is a near 100% recovery of the embedded watermark (shown in Fig. 6). Continuous testing shows that the tolerance is at approximately 40% to 50%. The 50% recovered embedded watermark is shown in Fig. 7.

![Fig. 6. CDMA Spatial, Recovered Watermark from 80% Scaling Attack](image6)

![Fig. 7. CDMA Spatial, Recovered Watermark from 50% Scaling Attack](image7)

Median and mean filter attacks for CDMA Spatial Domain have similar results, which is tolerable to a 3x3 averaging (i.e. mean) or median filter. At 3x3, the UBC watermark recovered is still visually distinguishable from the noise as shown in Fig. 8 and Fig. 9.

![Fig. 8. CDMA Spatial, Recovered Watermark from 3x3 Mean Attack](image8)

![Fig. 9. CDMA Spatial, Recovered Watermark from 3x3 Median Attack](image9)

B. Spatial – Bruyndonckx Algorithm

Bruyndonckx algorithm is a fingerprinting algorithm in the spatial domain. In terms of imperceptibility, Bruyndonckx algorithm is far superior as shown by the comparison between Fig. 10 and Fig. 11. The Bruyndonckx watermarked Lena image has only small visual differences compared to the original and the differences are hard to pick up. As shown by the two circles, the most obvious differences are the two dots near the Lena’s chin. Generally, fingerprinting algorithms have the property of making less visually perceptible watermarks. For this particular signature using Bruyndonckx algorithm, the PSNR ratio is 26407, an extremely high value.

![Fig. 10. Original Lena](image10)

![Fig. 11. Bruyndonckx Watermarked Lena](image11)
Another fundamental difference between fingerprinting and watermarking algorithm is the embedded information type. Previously with CDMA, an image was embedded. Fingerprinting algorithms uses a text signature generated according the algorithm’s specification. The signature used for this study is shown in Fig. 12.

To understand the results, one has to understand the signature comparison process. Besides from extracting the signature from the image, a correlation algorithm is used to determine if the extracted file is actually the signature. This correlation factor for the watermarked Lena image without any attack processing is 0.969178.

Subjecting the Bruyndonckx algorithm to the JPEG attack, the quality factor drop point is roughly 75, with a correlation factor of 0.904110. When the quality factor drops to 50, the correlation is 0.349315. Therefore, even through both Bruyndonckx and CDMA Spatial Domain Spread are both spatial domain techniques, and that Bruyndonckx is a fingerprinting technique, Bruyndonckx has lower resistance to JPEG attack than CDMA.

For geometric/affine transformation attacks, Bruyndonckx still suffers from lower resistance. As mentioned before, no algorithms were able to tolerate rotation without first perfectly negating the rotation. For downsampling, Bruyndonckx can tolerate only 60% downscaling, with a correlation factor of 0.791096. A 75% threshold was chosen as a reasonable cut-off point because deciphering correlation factor is a subjective process.

Median and mean filter attack resistance suffers from similar results. Median filter resistance, using the threshold mentioned above, has a tolerable point of 3x3 filter size, correlation factor 0.763699. Mean filter though, has a tolerable point of 2x2, correlation factor 0.818493. At 3x3 mean filter, the correlation factor drops to 0.681507, below the set threshold allowance.

C. DCT – Comparison-based Correlation in DCT Midband

The first Discrete Cosine Transform domain watermarking algorithm that will be examined is the comparison-based correlation in DCT midband. This algorithm is a watermarking but not fingerprinting algorithm so it is similar to the CDMA Spatial Domain Spread.

As shown in Fig. 13, the visual perception degradation is in the form of a patterned mask over the image. If viewed as a small picture, the pattern is not very noticeable. From a subjective visual comparison, this algorithm appears to have better imperceptibility performance compared to CDMA Spatial Domain Spread and it is confirmed by the PSNR ratio of 1933 versus the 335.4 of CDMA Spatial.

For JPEG resistance, the DCT domain algorithm performed extremely well. At quality factor of 28, most of the watermark, Fig. 15, is recovered with minor noise on the right side. By pushing the quality factor lower, the deterioration point is found to be about 23 with an approximate 40% noise coverage. The watermark recovery results from JPEG quality 23 is shown in Fig. 16.

Downscaling attack results has also shown that comparison-based correlation in DCT midband has better resistance than spatial domain algorithms. The recovered watermark from a 40% downscaling attack is shown in Fig. 17, which clearly shows the “UBC” silhouette with some noise. Recalling that the best spatial domain technique can tolerate only down to 50% downscaling, the 40% tolerance of this DCT algorithm is
clearly superior. We attempted to further push the downscaling attack to 30% but the results, Fig. 18, is too blurry to count as a successful watermark extraction.

The results from median and mean attacks exceeded expectations with the recovered watermark still clearly distinguishable after a 4x4 filter attack. Fig. 19 shows the recovered watermark from a 4x4 median attack. The “UBC” outline is still quite visible. Mean attack has similar results shown in Fig. 20. Unfortunately, 5x5 attack completely blurred the recovered watermark so the tolerance ends at 4x4.

DCT – Cox Algorithm

Similar to the comparison-based correlation in DCT midband, the Cox algorithm is a DCT domain watermarking algorithm as well. However, like Bruyndonckx algorithm, the Cox algorithm is a fingerprinting algorithm that uses a signature file instead of an image as the embedding file. The results from Cox algorithm can verify if DCT domain has superior resistance as indicated by the comparison-based correlation in DCT midband.

The first quality to look at, as before, is the perceptual degradation caused by watermarking. Fig. 21 shows the watermarked Lena image using Cox algorithm and Fig. 22 is the original Lena for comparison. Unlike previous algorithms, Cox algorithm embeds in the lower frequency zone so it is more difficult to spot the difference. The difference image between Fig. 21 and Fig. 22 is shown in Fig. 23. By looking at the difference image, one can tell that Cox algorithm is a low frequency, illumination embedding algorithm. Notice the difference in lightness near Lena’s nose and right eye (left from our view). Subjectively, Cox algorithm performs the best in the area of visual imperceptiveness. However, PSNR ratio indicates only 665.96 signal ratio. Therefore, the imperceptiveness criterion for Cox algorithm is debatable.

JPEG quality factor attack results are as impressive as the other DCT domain algorithm. At a quality factor of 20, the correlation factor is still 0.999073. Since saving in JPEG below 20 degrades the image to a non-commercially valuable form, it is sufficient to state that Cox algorithm is completely JPEG compression resistant.

Downscaling attack also exceeded expectation with correlation factor remaining 0.990587 at 20% of the original size. It is unusual though, that the correlation at 50% downsizing is 0.987915, meaning that at size 20% the correlation confidence level is actually higher than 50%. Compared with the results indicated in [8] on Cox algorithm, the results does not seem to agree. An examination of [8]’s analysis should be done to account for the inconsistency.

For median and mean filter attacks, we expected high resistance for Cox algorithm because median and mean filters would cause little change to the lower frequency zone which Cox algorithm embeds. As expected, for both median and mean Cox algorithm was able to tolerate a 10x10 size spatial filter. The correlation factor for median and mean are 0.949934 and 0.867709, respectively.
E. Wavelet – CDMA Wavelet Domain Spread Spectrum

The only wavelet algorithm that was examined is the CDMA Wavelet Domain Spread Spectrum. Unfortunately the wavelet fingerprinting algorithms’ implementations are unavailable on [], so there will be no comparison between watermarking and fingerprinting algorithms for wavelet domain.

The visual difference between a watermarked and non-watermarked image is shown in Fig. 24 and Fig. 25. One can notice a smooth noise over the entire image, almost as if the image was printed using a low dpi printer. Subjectively, one does not notice the difference unless compared with the original. The PSNR value for this algorithm using Lena image and UBC watermark logo is 665.8780, which is similar to the Cox algorithm’s PSNR value.

The JPEG resistance of the CDMA Wavelet Domain Spread Spectrum is at a quality factor of 20, similar to Cox algorithm. At quality factor of 30, the watermark logo can be recovered 100%. Fig. 28 shows the recovered watermark image after a quality factor 20 JPEG attack. There are evident noises throughout the entire image. Fig. 29 shows the recovered watermark image from a quality factor 30 attack and it has no apparent noise.

Downscaling attack on CDMA Wavelet Domain Spread Spectrum indicates a tolerance level of 50%. The resulting recovered watermark for 50% and 40% is shown in Fig. 30 and Fig. 31, respectively. Although one can argue that 40% downscaling attack recovery still contains a visible outline but for consistency with the downscaling results from comparison-based correlation in DCT midband, a 50% downscaling tolerance point was chosen instead.

F. Comparisons

As shown in Table 1, Bruyndonckx algorithm has the best PSNR ratio while Cox dominates the attack resistance values. Even looking at the Comparison-based Correlation in DCT Midband compared to other watermarking algorithms, CDMA Spatial and Wavelet Domain Spread Spectrum, the DCT domain algorithm is superior in downscaling, median, and mean resistance. Although JPEG compression value for Comparison-based Correlation in DCT Midband is 23 while CDMA Wavelet is 20, the JPEG Compression resistance value is subjective since acceptable values are gauged visually.
TABLE I
RESISTANCE COMPARISON OF FIVE WATERMARKING/FINGERPRINTING ALGORITHMS

<table>
<thead>
<tr>
<th></th>
<th>Spatial Domain</th>
<th>DCT Domain</th>
<th>Wavelet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CDMA Spatial Domain Spread Spectrum</td>
<td>Bruyndoncx Comparison-based Correlation in DCT Midband</td>
<td>Cox CDMA Wavelet Domain Spread Spectrum</td>
</tr>
<tr>
<td>PSNR</td>
<td>335.4</td>
<td>26407</td>
<td>665.96</td>
</tr>
<tr>
<td>JPEG Compression</td>
<td>50</td>
<td>75</td>
<td>23</td>
</tr>
<tr>
<td>Downscaling</td>
<td>50%</td>
<td>60%</td>
<td>40%</td>
</tr>
<tr>
<td>Median</td>
<td>3x3</td>
<td>3x3</td>
<td>4x4</td>
</tr>
<tr>
<td>Mean</td>
<td>3x3</td>
<td>2x2</td>
<td>4x4</td>
</tr>
</tbody>
</table>

VII. CONCLUSION

The algorithm with the best resistance to the attacks in this study is the Cox algorithm. The two DCT domain algorithms both showed superior resistance to JPEG compression, downscaling, median, and mean attacks. However, spatial domain algorithm has the best visual performance with the highest PSNR. A limitation of this study is that combined attacks are not analyzed. JPEG compression attack in conjunction with downscaling is not atypical of typical web image processing. Such conjunction attack may prove to be more successful at removing watermarks without deteriorating image quality.

One problem that was encountered is the rotational resistance. All the algorithms cannot tolerate even a one degree rotation. Although rotation is in fact a no loss process, an auto rotational reversal program must be written first in order to test rotational resistance. If the program is able to completely undo the rotation, then rotational resistance is not even an issue to consider. However, it is more likely that some error is introduced when undoing the rotation and resistance can then be analyzed based on the error.

Cropping resistance was not analyzed either because it would depend on preprocessing. By using a search algorithm to in the correct part of the original image which the resulting cropped watermark came from, one can replace that section and run the extraction algorithm to test. Pure no preprocessing cropping resistance is meaningless when other algorithms is able to provide better confidence level after preprocessing but is unable to handle pure cropping attacks.

ACKNOWLEDGMENT

J. C. Lee thanks Deepa Kundur and William Luh from Texas A&M University for helping explain some subtle watermarking and fingerprinting issues.

REFERENCES