Reversible watermarking scheme for image tampering detection and restoration

Gabriel Melendez Melendez, René Armando Cumplido Parra

Technical Report No. CCC-20-004

August 14, 2020

Luis Enrique Erro 1

Sta. Ma. Tonantzintla,

72840, Puebla, México.

Contents

Abstract

Image tampering is a deliberate attempt to add or remove some image object without leaving any obvious traces of the manipulation. Image tampering detection methods arose to identify and localize tampered regions of a suspicious image. Due to the lack of active tampering detection methods to prevent these attacks, most images are exposed to malicious people whose objective is modify and use them with unethical purposes.

Image watermarking is a hiding information technique that has been thoroughly investigated by the cybersecurity community in recent years. A traditional watermarking scheme permanently distorts an image in order to hide information, therefore, reversible watermarking schemes (RWS) emerged for application domains where original image is wanted. RWS have the capability to extract hidden information and simultaneously remove the distortion introduced, such that original image is recovered.

In this report, we propose to develop a reversible watermarking scheme for image tampering detection and restoration. Most RWS designed to perform it only protect certain regions of interest and do not provide image restoration. Therefore, it is proposed to design a RWS that strategically embeds and extracts a watermark including authentication information to detect and localize the tampered regions of a watermarked image and also include recovery information to approximate the original image when tampered regions of a watermarked image are detected.

Preliminary results suggest that a compact version of the image can be created and used to approximate it with optimal quality when the relationship among neighboring pixels is maintained. Therefore, this compact representation of the image could be used as part of the watermark to perform image restoration.

1 Introduction

With the development of infrastructure and technology to communicate people, along with the large number of mobile devices, nowadays the number of images in social networks and the Internet have significantly increased. Most images could be exposed to malicious people, who aims to use it with unethical purposes.

Digital image processing tools, such as Photoshop, Corel Paint Shop, Photoscape, Photo-Plus, etc., are used to apply manipulations to original photographs in a simple way. Commonly, a region of the image is selected to be replaced by an object of the same image or an external image, this manipulation is known as tampering (Haghighi et al., 2018), which is considered as a category of image forgery. The aim is to deceive the human eyes to add or hide an object in the real scene. An example of image tampering is the Kerry-Fonda photograph, see Fig. 1. In the 2004 US presidential election, a controversial image was released, it showed the candidate John Kerry together with the anti-Vietnam war activist Jane Fonda in a protest. The image was politically motivated to affect the candidate Kerry, who was considered as a traitor for many US citizens after the image was created.

Figure 1: Kerry-Fonda image controversy, 2004 US presidential election (Zheng et al., 2019).

At the present time, it is common sharing images in the social networks, however, they are not protected and could be used to be tampered with. Fig. 2 shows another example of image tampering. Here some facial features of two women are swapped, resulting in two new images, which can be seen on the right side of the figure. The above examples clearly show that if the original images had not been displayed, one could not notice that the right-side images have been tampered, so tampering detection techniques are required in order to identify tampering and provide protection against these attacks.

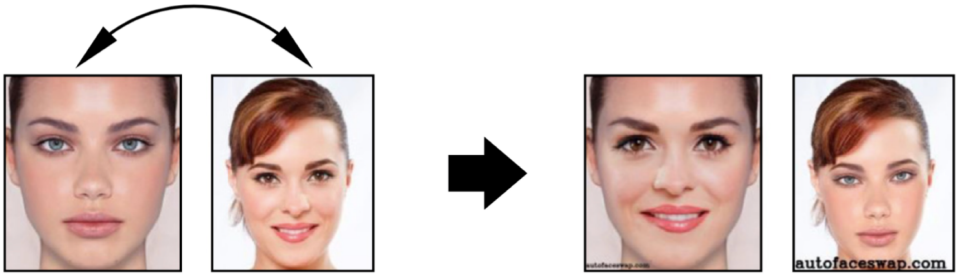

Figure 2: Face swapping example (Zheng et al., 2019).

Many tampering detection techniques have been identified in the literature, which can be classified into two approaches (Warif et al., 2016). On the one hand, the active approaches are used to obtain prior information about the image, using digital signatures (Lou and Liu, 2000; Xie et al., 2001) or image watermarking (Velumani and Seenivasagam, 2010; Wang et al., 2010; Lo and Hu, 2014). The obtained information is stored so that when it is required, it is used to verify if the corresponding image was tampered. On the other hand, the passive approaches (Warif et al., 2016; Zheng et al., 2019), which are able to identify a tampered image without knowing any previous knowledge about it, so image tampering is commonly identified by searching inconsistencies in the properties of the image using machine learning techniques.

Most image tampering detection methods are passive and they are used in digital forensics since most images are not protected, therefore, it is necessary to propose new active methods to prevent image tampering. Digital watermarking schemes offer three potential benefits over digital signature based schemes (Cox et al., 2002). First, the message is imperceptible. Sec-

ond, the message travels with its associated cover work and finally, the message undergoes the same transformations as the cover work in which it is embedded. These advantages are discussed in background section. Nevertheless, conventional digital watermarking schemes cause permanent distortion since the cover image is modified in order to embed the watermark. This problem can be overcome by using reversible watermarking approaches, which strategically embed the watermark to allow the full image restoration after the watermark is extracted.

Reversible watermarking schemes can be exploited by including recovery information as part of the watermark in order to approximate the corrupted regions of a watermarked image after it is tampered (Deng et al., 2013; Gao et al., 2018). These type of schemes could be very helpful to show evidence that a tampered image is not as it claims to be and it could be possible to show how the original image was like.

1.1 Report organization

This technical report is structured as follows: Section 2 provides the reader with some basics to better understand the topics addressed throughout this report including image tampering, digital watermarking and signal reconstruction. In Section 3 it is presented the state of the art on reversible watermarking field used to image tampering detection and restoration. Section 4 describes the research proposal. Finally, experiments, preliminary results and conclusions are discussed in Section 5.

2 Background

In this section it is first introduced what does image tampering means, then, digital watermarking subject is explained including main features, applications and the standard metrics used to evaluate the schemes. Finally, signal reconstruction techniques are briefly reviewed.

2.1 Image Tampering

According to Zheng et al. (2019), image manipulation refers to any modification that can be done to a digital image by software or applications using a computer or some digital device. Within image manipulation there are two well-known categories: steganography and image forgery, as it is shown in Fig. 3. On the one hand, steganography refers to the practice of undetectably altering an image to embed a secret message (Cox et al., 2002), so pixel-level operations are applied in such a way that the image is slightly altered and perceptual changes are not noticed by the human eyes when the original and modified images are compared, however, the changes can be perceived using an objective metric. On the other hand, image forgery emphasizes in content-level operations in order to deliver fake information about the scene in the image and trick the human eyes. Image tampering is an special type of image forgery typically used to hide an object of the scene or add a new object to the scene. There are three common types of image tampering:

1. **Copy-move**: A region of the image being tampered is duplicated on the scene. The duplicated region could undergo some geometrical operation such as scaling, rotation, etc.

2. **Cut-paste**: A region of an external image is pasted into the image being tampered, typically to hide some fact in the real scene.

3. **Erase-fill**: A region of the image being tampered is filled using near information of the scene (e.g., a texture, background, etc.).

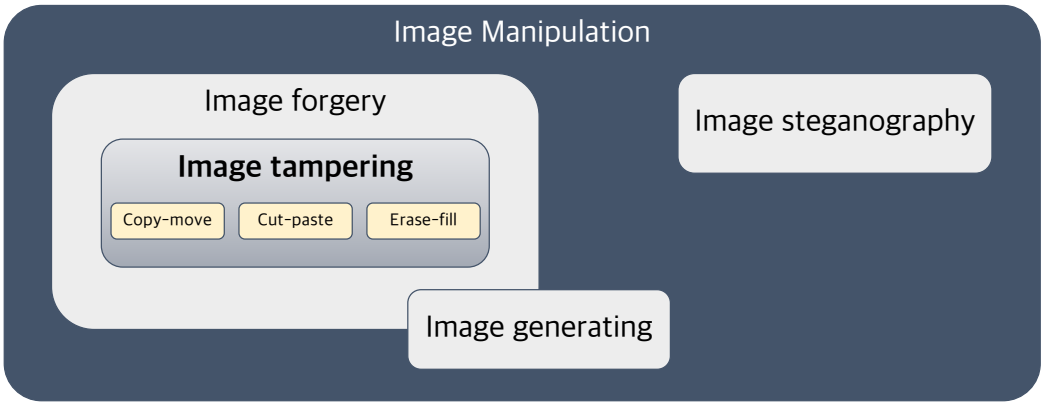

Figure 3: Image manipulation hierarchy.

Another type of image manipulation is image generating, which refers to the creation of images using computer software. As the created images might not show a real-world scene, image generating partially belongs to image forgery.

2.1.1 Tampering detection approaches

Tampering detection techniques can be classified into passive approaches and active approaches (Warif et al., 2016), as it is shown in Fig. 4. The passive approach analyzes the images to gather information of tampering. The problem can be seen as a binary classification task, where image's features are analyzed to find inconsistencies at image-level or region-level. Depending on the type of tampering applied, it is decided which characteristics of the image are analyzed. For example, copy-move is commonly detected by comparing blocks or key points of the image to find duplicated regions. To detect cut-paste, it is common to examine edge discontinuities, lighting and geometric inconsistencies, etc.

In the active approaches it is necessary to know prior information about the image, so image watermarking or digital signatures can be used to protect it. A digital signature is a cryptographic technique used to summarize some digital content depending on a user key, the signature must be stored so that when it is required the user can validate the integrity of the

digital content using the respective user key. Digital watermarking is used to hide a message into the image. The message is known as watermark and it provides useful information about the image.

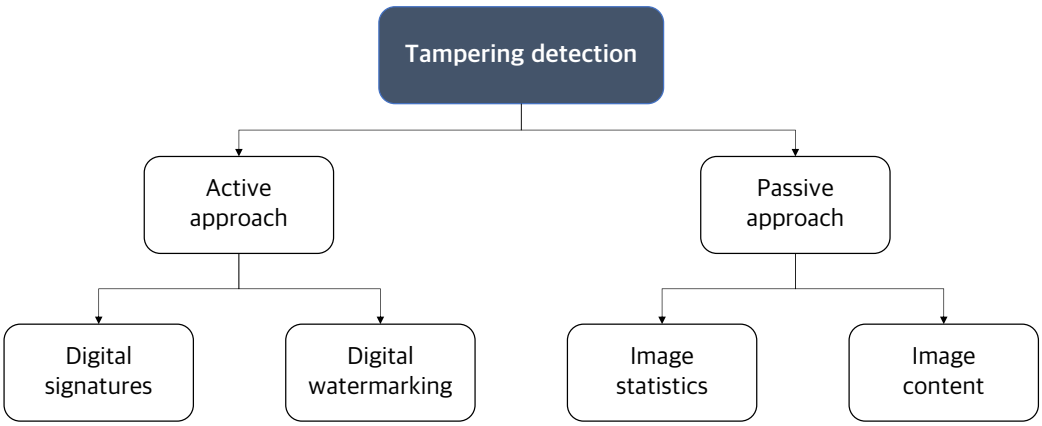

Figure 4: Tampering detection approaches.

Passive approaches are used in image forensics since they do not require any prior information about the image to detect tampering. On the other hand, active approaches are used to protect the images before they are tampered, nevertheless, as additional information can be added to the image using digital watermarking, a recent challenge is to include recovery information which is used to approximate those corrupted regions of a tampered image. Next sub-section explains digital watermarking subject.

2.2 Digital Watermarking

In recent years, progress related to communication technologies allows information exchange in a simple fashion. Thus, public and private organizations move part of their operations to an electronic environment. Protecting information in those communication systems is a task that can be addressed using security technologies such as cryptography, steganography or digital watermarking (Cox et al., 2008).

Digital watermarking is a research area which aims to embed a secret message into a digital content known as cover work, which can be an image, audio, video, document, etc. The message is a digital code known as watermark. Sometimes watermarking and steganography terms can be confused since both are information hiding techniques. However, there are some properties that identify each one. The most important difference is that in watermarking systems, the message provides useful information about its associated cover work, unlike steganographic systems where the message is not related to it (Cox et al., 2008). Here, an image is used as cover work, so a cover image with an embedded watermark is called a watermarked image.

Research interest related to digital watermarking techniques started in the 90s. Watermarking systems consist of two important processes: embedding and extraction. Fig. 5 shows a common digital watermarking system. Let w be a watermark and I_c the cover image in which w will be inserted using the embedding algorithm W_E, the corresponding watermarked image is defined to be $I_w = W_E(I_c, w)$. Typically, I_w is transmitted or recorded. At the receiver, I_w is received and the watermark w_r is obtained using the corresponding extraction algorithm W_D. So watermark is extracted as: $w_r = W_D(I_w)$, where $w_r = w$.

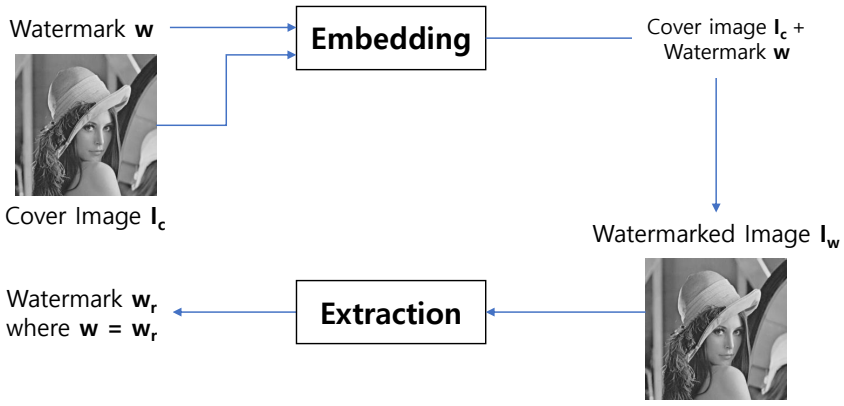

Figure 5: A general image watermarking system.

As additional information is stored into cover image, after embedding a permanent distortion is presented within the watermarked image. Although introduced distortion can be

minimal, in certain application domains such as medical, legal or military image processing, using the original image is mandatory for decision making and a permanent loss in the signal is prohibited (Shi et al., 2016). Reversible watermarking schemes (RWS) provide a solution to this problem since the embedding process is strategically performed, so that at the receiver, it is possible to extract the watermark and also the watermarked image can be recovered to its original form after extraction and restoration stages, using control information generated in the embedding process, see Fig. 6. Let RW_E be a reversible embedding algorithm. For a given cover image I_c and a watermark w, its corresponding watermarked image is defined to be $I_w = RW_E(I_c, w)$. At the receiver, both I_c and w are obtained by applying the corresponding reversible extraction algorithm as: $(I_r, w_r) = RW_D(I_w)$, where $I_r = I_c$ and $w_r = w$.

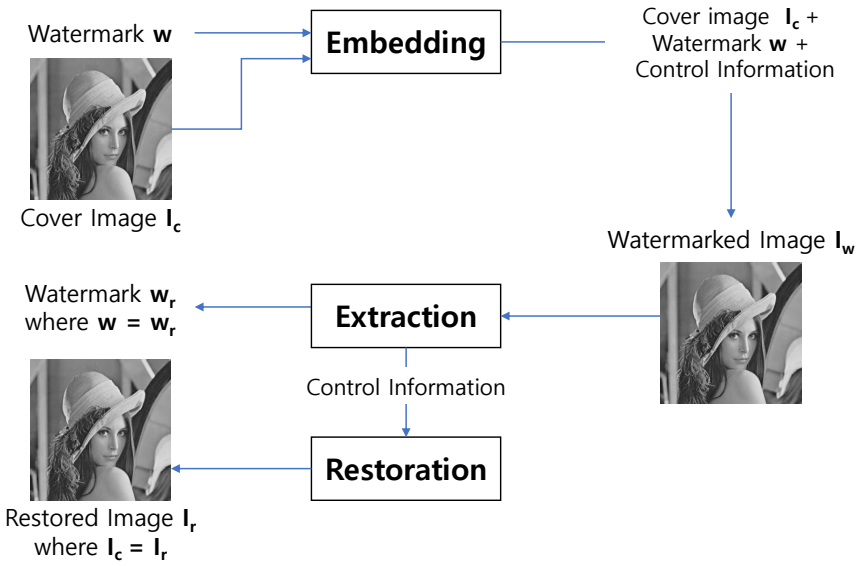

Figure 6: A general reversible image watermarking system.

2.2.1 Watermarking properties

Digital watermarking schemes must meet some requirements. Depending on the application for which the scheme is designed, some properties are prioritized over others. Below, the main properties that should be considered when designing a watermarking scheme are described (Cox et al., 2008).

- **Imperceptibility** - This property is very important since invisibility gives advantages to watermarking over other techniques to keep the information secret. Whenever a watermark is embedded into the cover image, it causes distortion. Watermarked image and cover image must be perceptually similar, thus introduced distortion should be the lowest possible.
- **Embedding capacity** - This property refers to the maximum number of information bits that can be embedded into the cover image.
- **Robustness** - Resistance level to attacks.

2.2.2 Applications

Watermarking techniques can be used in a wide variety of application domains since it is possible to hide additional information as part of the watermark. Watermarking is distinguished from other techniques by three important characteristics (Cox et al., 2008).

- Information embedding is performed in a way that cover media is minimally altered to avoid attracting the attention of malicious people.
- Watermark travels together with its associated cover work without the need of sending each item separately.
- The watermark undergoes the same transformations as the content in which it is embedded. So it is possible to learn something about transformations applied to the cover content, by analyzing the extracted watermark.

Therefore, watermarking schemes are ideal to be used in applications such as:

- **Broadcast monitoring** — In 1997, Japanese TV advertisers detected that thousands of their commercials were never shown on air despite having paid for them. Human observers were used to watch the broadcast and record what they saw or heard, however

it was costly and imprecise. Digital watermarking is an alternative to solve this problem. Identification information is embedded into TV and radio commercials to monitor transmissions using watermark detectors, which is less expensive and more effective for the advertisers.

- **Content Authentication** — Digital works could be easily altered by malicious people. Watermarking is used to verify the integrity of its associated cover work. Authentication is the process of identify if the received digital content is exact as it was sent (Saini and Shrivastava, 2014). This process can be performed using cryptographic techniques such as MACs or digital signatures, however as mentioned above, watermarking offers potential benefits over these techniques. In this case, watermarks are embedded into the cover work itself, removing any need of store or send the work and their authentication information separately.

- **Owner identification** — Digital content authors must protect their works (images, videos or songs) to claim them when copyrights are infringed. Watermarking schemes are used to embed the owner information into the works. By extracting the watermark, it is possible to identify the owner of a specific digital work.

- **Transaction tracking**- In this case, watermarks are embedded into a specific work in every transaction taking place in the copy history of the work itself. For instance, legal copies of a movie could be watermarked using a different watermark in each copy. If someone leaks a copy of the movie to the Internet, it could be possible to identify the user who leaked it by extracting the watermark.

- **Copy control**- Watermarking is also used to prevent illegal actions as copying of digital works. For example, a watermark could be embedded into a song or movie to notify a recorder (whit a watermark detector) that a copy is legal.

2.2.3 Evaluation metrics

Reversible watermarking schemes are evaluated considering imperceptibility and embedding capacity properties. Metrics to evaluate each property are described next.

Imperceptibility evaluation

As mentioned in previous section, embedding process causes distortion in the cover image, introduced distortion is noticeable when watermarked image quality is evaluated. Although sometimes a subjective analysis between cover and watermarked images is enough, in most cases it is necessary to use numerical metrics to measure distortion levels caused.

MSE

Distortion introduced in watermarking embedding process, can be obtained by computing the difference or distance between the original cover image and the watermarked image. Mean Squared Error or MSE is the simplest distortion metric to measure it (Eskicioglu and Fisher, 1995). The MSE is computed using Eq. 1.

$$MSE(I_c, I_w) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} (I_{c_{i,j}} - I_{w_{i,j}})^2 \tag{1}$$

Where:

I_c = Cover image

I_w = Watermarked image

N, M = Images size

If there are no differences between compared images, MSE will return 0, so a MSE closer to 0, implies a better watermark imperceptibility.

PSNR

Peak Signal-to-Noise Ratio or PSNR is the most commonly used metric to measure introduced distortion in decibels between cover and watermarked images (Khan et al., 2014). This metric is based on the MSE metric.

For a given cover image I_c and a watermarked image I_w of size $N \times M$, the PSNR between I_c and I_w is defined as Eq. 2.

$$PSNR(I_c, I_w) = 10log_{10}\frac{MAX^2}{MSE(I_c, I_w)} \tag{2}$$

Where MAX is the maximum value that image pixels can take, so for an 8-bit deep image this value is 255.

When I_c and I_w images are identical, MSE metric returns 0, therefore PSNR goes to infinite, so a higher PSNR value means a better watermark imperceptibility.

SSIM

Structural Similarity Index or SSIM is another objective metric usually used to evaluate image quality (Hore and Ziou, 2010).

SSIM is considered to be correlated to the human visual system (HVS), since this metric combine three components such as loss of correlation, luminance distortion and contrast distortion.

For a given cover image I_c and a watermarked image I_w, the SSIM measure between I_c and I_w is defined using Eq. 3.

$$SSIM(I_c, I_w) = l(I_c, I_w)c(I_c, I_w)s(I_c, I_w) \tag{3}$$

where:

$$l(I_c, I_w) = \frac{2\mu_{I_c}\mu_{I_w} + C_1}{\mu_{I_c}^2\mu_{I_w}^2 + C_1} \tag{4}$$

$$c(I_c, I_w) = \frac{2\sigma_{I_c}\sigma_{I_w} + C_2}{\sigma_{I_c}^2\sigma_{I_w}^2 + C_2} \tag{5}$$

$$s(I_c, I_w) = \frac{\sigma_{I_c I_w} + C_3}{\sigma_{I_c}\sigma_{I_w} + C_3} \tag{6}$$

Eq. 4 represents luminance mean. Eq. 5 is the contrast measured by standard deviation. Eq. 6 is the structure comparison, which measures correlation coefficient between cover and watermarked images. Finally, C_1, C_2 and C_3 are positive constants to avoid null denominator (Hore and Ziou, 2010).

SSIM metric is within $[0, 1]$ range. If there is no correlation between compared images, SSIM metric will return 0. On the other hand, if compared images are identical, SSIM will be 1. A SSIM value close to 1 corresponds to a better image quality.

Embedding capacity evaluation

Embedding capacity refers to a watermarking feature, which indicates the maximum number of information bits that can be embedded within a cover image.

Commonly this property is evaluated by directly indicating the number of embedded bits, however, it is possible to compute the number of bits per pixel or BPP that are inserted into the cover image (Cox et al., 2008).

Given a cover image I_c of size $N \times M$ with a maximum embedding capacity of P bits, the BPP is computed using Eq. 7.

$$BPP(I_c) = \frac{P}{N \times M} \tag{7}$$

Imperceptibility and embedding capacity are the most important properties when a reversible data hiding scheme is been designed Khan et al. (2014). Therefore, finding a trade-off among these two properties is essential, since the more the number of inserted bits, the introduced distortion increases affecting imperceptibility and the opposite occurs when less bits are inserted.

2.3 Signal reconstruction

Recovering a signal from a set of under-sampled measurements is a problem presented in this research since it is necessary to create a compact representation of the image, which will be embedded as part of a watermark to be used as recovery information when it is tampered. Due to embedding capacity limitations in reversible data hiding methods, the well-known sampling theorem (Nyquist, 1928; Shannon, 1949) is not suitable to create this representation, so another techniques need to be explored.

Compressed sensing (also known as compressive sensing) is a signal processing method used to acquire and fully reconstruct signals at sub-Nyquist rates (Donoho et al., 2006; Candes et al., 2006) considering some constrains. This problem can be stated as next.

Let $x \in \mathbb{R}^n$ be an unknown signal, which is sparse in a certain domain Ψ. In other words, x is defined as a K-sparse signal if only has K nonzero coefficients in the Ψ domain. Let $\Phi \in \mathbb{R}^{m \times n}$ be a measurement matrix that has i.i.d.(independent and identically distributed) inputs. Measurement matrix Φ contains far fewer rows than columns, that is to say, $m \ll n$. Compressed sensing measurements are obtained by using Eq. 8, where $y \in \mathbb{R}^m$.

$$y = \Phi x \tag{8}$$

The purpose of compressed sensing theory is to recover the sparse signal x using the set of linear measurements y, see Fig. 7.

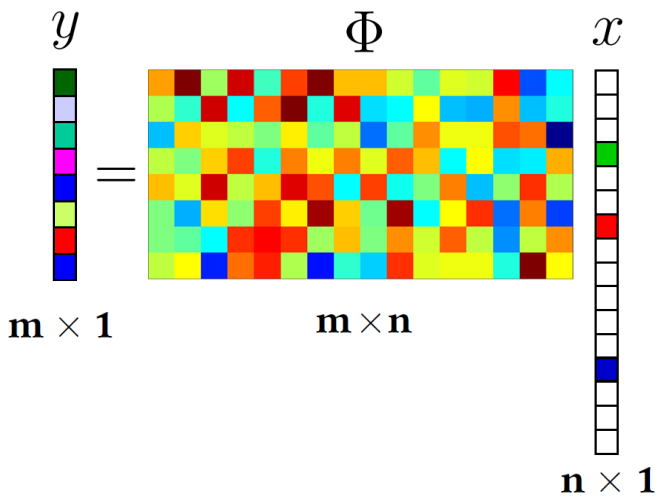

Figure 7: Compressed sensing measurements y obtained from the sparse signal x using a measurement matrix Φ.

Since y is a lower dimension vector compared to x, it is impossible using the inverse transform to get the original signal x, and consequently signal reconstruction has to be done by solving the optimization problem in 9.

$$\min \|\Psi^T x\|_1 \text{ subjected to } y = \Phi x \tag{9}$$

Therefore the reconstructed signal x is among all signals that generates the same measured data, which has transform coefficients with the minimal ℓ_1 norm.

This problem can be solved by using greedy algorithms. The most used greedy algorithms are basis pursuit(BP) (Ekanadham et al., 2011), matching pursuit(MP) (Mallat and Zhang, 1993) and orthogonal matching pursuit (OMP) (Tropp et al., 2007).

Next section describes the principal reversible watermarking schemes used to image tampering detection and restoration identified in the literature.

3 State of the art

In this section some active tampering detection and image restoration methods based on reversible watermarking are presented. Reversible watermarking schemes have received increasing research interest in recent years (Khan et al., 2014). The first solution when addressing the hiding information problem in a reversible way was proposed in Barton (1997). The patent exposed a method to embed authentication information within a digital data block. So an image is partitioned in pixel blocks to compute a digital signature from the high-order bits. The LSBs (Least Significant Bits) of block pixels are modified to embed a bit string composed by a block's digital signature and a compressed version of the original bits that were altered. The insertion of compressed original bits allows reversibility while digital signatures provide authentication. Since then, many reversible watermarking schemes have been proposed to be used in different application domains, using new, extended or improved versions of existing watermarking embedding techniques. Fig. 8 shows the identified works to perform tampering detection, some of which include image restoration.

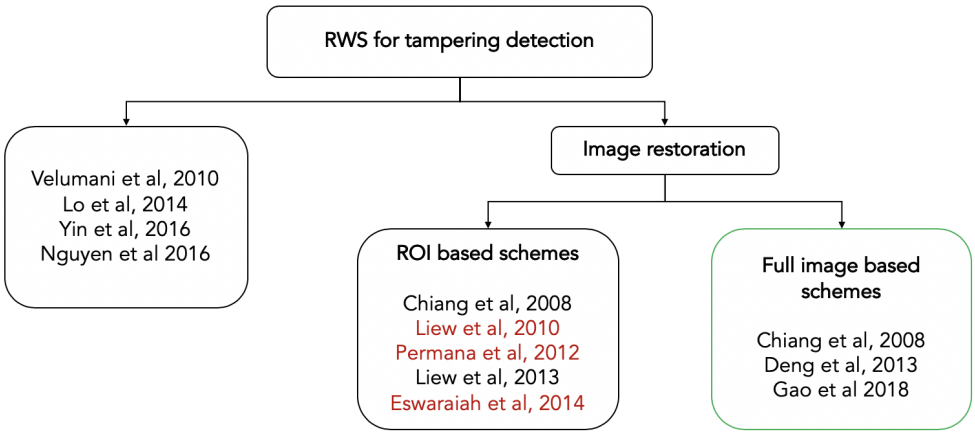

Figure 8: Reversible watermarking works for tampering detection.

On the one hand, works that only perform tampering detection usually divide the cover image into blocks to embed a watermark containing authentication information, which can be generated using seeded pseudo-random number generators (Lo and Hu, 2014; Yin et al.,

2016; Nguyen et al., 2016), or by using an image as a fragile watermark[1] (Velumani and Seenivasagam, 2010).

On the other hand, works that perform both, tampering detection and image restoration include some information about the image as a part of the watermark in order to be recovered and used to approximate the corrupted regions of a tampered image. Some of this works protect only regions of interest (ROI) and less works protect the entire image. So ROI-based works are usually used in the medical fiel. This works divide the cover image into regions of interest and regions of not interest (RONI), then, ROIs and RONIs are divided into blocks of pixels. Every RONI block contains information about a respective ROI, so these regions are protected. Liew and Zain (2010) include as as part of the watermark an authentication bit, a parity bit and the pixel average of a ROI block as recovery information. Permana et al. (2012) embed the watermark into ROI blocks, and RONI blocks are used to back up the original ROI values. A sequence $A \rightarrow B \rightarrow C \rightarrow ... \rightarrow A$ is created using ROI blocks, so the authentication information of block A is embedded into block B, the authentication information of block B is embedded into block C and so on. The LSB of ROI blocks are recorded into two LBSs of RONI blocks to ensure the restoration of the image. Liew et al. (2013) proposed a ROI-based scheme using multilevel authentication. The watermark is created with authentication codes by computing the cyclic redundancy check (CRC) for each ROI block. ROI original values are embedded into the LSBs of the RONI blocks to perform image restoration.

Some identified drawbacks on ROI-based works are the next. First, if a RONI of the watermarked image is modified, then, the changes on it could not be detected since only ROIs are protected, so it is mandatory to protect the complete image regions. Second, it is necessary to have smaller ROIs than RONIs to ensure the protection of ROIs by embedding the authentication information or backups into the RONIs, then, the number of protected ROIs is as small as possible. Finally, ROI-based schemes are not completely reversible since the information in RONIs is not restored at all, so it is desired to remove the distortion caused by the embedding algorithm in the entire image.

[1]The watermark is distorted when the watermarked image is modified.

To overcome the above mentioned issues, some full image based reversible watermarking schemes for tampering detection and restoration have been proposed. Chiang et al. (2008) divide a cover image into blocks of size 4×4. The watermark is created by computing the pixel average by block as recovery information and a pseudo random binary sequence as authentication information. The watermark is embedded into the low frequency coefficients of the integer wavelet transform (IWT) using prediction error expansion technique. Deng et al. (2013) partition a cover image into blocks using a quad tree decomposition, which return homogeneous pixel blocks. Two watermarks are created, the first watermark incorporates the recovery information by computing linear interpolations of the image. The second watermark contains the quad tree information and essential parameters to be used in the extraction algorithm. The watermarks are embedded into the frequency and spatial domains respectively. Recently, Gao et al. (2018) propose a reversible scheme based on compressed sensing reconstruction. A cover image is partitioned into non-overlapping blocks of size 32×32. The watermark is created by incorporating reference values obtained from the coefficients of the discrete cosine transform (DCT) as recovery information. To detect and localize tampering, a digest obtained by using SHA-256 hash function is added to the watermark. The watermark is embedded into the spatial domain by using prediction error expansion embedding technique. The scheme allows image restoration by extracting the watermark and using the non-tampered reference values as input to the basis pursuit algorithm to approximate the original reference values.

Author	Year	Tampering	Restoration	Embedding domain
Chiang et al.	2008	Yes	Yes	Frequency, IWT
Velumani and Seenivasagam	2010	Yes	No	Space
Liew and Zain	2010	Yes	ROI	Space
Permana et al.	2012	Yes	ROI	Space
Liew et al.	2013	Yes	ROI	Space
Deng et al.	2013	Yes	Yes	Frequency, IWT - Space
Lo and Hu	2014	Yes	No	Space
Eswaraiah and Reddy	2014	Yes	ROI	Space
Yin et al.	2016	Yes	No	Space
Nguyen et al.	2016	Yes	No	Frequency, DWT
Gao et al.	2018	Yes	Yes	Space, PEE

Table 1: Identified related work.

Table 1 shows a summary of the identified state of the art works. From the above, it is showed that reversible watermarking schemes are a prominent approach to the tampering detection and image restoration tasks. There are areas of opportunity to work in the subject such as, increasing the size of tampered regions, improving the approximation of a restored image after it was tampered and improving the accuracy of tampering detection.

4 Research Proposal

4.1 Problem Statement

In recent times, a large number of digital image editing tools are used to apply manipulations in images, specifically image tampering (Zheng et al., 2019). Therefore, images can be easily manipulated to be used with unethical purposes.

Due to the lack of active image tampering detection methods, most images are not protected against these attacks, therefore, most digital forensics use passive approaches. However, passive methods are less accurate when detecting tampering because the criteria to decide if an image was tampered or not rely on the analysis of image features and statistics.

Reversible watermarking schemes have attracted research interest in recent years to provide image tampering detection as it was presented in Section 3. Reversible watermarking schemes are desired since they are able to remove all distortion introduced by the embedding process. More advantages of using a reversible watermarking scheme for image tampering detection are listed bellow.

- **More effectiveness in tampering detection problem**. As authentication information about the image is included into the watermark, tampering detection is performed using this knew information.
- **Detection of different tampered regions**. A reversible watermarking scheme can be designed to protect all image regions. So, if different regions of the image are tampered with, the method can detects those tampered regions by using the corresponding authentication information, unlike passive approaches where typically fail when detecting two attacks of the same type in different regions of the image.
- **Different attacks detection**. Copy-move, cut-paste and erase-fill attacks could be detected using the same reversible watermarking method, without the need of designing

a different method for each type of attack as happen with passive approaches.

- **Restoration of corrupted regions**. As additional information can be included in the watermark, a reversible watermarking scheme can be designed to include recovery information to approximate the damaged regions of a tampered image.

New reversible watermarking schemes are desired to protect images in applications domains where permanent distortion is prohibited. Developing a new reversible watermarking scheme for image tampering detection and restoration is challenging because it must meets with next requirements.

- **Reversibility** - All watermark bits embedded into the cover image must be removable from the watermarked image to ensure reversibility.
- **Tampering detection** - All image regions must be protected so that when watermarked image is tampered the proposed tampering detection strategy can accurately localize all damaged regions of the image.
- **Image restoration** - On the one hand, embedding and extraction strategies need to be designed in order to include the enough recovery information as part of the watermark, considering the limitations regarding to embedding capacity offered by existing reversible data hiding techniques. On the other hand, it is required to design an image reconstruction technique, which is able to approximate damaged regions of a tampered image using the recovery information obtained by the watermark extraction algorithm.

As it is discussed in Section 3, most reversible watermarking schemes only protect one or more Regions of Interest (ROI) in the image and do not provide image restoration. Considering the above, it is necessary to develop a new scheme allowing both, the complete image protection and the restoration of corrupted regions when it is tampered.

4.2 Hypothesis

It is possible to improve existing image tampering detection and restoration methods based on reversible watermarking by designing a proper watermark including authentication information and a new compact representation of the image which is strategically embedded and extracted to be used as recovery information together with a signal reconstruction technique.

4.3 Research questions

- Which image's information should be considered to be embedded as a watermark to allow image tampering detection and restoration?
- Which reversible watermarking technique provides enough embedding capacity to incorporate the designed watermark into the image?
- How does the watermark should be embedded into the image and extracted from it in such a way that if a watermarked image is tampered, as much information as possible could be recovered?
- Which signal reconstruction technique should be used together with the extracted recovery information to perform image restoration?
- What is the maximum percentage level of tampering attacks supported by the scheme?

4.4 General objective

- To design and develop a reversible watermarking scheme to allow image tampering detection and restoration of corrupted regions within tampered images.

4.4.1 Specific objectives

- Analyze and select the image information that will be used as a watermark to perform tampering detection and restoration.
- Revise, analyze and select a reversible watermarking technique that provides the enough embedding capacity to incorporate the designed watermark.
- Develop a reversible watermarking scheme to perform image tampering detection.
- Design and develop a strategy to perform image restoration after tampered regions in the watermarked image are detected.

4.5 Methodology

The following methodology is proposed to achieve the objectives of this research.

1. **Design of the watermark.**
   In this step, it will be studied the information that could be used as a watermark w to be embedded into the cover image I_c in order to provide tampering detection and image restoration.

   - **Recovery information**: The recovery information w_r must be a compact representation obtained from I_c that can be used to approximate it. As recovery information w_r will be the largest part of w, recovery strategies will be first explored. Recovery strategies will be evaluated by reconstructing the original image using the corresponding compact information. Best recovery strategy will be selected according to the quality between original and recovered images using PSNR and SSIM metrics.
   - **Authentication information**: In order to perform tampering detection with high localization precision, the watermark must include region level authentication information w_a. A cryptographic technique can be applied to image regions to create

a code w_a, in such a way that if some image region is modified, the same cryptographic technique will return a different code.

At the end of this step, $w = w_r \oplus w_a$ will provide the embedding capacity necessary by a reversible data hiding scheme to incorporate the watermark into the image.

2. **Review and selection of reversible embedding and extraction algorithms.**
The next step is to review reversible data hiding schemes. As designed watermark contains recovery and authentication information, a high embedding capacity reversible data hiding scheme is required. Data insertion and extraction algorithms that will be used in the proposed scheme, will be selected according to embedding capacity and imperceptibility properties. The designed watermark w must be full incorporated into I_c using the selected reversible embedding algorithm.

3. **Develop a reversible watermarking scheme for image tampering detection.**

Here a tampering detection scheme will be designed and implemented using the selected reversible embedding and extraction algorithms and the authentication codes w_a designed to be included into w.

In this step the authentication information w_a is embedded into the cover image I_c using the selected reversible embedding algorithm to obtain the watermarked image I_w. An strategy to embed the watermark will be developed in order to provide tampering detection. Watermark imperceptibility levels will be obtained by using PSNR and SSIM metrics between I_c and I_w.

Next I_w will undergo copy-move, cut-paste and erase-fill attacks at different percentages regarding to the image size to obtain $I_w^{'}$. PSNR and SSIM values between I_w and $I_w^{'}$ will be obtained to measure the distortion introduced by the attacks.

The corresponding extraction algorithm will be used to obtain a watermark $w^{'}$. Finally, w' will be used to identify the damaged regions of $I_w^{'}$. The quality between embedded w and extracted $w^{'}$ will be measured using bit error rate (BER) metric. Finally, the image tampering detection accuracy will be evaluated by using false positive rate metric.

4. **Include image restoration.**

   In this step, the reversible watermarking scheme designed to tampering detection will be extended to incorporate the recovery information w_r designed in step 1 as a part of the watermark, which will be used to recover $I_w^{'}$ after tampered regions are detected.

   First, the watermark w will be extended to include w_r, i.e., the compact representation of I_c. Embedding capacity required to incorporate it was tested in step 2, therefore, all bits of w can be embedded into I_c to obtain I_w. An strategy to embed w_r will be proposed in order to recover as much information as possible when I_w is tampered. Distortion levels will be obtained by using PSNR and SSIM metrics between I_c and I_w.

   Next I_w will undergo the same attacks as in step 3 to obtain $I_w^{'}$ and damaged image regions will be identified. PSNR and SSIM metrics will be used to measure the distortion introduced by the attacks.

   Once corrupted regions are detected, the corresponding recovery information w_r is used together with the image reconstruction strategy selected in step 1 to approximate these damaged regions. The similarity between the original image I_c and the recovered image I_n will be computed using PSNR and SSIM metrics. Finally, the proposed scheme will be compared against related works, considering tampering detection accuracy and imperceptibility among the cover image, the watermarked image and the reconstructed image.

4.6 Expected contributions

The expected contributions of this research are:

- A new strategy to create a compact representation of the cover image to be used as recovery information. Although this representation is compact, it must contain enough information needed to approximate damaged regions of a tampered watermarked image with high accuracy.

- A new reversible watermarking scheme for image tampering detection with embedding and extraction strategies to protect the complete image against copy-move, cut-paste and erase-fill attacks.

4.7 Limitations

This research is mainly focused on overcome the tampering detection and image restoration problems by using a reversible watermarking approach. Therefore, reversible data embedding and extraction algorithms will not be designed, instead existing algorithms will be reviewed and selected to be strategically used.

Experiments will be applied into standard natural images[2], therefore, this research will not be linked to another image domain such as encrypted images, synthetic images, etc.

4.8 Activities schedule

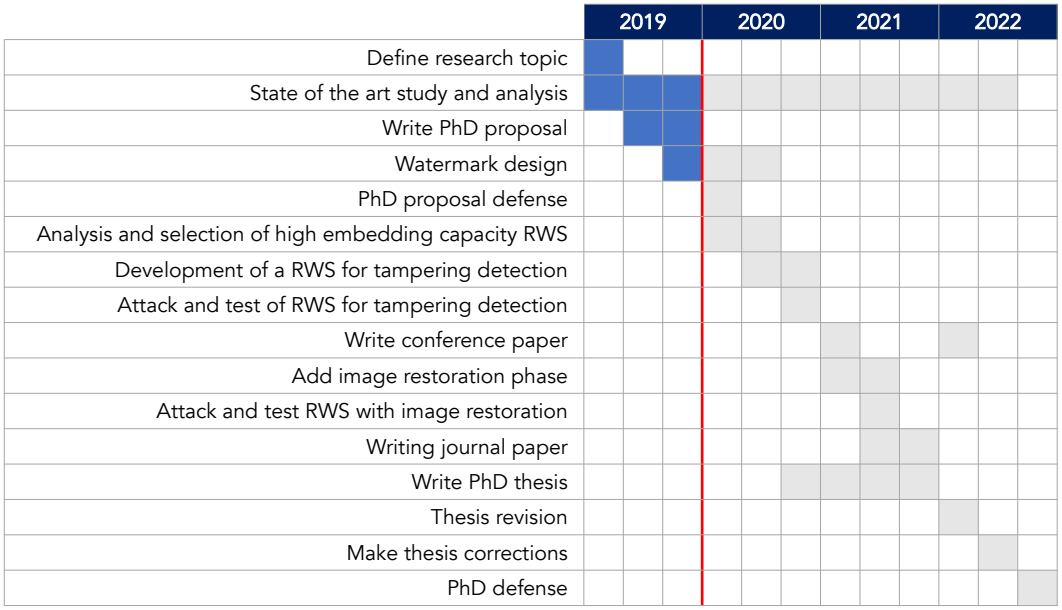

Figure 9: Activities schedule.

[2]Images that capture somewhere in the world by using a common digital camera (Hyvärinen et al., 2009)

5 Preliminary Results

This section presents the experiments and achieved preliminary results for this technical report. The next experiments focuses on the recovery information that could be included as part of the watermark to approximate damaged regions of the image.

Due to the limitations related to the embedding capacity of existing reversible watermarking schemes, it is necessary to get an image sampling as small as possible to be used as recovery information. Most reversible watermarking schemes use block pixel average to perform it. However, as block size increases, more information is lose. Compressed sensing is a recent theory that can be used to signal recovery (Donoho et al., 2006). Therefore, we explore compressed sensing theory to get a highly incomplete representation of an image and use this representation to approximate the original one.

In these experiments, the signal $x \in R^n$ is recovered from a small number of linear measurements $y = \Phi x$, where $y \in R^m$, Φ is the measurement matrix and $m \ll n$. A 256×256 grayscale image is pre-processed to create pixel vectors of size n following three different image scanning strategies, see Fig.10. Each vector is used as input to the compressed sensing algorithm. Then a smaller representation y is obtained. Image recovery is performed by using l_1-magic algorithm (Candes and Romberg, 2005), which solves the linear system Φb$= y$ using a convex optimization algorithm that minimizes the l_1 norm of b to approximate the original signal x, this algorithm is also known as basis pursuit. PSNR and SSIM metrics are used to evaluate the quality between the original and recovery images.

Figure 10: Image *cameraman* used in the experiments.

5.1 Non overlapping blocks scanning

In this experiment, the image is divided into non-overlapping blocks of fixed sizes. Every block is reshaped to a one-dimensional array, which is used as input to the compressed sensing algorithm.

Figures 11,12, 13, 14 and 15 show some recovered images using blocks of sizes 128×128, 64×64 and 8×8 according to different number of linear measurements (m) used to recover original image. On the left side, it can be observed the approximations by least squares method. On the the right side, it is shown the basis pursuit algorithm approximation.

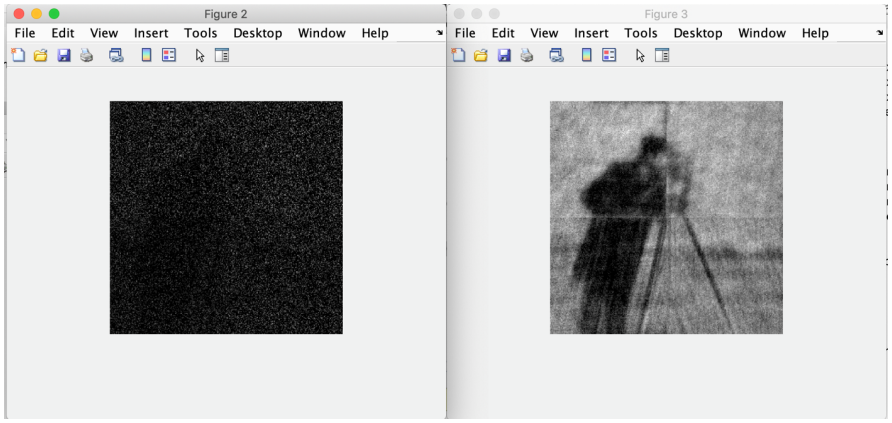

Figure 11: Recovered images setting $n = 16,384$ and $m = 1,600$.

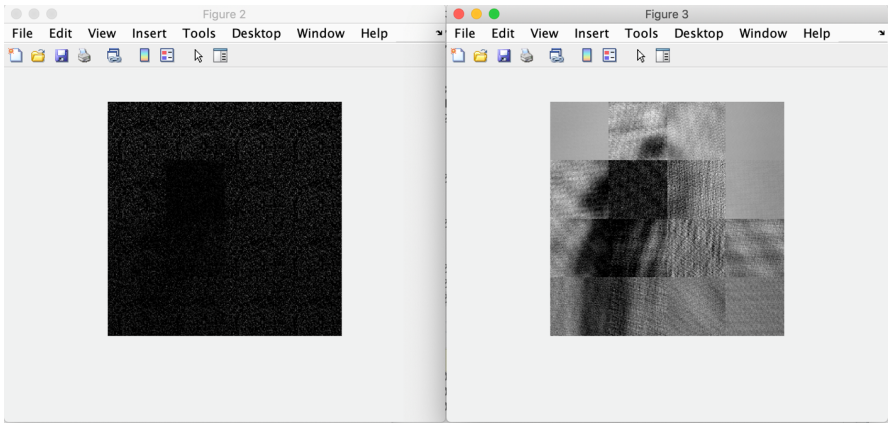

Figure 12: Recovered images setting $n = 4,096$ and $m = 200$.

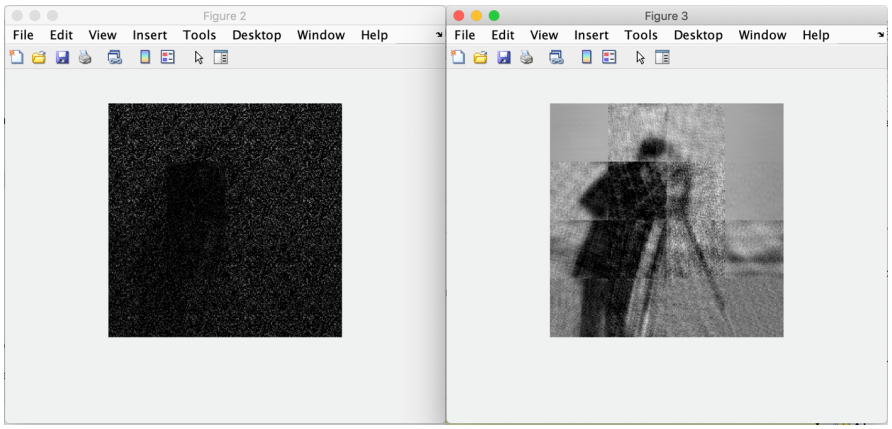

Figure 13: Recovered images setting $n = 4,096$ and $m = 400$.

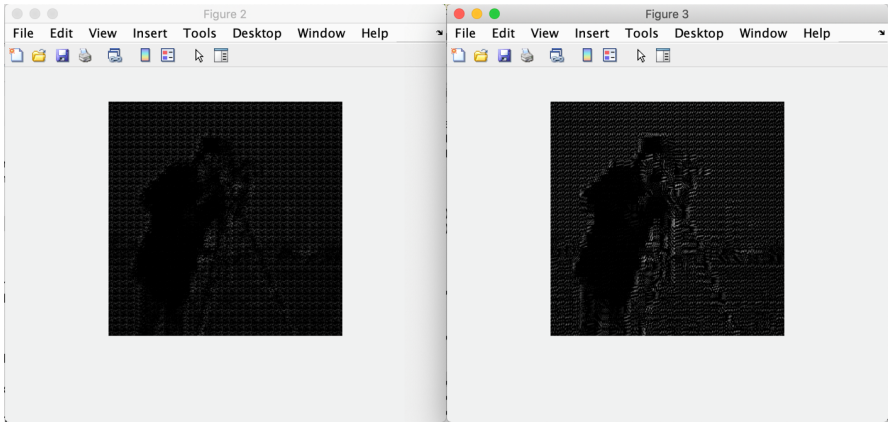

Figure 14: Recovered images setting $n = 64$ and $m = 6$.

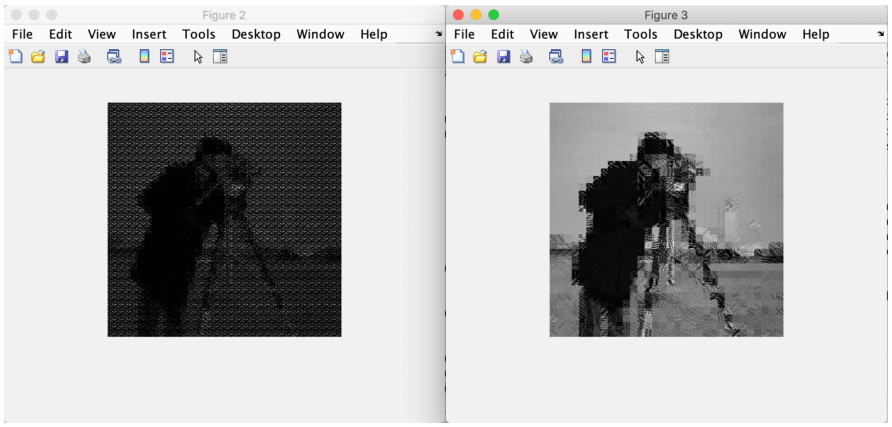

Figure 15: Recovered images setting $n = 64$ and $m = 12$.

Block size	n	m	Least squares PSNR	Least squares SSIM	Basis pursuit PSNR	Basis pursuit SSIM
128×128	$16,384$	$1,600$	6.8741	0.0473	18.3158	0.2236
64×64	$4,096$	200	6.4023	0.0598	16.9675	0.2968
64×64	$4,096$	400	6.8442	0.0546	18.1831	0.3506
32×32	$1,024$	180	7.6276	0.0766	19.6495	0.5287
32×32	$1,024$	280	7.9681	0.0861	21.6736	0.6097
8×8	64	6	6.3270	0.0512	6.9393	0.0669
8×8	64	12	7.8549	0.0979	18.7946	0.6140

Table 2: Quality of recovered images in terms of PSNR and SSIM using non-overlapping blocks scanning.

The results in terms of PSNR and SSIM are shown in Table 2. From the above, it can be observed that when more linear measurements are used, the recovered images' quality is increased. Better PSNR values with minor variations are obtained when m is about 20% of n. However, SSIM metric indicates that a better approximation is achieved when image is divided into 32×32 sized blocks and 280 linear measurements are obtained by compressed sensing algorithm.

5.2 Hilbert curve scanning

In this experiment image is scanned by using a continuous fractal space-filling technique called Hilbert curve (Lawder, 2000). Hilbert curve is used to pass through every pixel of the image once and in some particular order, see Fig. 16. First, pixels are mapped to a one-dimensional vector, which is divided into non-overlapping sets of size n to be used as input to compressed sensing algorithm. Later, obtained representations of size m are used to recover the image. Figures 17, 18, 19 and 20 show some recovered images obtained by solving the linear systems by least squares and basis pursuit algorithms.

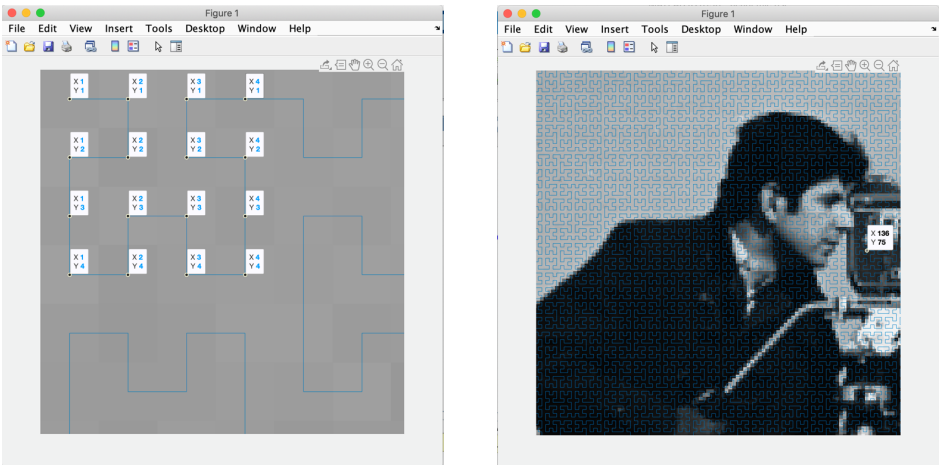

Figure 16: Hilbert curve scanning in *cameraman* image.

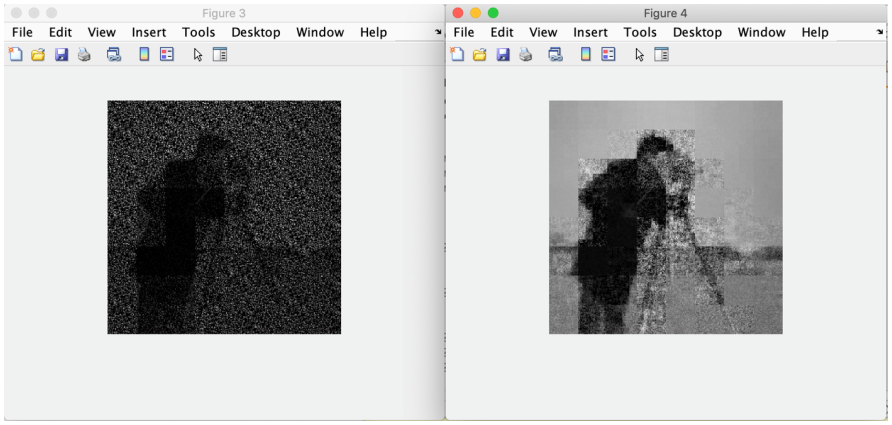

Figure 17: Recovered images setting $n = 1,024$ and $m = 180$.

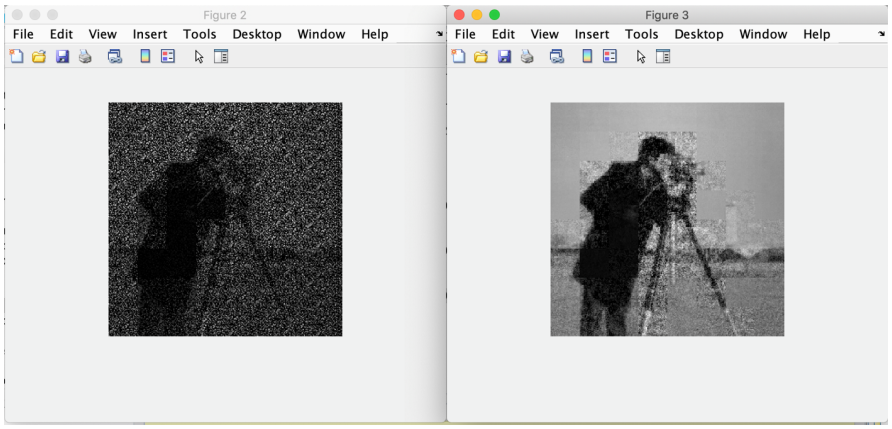

Figure 18: Recovered images setting $n = 1,024$ and $m = 280$.

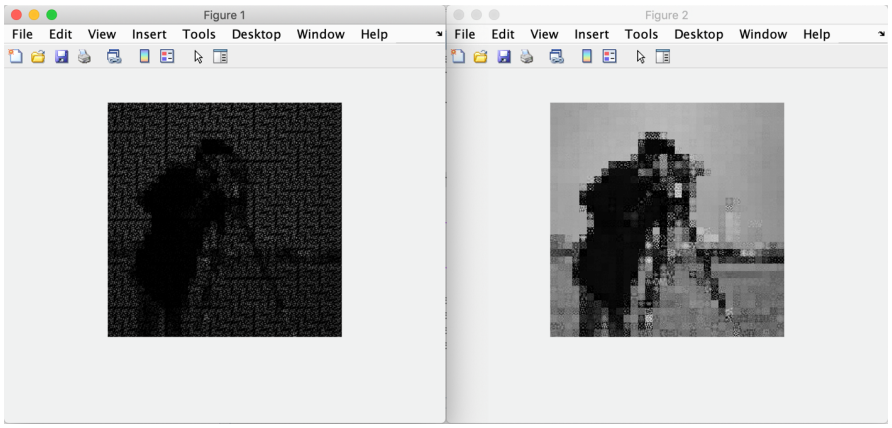

Figure 19: Recovered images setting $n = 64$ and $m = 6$.

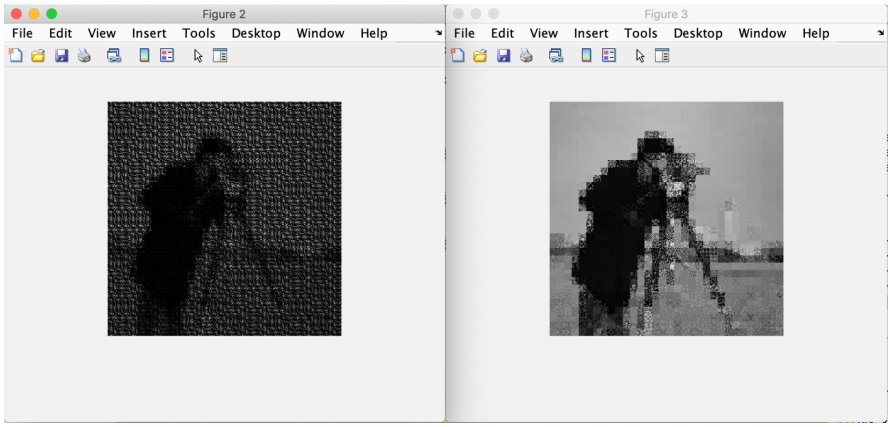

Figure 20: Recovered images setting $n = 64$ and $m = 12$.

		Least squares		Basis pursuit	
n	**m**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
$16,384$	$1,600$	6.9077	0.0499	17.4198	0.1817
$4,096$	200	6.4369	0.0609	16.1670	0.2620
$4,096$	400	6.8928	0.0558	17.2609	0.3012
$1,024$	180	7.6726	0.0768	18.8093	0.4903
$1,024$	280	8.1740	0.0903	20.3166	0.5553
64	6	7.0715	0.0753	17.0998	0.5313
64	12	7.7504	0.0971	18.2872	0.5861

Table 3: Quality of recovered images in terms of PSNR and SSIM using Hilbert curve scanning.

Table 3 shows the image recovery results in terms of PSNR and SSIM. From the above results, it can be observed that using Hilbert curve scanning provide similar reconstruction results with minor variations as those obtained by using directly non-overlapping blocks. However, when sets of size $n = 64$ are used to create $m = 6$ linear measurements, the PSNR and SSIM values significantly increase compared with non-overlapping blocks scanning. The best PSNR and SSIM values are obtained when sets of size $1,024$ and 280 linear measurements are used to recover image.

5.3 Pseudo-random scanning

In this experiment, all image pixels are mapped row-by-row into a one-dimensional array. The obtained array is pseudo-randomly permuted and divided into sets of size n, which are used as input to compressed sensing algorithm to obtain representations of size m. Then, every representation is used to approximate their corresponding part of the image. Figures 21, 22 and 23 show some recovered images by using least squares and basis pursuit algorithms to solve the linear systems.

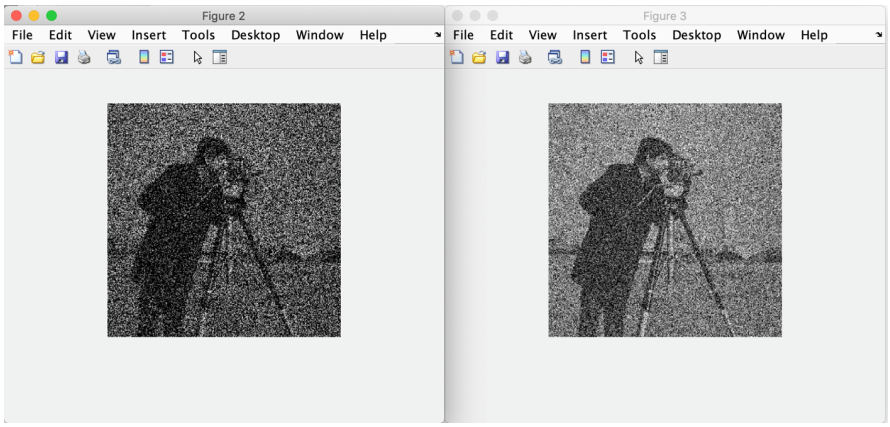

Figure 21: Recovered images setting $n = 1,024$ and $m = 400$.

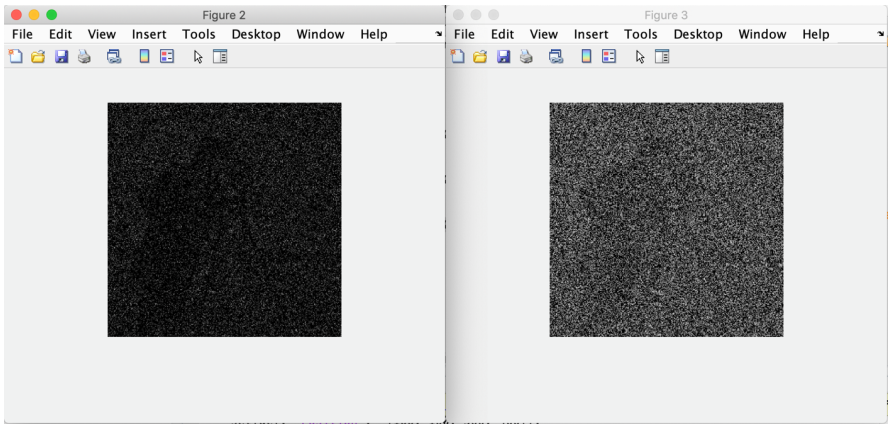

Figure 22: Recovered images setting $n = 64$ and $m = 6$.

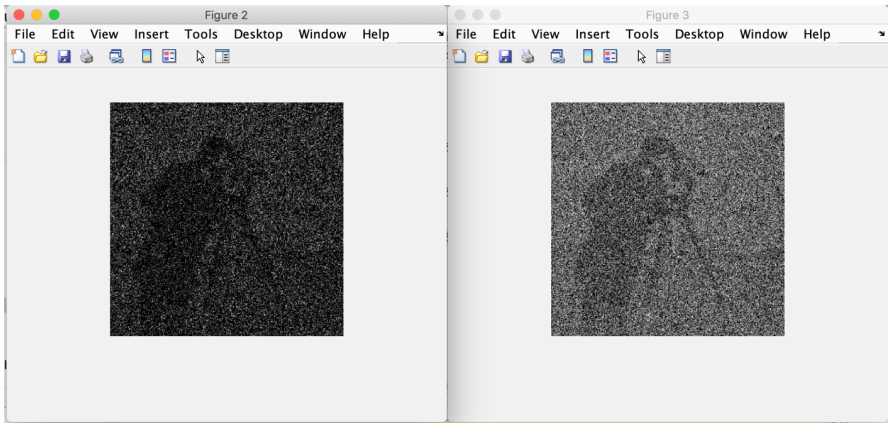

Figure 23: Recovered images setting $n = 64$ and $m = 12$.

		Least squares		Basis pursuit	
n	**m**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
$16,384$	$1,600$	6.8733	0.0430	11.4033	0.0540
$4,096$	200	6.4129	0.0498	11.2331	0.0478
$4,096$	400	6.8945	0.0407	11.3329	0.0513
$1,024$	180	7.5423	0.0433	11.5789	0.0640
$1,024$	280	8.0844	0.0510	12.0984	0.0806
$1,024$	400	9.9639	0.0843	13.3337	0.1274
64	6	6.8587	0.0449	8.9910	0.0267
64	12	7.7427	0.0473	10.7878	0.0575

Table 4: Recovered images quality in terms of PSNR and SSIM using pseudo-random scanning.

Table 4 shows the results in terms of PSNR and SSIM when random scanning is applied.

Compared with previous scanning strategies, it can be observed that using random scanning provides worse results, this could happen because the relationship between neighboring pixels is lost in the pseudo-random scanning process.

5.4 Conclusions

Preliminary results strongly suggest creating input vectors to the compressed sensing algorithm by using an image scanning strategy that takes advantage of the existing relationship among neighboring pixels such as non-overlapping blocks scanning or Hilbert curve scanning, since these strategies provide better PSNR and SSIM values.

Regarding to input vectors size (n), it can be observed that using smaller vectors provides better image approximations. It happens because there is greater pixel redundancy into a small set of pixels and as more pixels in the set are considered, the redundancy among them decreases.

Experiments show that using compressed sensing algorithm to create a highly incomplete representation of the image can be used to approximate it with optimal quality. Instead of using block pixels average as recovery information, compressed sensing algorithm can be use to create a compact representation of the image, which can be included as part of a watermark to use it as recovery information.

It has been identified in the literature that recent studies related to compressed sensing theory incorporate deep learning models to improve signal reconstruction performance (Adler et al., 2016; Mousavi and Baraniuk, 2017; Zou and Yang, 2019). Therefore, deep learning approaches will be also studied to be considered as a possible solution to reconstruction problem.

References

Adler, A., Boublil, D., Elad, M., and Zibulevsky, M. (2016). A deep learning approach to block-based compressed sensing of images. *arXiv preprint arXiv:1606.01519*.

Barton, J. M. (1997). Method and apparatus for embedding authentication information within digital data. US Patent 5,646,997.

Candes, E. and Romberg, J. (2005). l1-magic: Recovery of sparse signals via convex programming. *URL: www. acm. caltech. edu/l1magic/downloads/l1magic. pdf*, 4:14.

Candes, E. J., Romberg, J. K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223.

Chiang, K.-H., Chang-Chien, K.-C., Chang, R.-F., and Yen, H.-Y. (2008). Tamper detection and restoring system for medical images using wavelet-based reversible data embedding. *Journal of Digital Imaging*, 21(1):77–90.

Cox, I., Miller, M., Bloom, J., Fridrich, J., and Kalker, T. (2008). *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edition.

Cox, I. J., Miller, M. L., Bloom, J. A., and Honsinger, C. (2002). *Digital watermarking*, volume 53. Springer.

Deng, X., Chen, Z., Zeng, F., Zhang, Y., and Mao, Y. (2013). Authentication and recovery of medical diagnostic image using dual reversible digital watermarking. *Journal of nanoscience and nanotechnology*, 13(3):2099–2107.

Donoho, D. L. et al. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Ekanadham, C., Tranchina, D., and Simoncelli, E. P. (2011). Recovery of sparse translation-invariant signals with continuous basis pursuit. *IEEE transactions on signal processing*, 59(10):4735–4744.

Eskicioglu, A. M. and Fisher, P. S. (1995). Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12):2959–2965.

Eswaraiah, R. and Reddy, E. S. (2014). Medical image watermarking technique for accurate tamper detection in roi and exact recovery of roi. *International journal of telemedicine and applications*, 2014:13.

Gao, G., Cui, Z., and Zhou, C. (2018). Blind reversible authentication based on pee and cs reconstruction. *IEEE Signal Processing Letters*, 25(7):1099–1103.

Haghighi, B. B., Taherinia, A. H., and Harati, A. (2018). Trlh: Fragile and blind dual watermarking for image tamper detection and self-recovery based on lifting wavelet transform and halftoning technique. *Journal of visual communication and image representation*, 50:49–64.

Hore, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369.

Hyvärinen, A., Hurri, J., and Hoyer, P. O. (2009). *Natural image statistics: A probabilistic approach to early computational vision.*, volume 39. Springer Science & Business Media.

Khan, A., Siddiqa, A., Munib, S., and Malik, S. A. (2014). A recent survey of reversible watermarking techniques. *Information Sciences*, 279:251–272.

Lawder, J. K. (2000). Calculation of mappings between one and n-dimensional values using the hilbert space-filling curve. *School of Computer Science and Information Systems, Birkbeck College, University of London, London Research Report BBKCS-00-01 August*.

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, 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.

Lo, C.-C. and Hu, Y.-C. (2014). A novel reversible image authentication scheme for digital images. *Signal processing*, 98:174–185.

Lou, D.-C. and Liu, J.-L. (2000). Fault resilient and compression tolerant digital signature for image authentication. *IEEE Transactions on Consumer Electronics*, 46(1):31–39.

Mallat, S. G. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415.

Mousavi, A. and Baraniuk, R. G. (2017). Learning to invert: Signal recovery via deep convolutional networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2272–2276. IEEE.

Nguyen, T.-S., Chang, C.-C., and Yang, X.-Q. (2016). A reversible image authentication scheme based on fragile watermarking in discrete wavelet transform domain. *AEU-International Journal of Electronics and Communications*, 70(8):1055–1061.

Nyquist, H. (1928). Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644.

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.

Saini, L. K. and Shrivastava, V. (2014). A survey of digital watermarking techniques and its applications. *arXiv preprint arXiv:1407.4735*.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Shi, Y.-Q., Li, X., Zhang, X., Wu, H.-T., and Ma, B. (2016). Reversible data hiding: advances in the past two decades. *IEEE Access*, 4:3210–3237.

Tropp, J., Gilbert, A. C., et al. (2007). Signal recovery from partial information via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 53(12):4655–4666.

Velumani, R. and Seenivasagam, V. (2010). A reversible blind medical image watermarking scheme for patient identification, improved telediagnosis and tamper detection with a facial image watermark. In *2010 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–8. IEEE.

Wang, W., Dong, J., and Tan, T. (2010). Image tampering detection based on stationary distribution of markov chain. In *2010 IEEE International Conference on Image Processing*, pages 2101–2104. IEEE.

Warif, N. B. A., Wahab, A. W. A., Idris, M. Y. I., Ramli, R., Salleh, R., Shamshirband, S., and Choo, K.-K. R. (2016). Copy-move forgery detection: survey, challenges and future directions. *Journal of Network and Computer Applications*, 75:259–278.

Xie, L., Arce, G. R., and Graveman, R. F. (2001). Approximate image message authentication codes. *IEEE Transactions on Multimedia*, 3(2):242–252.

Yin, Z., Niu, X., Zhou, Z., Tang, J., and Luo, B. (2016). Improved reversible image authentication scheme. *Cognitive Computation*, 8(5):890–899.

Zheng, L., Zhang, Y., and Thing, V. L. (2019). A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation*, 58:380–399.

Zou, C. and Yang, F. (2019). Deep learning approach based on tensor-train for sparse signal recovery. *IEEE Access*, 7:34753–34761.