Watermarking Technique for self-correcting image

Term project for Information Theory – Liu, Tse-Wei D10922035

December 12, 2022

Abstract

Digital watermarking technology enables image authors to protect copyright and ensure the correctness of images, however, there is also a growing threat of new tampering methods such as deepfake. This study reviews the existing methods in terms of their reversibility, non-removability and robustness, in hope of providing a potential framework that could help to prevent deepfake modification. Finally, we proposed the hybrid methods which combines the selection on region of interest (ROI), discrete wavelet transform (DWT), discrete cosine transform (DCT) with singular value decomposition (SVD) or discrete Gould transform (DGT), and further point out the challenges in each step.

1 Introduction

In the past years, watermarking techniques have been implemented into several different applications. For example, they could be used to ensure the correctness of pictures and prevent human malicious tampering [Kumar and Singh, 2021, Gupta and Rama Kishore, 2022, Sun et al., 2022], to secure the copyright of the pictures [Kumar and Singh, 2021, Gupta and Rama Kishore, 2022, Sun et al., 2022, Fei et al., 2022], or as a steganography method to deliver sensitive messages through a common image which could be seen everywhere in the daily life [Singh and Gehlot, 2022]. They are also being used in different scenarios, such as the medical field [Sun et al., 2022] or further integrate with other techniques such as deep learning [Fei et al., 2022].

However, with the growing demand for image transmission and the increasing threat of new tampering technologies such as deepfake, it is critical to review existing methods and see if there is a way to better ensure the correctness, prevent malicious tampering, and go a step further - restore the tampered image and bring it back to the original image.

The goal of this study is to provide a framework for self-correcting watermarked images by utilizing the two characteristics of the watermarking techniques, that is, the ability to transmit hidden information and hard to be removed without damaging the original image. In the framework, the reversible features of the input image will first be extracted, the features will then be used as the watermark and embedded into the original image. When the image is sent, the image features in the watermark can be extracted to authenticate whether the delivered image has been tampered with. Once the image is regarded as tampered image, the reversible message of the extracted watermark could be used to rebuild the original image. The feasibility and effectiveness of the proposed framework is determined through the following points, that is:

- Robustness of the watermark: the watermark has to perform a certain level of robustness, the higher robustness indicates better resistance under different attacks and maintains good integrity. This is crucial in the self-correcting step because if the reversible features in the watermark are easily lost through the attack, we cannot expect the restoration to be applicable.

- Reversibility of the feature: the reversibility of the extracted feature is one of the most crucial steps in the proposed framework, the framework needs to ensure that the feature contains enough information for restoration, while the watermark needs to maintain invisible to human eyes.

- Non-removability of the watermark: the watermark has to be non-removable, which means it couldn't be removed without damaging the original image. This could help in both the image authentication step and the self-correcting step, as if the watermark could be removed easily, the validity of the authentication is going to be questioned.

2 Method

In this section, I investigate the possibilities of the proposed framework through a literature review of the recently proposed techniques, and to see if there is an approach that can correspond to our use case. I will focus on the three points stated above, which are reversibility, non-removability, and robustness. Hope we could find some potential methodologies that match our expectations.

2.1 Survey on the robustness and imperceptibility

For watermark embedding methods, when the capacity, i.e., the number of watermarks that can be embedded into an image is fixed, the imperceptibility and robustness of watermarks are usually negatively correlated [Wan et al., 2022]. Therefore, our goal in this section is to find a single or hybrid embedding procedure that maximizes the imperceptibility and robustness.

The existing watermark techniques could be divided into two main groups, which are the spatial domain and the transform domain. The spatial domain embeds the watermark directly into pixels, which gives it the property of time-saving, easy embedding and extracting. However, because it directly works on pixels, the watermark could not be robust against attacks and noises. The transform domain, on the contrary, focuses on adding the watermark information into the frequency domain of the image, which gives these methods good imperceptibility as the watermark could be spread out through the entire image. Although the embedding process is more complex in the transform domain, more and more research and methods are focusing on this domain and reaching a higher popularity compared to the spatial domain.

In the following section, I will introduce several methods from both the spatial domain and the transform domain, and emphasize the possibility that the methods adapt to our framework.

- Spatial domain - Least Significant Bit (LSB)

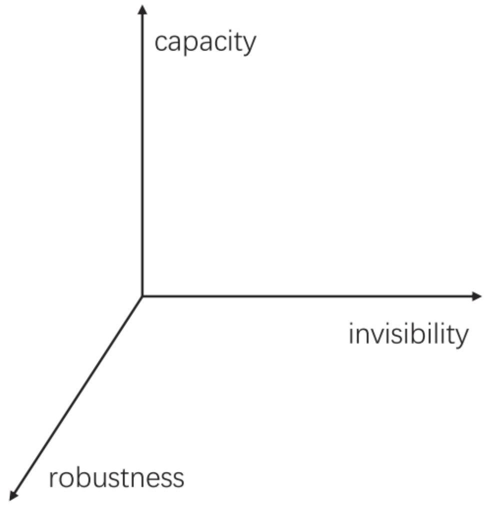


Figure 1: The trade-offs among imperceptibility, robustness, and capacity. (Picture retrieved from: [Wan et al., 2022])

LSB is one of the earliest proposed methods to implement the invisible watermark, it utilizes the different weighting inside the bits of an image, and embeds the watermark into those bits with less weighting so that the difference is hard to be observed by humans.

There are several advantages when it comes to embedding the watermark with LSB, the two major advantages are the simplicity of the watermark extraction and the imperceptibility of the watermark. However, since the watermark is directly embedded into the image pixels, any of the attacks and noises could be harmful to the watermark. Despite the efforts of combining LSB with different methods to try to reach the balance between imperceptibility and robustness, the robustness of the LSB watermark is still unsatisfactory and could only withstand with simple attacks such as cropping and noises, and LSB have been gradually ignored in the field of digital watermarking.

In our framework, as stated before, the robustness of the implemented watermark is one of the key requirements in our framework, so the LSB might not be an appropriate option to embed our watermark.

- Transform domain - Overview

  In recent years, methods in the transform domain have received more attention than those in the spatial domain, and the main difference between the two lies in the process of embedding the watermark. Unlike the methods in the spatial domain, which directly embed the watermark into the bits, methods in the transform domain must first convert both the host image and the watermark into frequencies, and then embed the watermark into the host image by choosing a certain frequency pattern. Some methods are often mentioned in the recent study, such as discrete cosine transform (DCT) and discrete wavelet transform (DWT).

- Transform domain - Discrete Wavelet Transform (DWT)

  Discrete Wavelet Transform (DWT), similar to other wavelet transform members, decomposes an image into a set of band-limited components that can be reassembled to reconstruct the original image without error. DWT is often used in the watermark embedding step to filter out certain frequency domains which have higher invisibility or robustness.

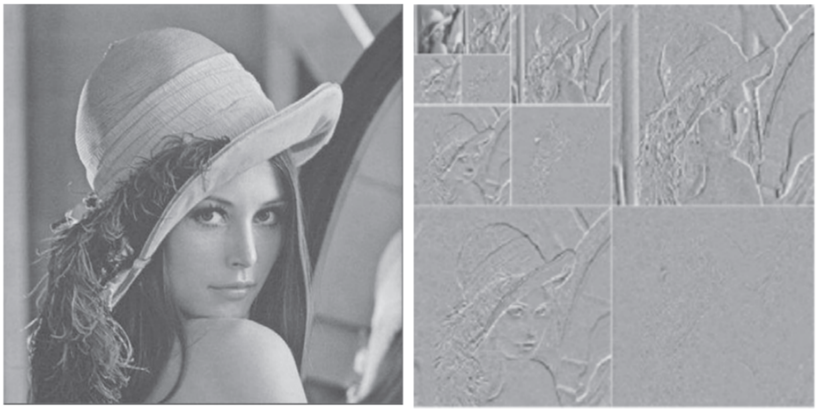


Figure 2: Lena with 3 levels DWT (Picture retrieved from: [Wan et al., 2022])

DWT can operate on both 1-D signals and 2-D images. A 2-D DWT can be seen as the combination of the two 1-D DWT in vertical and horizontal directions respectively. In each direction, a 1-D DWT scan through the image with both high pass filter and low pass filter and output two filtered results. After scanning through both vertical and horizontal directions, the input image could be divided into four non-overlapping multi-resolution sub-bands, which are Lower resolution approximation image (LL), Vertical (LH), Horizontal (HL), and Diagonal (HH).

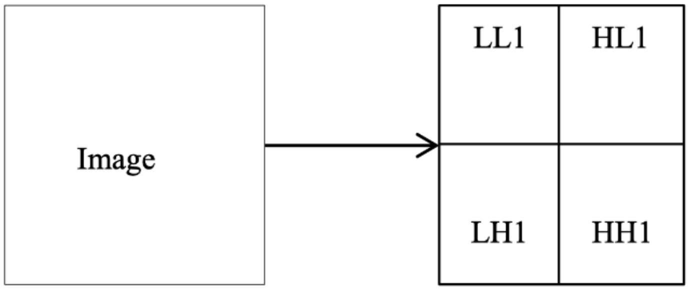


Figure 3: Concept of a single level DWT operation (Picture retrieved from: [Khan et al., 2013])

Since the sub-bands are generated through the combination of low pass and high pass filter, they remain some of the characteristics from the original image.

- Lower resolution approximation image (LL): Both row and column directions are scanned with low pass filter, which makes the result similar to the down-sampled (by a factor of two) version of the original image.
- Vertical (LH): The input image are scanned with low pass filter on horizontal direction and high pass filter on vertical direction, which makes the output preserve localized vertical features in the original image.

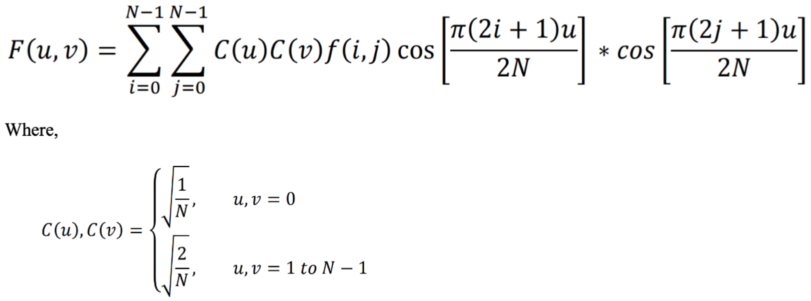
  - Horizontal (HL): The input image are scanned with high pass filter on horizontal direction and low pass filter on vertical direction, which makes the output preserve localized horizontal features in the original image.
  - Diagonal (HH): Both row and column directions are scanned with high pass filter, which isolates the high-frequency features in the image.

The human visual system (HVS) is sensitive to the changes in the smooth region of the image, and not sensitive to the tiny changes in edges. And the sharp variations (discontinuities) such as edges induce significant components in the high frequencies of the spectrum, which could be easily identified in the high frequency band through DWT, this is the reason why more studies use DWT to embed the watermark in the high level HH sub-band to ensure a better invisibility of the watermark.

DWT enables the selection of a specific part of the image which is hard to be perceived by the human eye. Our framework could strengthen invisibility by combining it with DWT. However, as the information of the watermark is going to be stored on the edges, it's questionable whether the framework can withstand a huge modification such as deepfake. One potential alternative is to implement different levels of DWT and see which level can be better against the attacks.

- Transform domain - Discrete Cosine Transform (DCT)

  Discrete Cosine Transform (DCT) is one of the most popular methods in the field of watermarks. Like DWT, DCT could be seen as a way to represent the original image as the sum of many cosine waves, which could be added back to the original image. Figure 4 shows the formula of a 2-D DCT, where f(i, j) denotes the input image, the input image will be transformed with cosine function in both vertical and horizontal direction, giving N $* N$ decomposite cosine waves. The result of performing DCT on Lena could be seen in Figure 5.



$$F(u,v) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} C(u)C(v)f(i,j) \cos\left[\frac{\pi(2i+1)u}{2N}\right] * cos\left[\frac{\pi(2j+1)u}{2N}\right]$$

Where,

$$C(u), C(v) = \begin{cases} \sqrt{\frac{1}{N}}, & u, v = 0 \\ \sqrt{\frac{2}{N}}, & u, v = 1 \ to \ N-1 \end{cases}$$

Figure 4: Equation of 2-D DCT (Picture retrieved from: [Khan et al., 2013]

One of the main differences between DWT and DCT in terms of watermark embedding is that DWT downsamples the original image so that the output of each sub-band is a quarter of the original size, while in each cosine wave generated by DCT, the size remains is the original input image. This is the reason why the studies such as [Khan et al., 2013] including DWT usually embed four watermarked images into the host image.

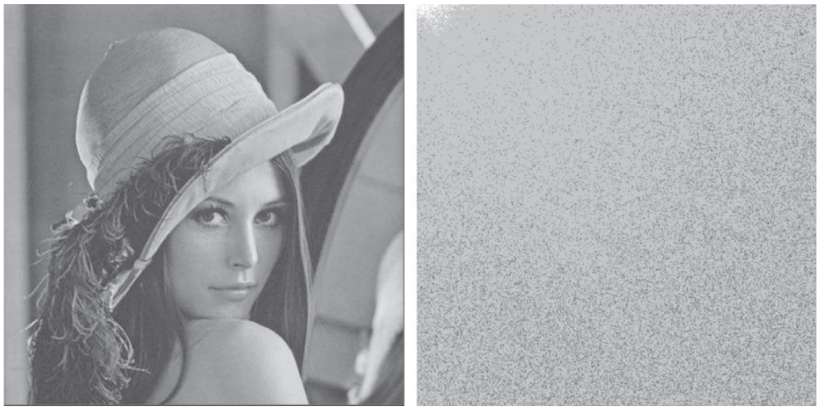


Figure 5: DCT on Lena (Picture retrieved from: [Wan et al., 2022]

Both DWT and DCT have good imperceptibility and robustness, which can meet the requirements of our framework. Some studies, such as [Khan et al., 2013], further combine the two and propose a new hybrid method that enhances imperceptibility and robustness.

2.2 Survey on the reversibility

The reversibility of the watermark was first proposed by [Barton, 1997] in 1997 by embedding the authentic information into the bits of the digital data block, similar to LSB method introduced above. Since then, a great deal of research in the field of reversible watermarking has been proposed, especially in medical and military domains where restoration of received data such as patient information is crucial [Jyothsna Devi et al., 2023, Devi et al., 2022].

The work of reversible watermarking technology can be roughly divided into two categories, one kind stores only the reversible information of the region of interest (ROI), and the other kind stores the information of the entire image. Each of them can be integrated with methods in the spatial domain or the transform domain depending on the requirements.

The ROI-based reversible watermark, usually used in the medical field, divides the input image into the region of interest (ROI) and the region of non-interest (RONI), and then embeds the information of ROI into the RONI area, so once the ROI is attacked, the hidden information in RONI can be extracted and restore the image. For example, [Liew and Zain, 2010, BW et al., 2012, Liew et al., 2013] all use the RONI to store the information in ROI. [Liew and Zain, 2010] divides the ultrasound image into several blocks, each ROI block corresponding to a RONI block, and the RONI block stores the information through both authentication bits and recovery bits. However, this method can only be used in predefined images such as ultrasound images and cannot be generalized. Same for [BW et al., 2012] and [Liew et al., 2013] which also use ultrasound image with LSB method.

However, the disadvantages of ROI-based reversible watermark is very obvious, if the modification is done in the RONI region, the image cannot be recovered, and what is worse, if the modification is only done in the RONI region, it cannot be detected as only the ROI region is protected.

Regarding the drawbacks in ROI-based reversible watermark, the full image based reversible watermark is proposed, [Selvam et al., 2017] combines integer wavelet transform (IWT) with discrete Gould transform (DGT). IWT is similar to DWT, which can divide the image into four sub-bands. DGT transforms an image using a lower triangular matrix, which releases one pixel of information each time DGT is applied to a 2 by 2 block, and the pixel information can be fully restored through inverse DGT.

It is worth noting that DWT is not considered a reversible watermarking technology, because in the process of embedding, the original integer value will be converted into a floating point number, which may be truncated during the process and cannot be guaranteed to be lossless. This is why IWT is used more than DWT for reversible watermarking.

2.3 Survey on the non-removability

The non-removability of the watermark is also considered to be an important feature in the proposed framework, which means that the watermark cannot be removed without destroying the original image. With this goal in mind, a study relates to the non-removability is conducted. However, few studies have emphasized the removal of watermarks, since watermarks themselves are designed to be difficult to remove, and is even more difficult with more complex, hybrid methods are proposed. Take [Khan et al., 2013] as an example, the watermark is ebedded in the transform domain which passed through multiple steps of filtration. First they use DWT and select only the HH sub-band, then the DCT is performed on the sub-band, the coefficient of the DCT is collected by zigzag scanning, making it spatially more disperse, and finally by the singular value decomposition (SVD). It is nearly impossible to extract or remove the watermark without knowing the exact embedding steps.

From the research we have done above, we can still summarize some learnings here. Since the methods in the spatial domain directly embed into the pixel bits, it's easier to be removed by scanning the entire image, while the methods in transform domain disperse the information in certain features, such as edges, textures, etc, which is difficult to capture all the watermarked pixels by scanning through the image. So if we aim to increase the difficulty of removing the embedded watermark, it's better to choose the method from transform domain than spatial domain.

3 Conclusion of the survey

This study investigates on several popular methods in the field of watermark in terms of their non-removability, reversibility and robustness. Through the investigation, we hope to propose a framework that can withstand the unethical use of the new tampering methods such as deepfake.

In terms of non-removability, methods in the transform domain are more suitable than those in the space domain. Most of the watermarks embedded with transform domain are difficult to remove without performing the embedding step in reverse, since the watermark information is actually assigned in certain features and scattered throughout the entire image. It is also important to know that most hybrid methods can increase the non-removability of the watermark, because it is more difficult to remove the watermark just by scanning the image.

In terms of reversibility, we find that a reversible watermark means that it is lossless after extraction. However, the main problem raised by deepfake is not image quality but unethical use, and while a detailed comparison of IWT and DWT would be great, there is no need to pursue fully reversible watermarking in this framework, which can also give us more flexibility in choosing the method.

Also, the concept of ROI and RONI from reversible watermarking is appealing, since the deepfake emphasizes on the modification of the facial features, it could be applicable if we can separate and embed watermarks in the background regions. This is also true for deepfake-modified videos, as we can filter the background regions by extracting the pixels that have not moved in the video.

Finally, in terms of the robustness of the watermark, most methods in the transform domain are resistant to some degree of attacks. However, most of the test cases in the studied journal were only slightly modified, i.e. the images were only tampered within a limited percentage of the entire image. This brings a lot of uncertainty to our framework as deepfake modifies a great portion of the image, and this issue needs to be testify carefully in the future.

We investigate different existing methods including LSB, DCT, DWT, IWT, DGT combined with other supporting methods such as ROI and RONI selection, singular value decomposition (SVD). [Khan et al., 2013] shows that by combining different methods together, the gains in robustness and invisibility can be huge, so we propose to combine different methods to improve performance.

The proposed framework is as follow:

1. Separate the image background from the objects as the region of interest (Note that it is important to design a good pairing between ROI and RONI so that the watermark does not require an additional key in the recovery step)
2. Perform different levels of DWT on the separated ROI (Note that every time DWT is performed, the output area will be down-sampled 4 times, so the balance between the level of DWT and the size of the watermark needs to be carefully determined)
3. Select the HH sub-band and perform DCT on it.
4. Further compress the information through DGT or SVD.

Here, I list interesting topics and learnings that can be left for future work.

- To defend against deepfake, it's not necessary to use reversible watermarking technology and pursue for lossless restoration. However, it's still good to investigate the difference between IWT and DWT.
- In addition to embedding techniques, other supporting methods, such as block selection and compression via SVD, DGT, etc, should also be emphasized as they have massive impact on the performance.
- Idea of the ROI, RONI may be a good way to prevent deepfake modification. However, it is important to find a general method that can be adapted to most scenarios, otherwise it is difficult to apply to the real world.

References

[Barton, 1997] Barton, J. M. (1997). Method and apparatus for embedding authentication information within digital data. *United States Patent, 5 646 997*.

[BW et al., 2012] BW, T. A., Permana, F. P., et al. (2012). Medical image watermarking with tamper detection and recovery using reversible watermarking with lsb modification and run length encoding (rle) compression. In *2012 IEEE International Conference on Communication, Networks and Satellite (ComNetSat)*, pages 167–171. IEEE.

[Devi et al., 2022] Devi, K. J., Singh, P., and Thakkar, H. K. (2022). Dual secured reversible medical image watermarking for internet of medical things. In *Connected e-Health*, pages 457–473. Springer.

[Fei et al., 2022] Fei, J., Xia, Z., Tondi, B., and Barni, M. (2022). Supervised gan watermarking for intellectual property protection. *arXiv preprint arXiv:2209.03466*.

[Gupta and Rama Kishore, 2022] Gupta, M. and Rama Kishore, R. (2022). Secured blind image watermarking using entropy technique in dct domain. In *Proceedings of Second Doctoral Symposium on Computational Intelligence*, pages 31–47. Springer.

[Jyothsna Devi et al., 2023] Jyothsna Devi, K., Singh, P., Santamaría, J., and Patel, S. (2023). Robust and secured reversible data hiding approach for medical image transmission over smart healthcare environment. In *Predictive Data Security using AI*, pages 119–131. Springer.

[Khan et al., 2013] Khan, M. I., Rahman, M., Sarker, M., Hasan, I., et al. (2013). Digital watermarking for image authenticationbased on combined dct, dwt and svd transformation. *arXiv preprint arXiv:1307.6328*.

[Kumar and Singh, 2021] Kumar, S. and Singh, B. K. (2021). Dwt based color image watermarking using maximum entropy. *Multimedia Tools and Applications*, 80(10):15487–15510.

[Liew et al., 2013] Liew, S.-C., Liew, S.-W., and Zain, J. M. (2013). Tamper localization and lossless recovery watermarking scheme with roi segmentation and multilevel authentication. *Journal of digital imaging*, 26(2):316–325.

[Liew and Zain, 2010] Liew, S.-C. and Zain, J. M. (2010). Reversible medical image watermarking for tamper detection and recovery. In *2010 3rd International Conference on Computer Science and Information Technology*, volume 5, pages 417–420. IEEE.

[Selvam et al., 2017] Selvam, P., Balachandran, S., Iyer, S. P., and Jayabal, R. (2017). Hybrid transform based reversible watermarking technique for medical images in telemedicine applications. *Optik*, 145:655–671.

[Singh and Gehlot, 2022] Singh, S. and Gehlot, A. (2022). A data hiding technique for digital videos using entropy-based blocks selection. *Microsystem Technologies*, pages 1–10.

[Sun et al., 2022] Sun, T., Wang, X., Zhang, K., Jiang, D., Lin, D., Jv, X., Ding, B., and Zhu, W. (2022). Medical image authentication method based on the wavelet packet and energy entropy. *Entropy*, 24(6):798.

[Wan et al., 2022] Wan, W., Wang, J., Zhang, Y., Li, J., Yu, H., and Sun, J. (2022). A comprehensive survey on robust image watermarking. *Neurocomputing*.

1.4