Symmetric Digital Watermarking

Nguyen Quang Vinh, Chan Weizhong,
Ng Jong Ray Edward, and Tham Jin Lin

National University of Singapore

Abstract. The rapid growth of the Internet in general has propelled the development of digital media. As websites hosting digital media got popular, there exists a need to protect digital media creators' content and claim their rightful ownership of such content. Without watermarking, valuable digital assets can be susceptible to content theft or unauthorized use. In this paper, we deal with one aspect of securing digital media - securing digital images with the use of watermarking. The paper will first give the definition a digital watermarking scheme, then the definiton of properties along with associated games, of a secure digital watermarking scheme. Finally, we will give our digital watermarking construction and prove some properties of that construction.

Keywords: Digital Watermarking · Digital Media

1 Introduction

Writers, photographers, musicians, and artists are among those who have taken advantage of the worldwide publishing opportunities provided by the Internet, yet these same people - all considered "authors" under international regulations - are frequently being taken advantage of by online pirates. Given the ease with which audio and visual files can be duplicated, it is no surprise that such duplication on the Internet regularly occurs without the rightful owners' permission. This has cost substantial financial damage to the owners of digital content. The situation has gone worse with the rising popularity of piracy sites - most notable is `ThePirateBay`, which are mainly used for quick and convenient sharing of pirated content. This is where digital watermarks come in to mitigate the damage that the online digital pirates may cause.

A digital watermark is a kind of marker covertly embedded in a noise-tolerant signal such as audio, video or image data. It is typically used to identify ownership of the copyright of such signal. "Watermarking" is the process of hiding digital information in a carrier signal; the hidden information should, but does not need to, contain a relation to the carrier signal. Digital watermarks may be used to verify the authenticity or integrity of the carrier signal or to show the identity of its owners. It is prominently used for tracing copyright infringements and for banknote authentication.

Like traditional physical watermarks, digital watermarks are often only perceptible under certain conditions, like after using some algorithm. If a digital

watermark distorts the carrier signal in a way that it becomes easily perceivable, it may be considered less effective depending on its purpose. Traditional watermarks may be applied to visible media (like images or video), whereas in digital watermarking, the signal may be audio, pictures, video, texts or 3D models. A signal may carry several different watermarks at the same time. Unlike metadata that is added to the carrier signal, a digital watermark does not change the size of the carrier signal.

Digital watermarking generally operates on different digital media or cover objects (e.g., image, audio, video) and is considered to have three major components watermark generation, embedding, and detection. Watermark generation yields the desired watermark, which can optionally depend on some keys. The generated watermark is embedded into the cover object by the watermark embedding, sometimes based on an embedding key. During detection, the embedded watermark in a cover object is extracted and verified.

As stated above, the digital media space is vast - image, audio and video content is included. The nature of different types of digital media also greatly varies, therefore coming up with a scheme that applies to all types of content in the digital media space is extremely difficult. Our group, hence, has decided to provide a digital watermarking model that strictly follows the systematic structure seen in CS4236. The paper is organized as follows: Section 2 provides the formal definition of a digital watermark scheme. Section 3 presents the security properties related to the scheme. In section 4, a construction of the image watermarking scheme is given. The conclusions are given in section 5.

2 Digital Watermarking scheme for images

Before giving the formal definition of the scheme, we first need to give the notion of a image that is used throughout the paper. A 2D image of size n bits is defined as a n-bit string. We prefer this definition as this simplifies the definition of a "flattened" image as this makes defining properties and construction easier, without having to get complex mathematics involved. Also, a 2D image of size $n = a \times b$ can be reconstructed trivially from a flattened string by taking a chunks of data sized b bits.

Definition 1. *Symmetric Digital Watermarking scheme*

Symmetric Digital Watermarking scheme contains 5 probabilistic polynomial-time algorithms $(\texttt{KeyGen}, \texttt{WtmkGen}, \texttt{Emb}, \texttt{Ext}, \texttt{Vrfy})$ such that:

1. $\texttt{KeyGen}(\mathbf{1^n})$: the key-generation algorithm takes as input the parameter 1^n and outputs 2 keys g, e $(|g|, |e| \geq n)$ as the watermark generation key and embedding key respectively.
2. $\texttt{WtmkGen}_{\boldsymbol{g}}(\boldsymbol{i}, \boldsymbol{m})$: the watermark-generation algorithm takes as input the original image $i \in \{0,1\}^I$, a unique message that we want to embed into the image $m \in \{0,1\}^M$, the watermark generation key g. The output is a watermark $w \in \{0,1\}^W$; $w := \texttt{WtmkGen}_g(i, m)$

3. $\texttt{Emb}_{\boldsymbol{e}}(\boldsymbol{i}, \boldsymbol{w})$: the embedding algorithm which takes as input the original image $i \in \{0,1\}^I$, the embedding key e, the watermark w, and outputs the watermarked image $\bar{i} \in \{0,1\}^I$; $\bar{i} := \texttt{Emb}_e(i, w)$
4. $\texttt{Ext}_{\boldsymbol{g},\boldsymbol{e}}(\bar{\boldsymbol{i}})$: the extraction algorithm takes as input a watermarked image $\bar{i} \in \{0,1\}^I$ and the generation and embedding key g, e, and outputs the estimated message $\widetilde{m} \in \{0,1\}^M$ and estimated watermark $\widetilde{w} \in \{0,1\}^W$. $(\widetilde{m}, \widetilde{w}) := \texttt{Ext}_{g,e}(\bar{i})$
5. $\texttt{Vrfy}_{\boldsymbol{g},\boldsymbol{e}}(\boldsymbol{i}, \bar{\boldsymbol{i}}, \boldsymbol{m})$: the verification algorithm $\texttt{Vrfy}$ takes as input the original and watermarked image $i, \bar{i} \in \{0,1\}^I$, the embedded message $m \in \{0,1\}^M$, embedding key e and watermark generation key g. $\texttt{Vrfy}$ outputs 1 iff $m = m'$ and $w = \widetilde{w}$

Under no tampering from some adversary, or bit flip errors during transmission of the watermarked image, for any security parameter n, keys g, e output by $\texttt{KeyGen}(1^n)$, image i and message m, this equation must hold:

$$\texttt{Ext}_{g,e}(\texttt{Emb}_e(i, \texttt{WtmkGen}_g(i, m))) = (m, w) \tag{1}$$

3 Properties

3.1 Imperceptability

An underlying property which enables much of the security in watermarks is the concept of imperceptibility. Taking into consideration that an image consists of relatively redundant information, we define imperceptibility by comparing 2 images and determining if they are "sufficiently similar" to each other. This concept comes from the idea that distortion of the original image is inevitable, especially when trying to embed watermarking information into the image. Being able to justify if 2 images are imperceptible with respect to each other is crucial in enabling comparison between images and their data.

To quantify the idea of "sufficiently similar", we introduce two concepts: similarity measures - what a party can use to verify the similarity of images, and threshold - the tolerable bound for two images to be perceptually similar using the given metric.

Definition 2. *Imperceptibility*

Any two images $i_1, i_2 \in \{0,1\}^*$ are said to be (F, T) imperceptible if for all similarity measures $f_j \in F \equiv \{f_1, f_2, ..., f_n\}$ and corresponding threshold $t_j \in T = \{t_1, t_2, ..., t_n\}$, the following holds:

$$f_j(i_1, i_2) \leq t_j \tag{2}$$

The definition stems from the fact that various measures are used in practice to quantify the changes made to a image. Some notable examples are peak or weighted signal to noise ratio, mean square error, structural similarity index. From our own research, there does not exist any universally accepted standard of measuring perceptible changes to a image.

3.2 Robustness

Building on the definition of Imperceptibility, we are able to now introduce the concept of a robust watermark. Informally, a robust watermark is an embedded watermark that is able to withstand image processing techniques such that the watermark can still be extracted from the processed image and this processed image is imperceptible to its original image (similar to Definition 2). This allows us to analyse the impact of image processing on a watermarked image.

An adversary can perform a set of processing techniques, using various operations/transforms to distort or completely destroy the message embedded in the digital watermark of the image. Thus, we need to formally define this ability of the adversary.

Definition 3. *Processed image*

Define P is the set of all applicable processing techniques for an image $i \in \{0,1\}^*$. A processed image is an image that is not essentially perceptually similar to its original, but a certain amount of distortion, δ is incurred by a processing technique $p \in P$.

That is, if an image $i \in \{0,1\}^*$ is processed by p then, for the processed image $i' \in \{0,1\}^*$, the following equation must hold:

$$i' = p(i) = i + \delta \tag{3}$$

Design rationale: There are many different processing techniques, such as compression, de-noising. Each technique has different parameters, like compression ratio, down sampling rate, type and rank of filter. These parameter settings give different strengths (or invasiveness) to a processing technique. Therefore, a technique p in our definition means that p is defined with its all essential parameter settings. This gives an adversary more freedom in choosing the appropriate parameters to successfully pull off an attack.

With definition 3, we can now define robustness of a digital watermarking scheme. Informally, an adversary would aim to find a a processing technique p that will apply a distortion on the watermarked image $\bar{i}$ such that $p(\bar{i}) \simeq i$ (in words, the change to the processed image is imperceptible to the original image i) and $\texttt{Ext}_{g,e}(\bar{i}) \neq m$.

The first condition, $p(i) \simeq i$ is crucial to restrict the adversary into create a meaningful image. Otherwise the adversary can trivially create a processing technique that will distort the entire image to remove the watermark and wins the game by destroying the original image.

The second condition is to describe a case where the adversary is able to apply a processing technique to change the watermark into another valid watermark but with a different message. An example can be an attack to change the author of the watermarked image.

When designing the robustness game, it is natural to create a naive game where the adversary chooses an image-message pair i,m and pass that into a

watermarking oracle to get back the watermarked image $\bar{i}$. However with this information, the adversary, knowing i, is trivially able to find a processing technique p that can distort the watermarked image to look similar to the original image. The adversary thus can win by meeting the 2 conditions $p(\bar{i}) \simeq i$, and $\texttt{Ext}_{g,e}(\bar{i}) \neq m$ as the original image i will not have any watermark to extract. Hence, a proper robustness game has to randomly generate the image-message pair i, m and only give the adversary the watermarked image $\bar{i}$.

***Watermark − Robust*$_{\mathcal{A},\Pi}$**

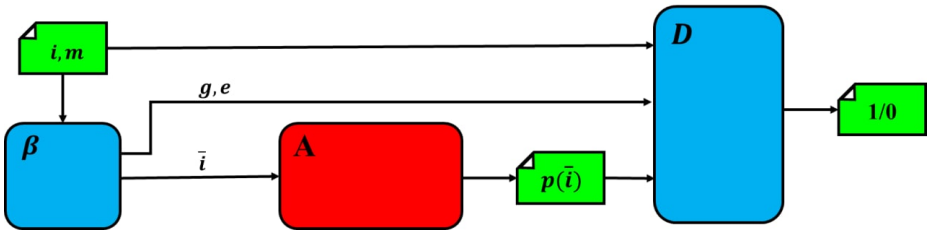

Fig. 1. High-level view of the $Watermark - Robust_{\mathcal{A},\Pi}$ game

Definition 4. *Robustness game*

The Robustness game $Watermark - Robust_{\mathcal{A},\Pi}$ has the following procedures:

1. The challenger randomly choose an image $i \in \{0,1\}^I$ and a message $m \in \{0,1\}^M$ and pass it into the watermarking engine β.
2. The watermarking engine β outputs a watermarked image $\bar{i} \in \{0,1\}^I$ and the watermark generation key g, and embedding key e. The watermarked image $\bar{i}$ is sent to the adversary $\mathcal{A}$.
3. The adversary $\mathcal{A}$ outputs a distorted (processed) watermarked image $p(\bar{i})$ and sends it to detector $\mathcal{D}$.
4. The adversary $\mathcal{A}$ wins if and only if $D(i, m, g, e, p(\bar{i})) = 1$.

Definition 5. *Watermarking engine* β

The watermarking engine β executes these procedures:

1. Use `KeyGen` as a subroutine, generate watermark generation key g and embedding key e.
2. Upon input of image $i \in \{0,1\}^I$, and $m \in \{0,1\}^M$, generate a watermark $w := \texttt{WtmkGen}_g(i, m)$.
3. With the watermark w, generate a watermarked image $\bar{i} := \texttt{Emb}_e(i, w)$.
4. Outputs watermark generation key g, embedding key e and watermarked image $\bar{i}$.

Watermark engine β

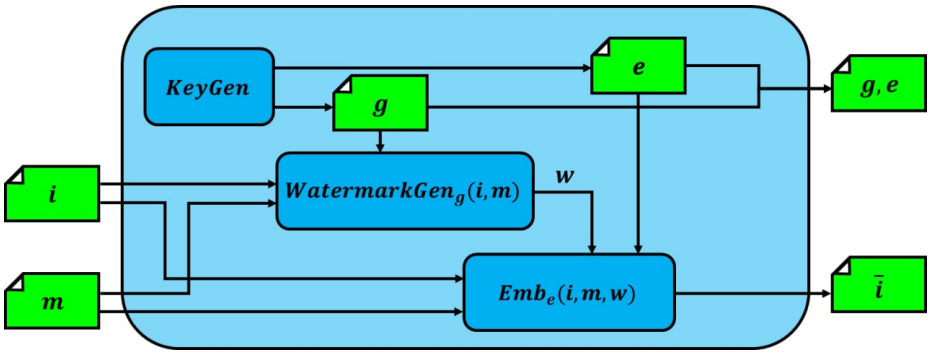

Fig. 2. Schematics of the watermark engine β

Definition 6. *Detector D*

The detector D executes these procedures:

1. Receive the image $i \in \{0,1\}^I$, the message $m \in \{0,1\}^M$, watermark generation key g, embedding key e, and adversary processed image $p(\bar{i})$ as input.
2. Use its own set of similarity measures F and corresponding threshold T, to check whether $p(\bar{i}) \simeq i$, or in words, whether the changes done on the watermarked image is imperceptible to the original image i.
3. Call the subroutine $Ext_{g,e}(p(\bar{i}))$ to obtain the estimated message m'. Afterwards, check whether $m \neq m'$.
4. If both conditions are True (using the `AND` operator), output 1. Otherwise, output 0.

Detector *D*

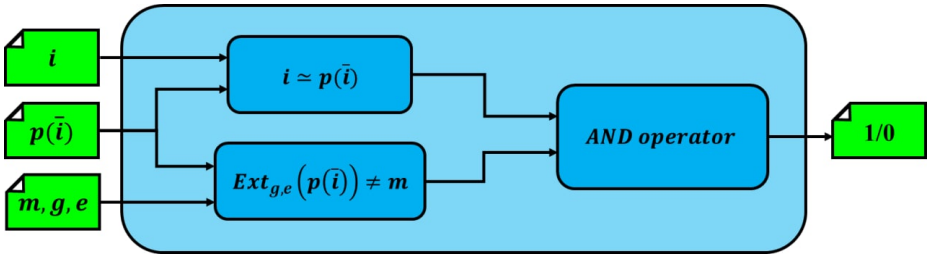

Fig. 3. Schematics of the detector D

Based on the *Robustness* game, we can define the *Robustness* property.

Definition 7. *Robustness*

A digital watermark Π = (`KeyGen, WtmkGen, Emb, Ext, Vrfy`) is robust if for every processing technique p in the set of all applicable processing techniques P that the PPT adversary $\mathcal{A}$ has access to, there is a negligible function $negl$, such that:

$$Pr[Watermark - Robust_{\mathcal{A},\Pi}(n) = 1] \leq negl(n) \tag{4}$$

A digital watermark is robust if it is robust under all image processing techniques available to the adversary $\mathcal{A}$. This definition is based on our literature review on digital watermarking. There does not exist any absolute robustness for watermarking, since taking all known/available processing techniques into consideration (for robustness) is not realistic to the adversary $\mathcal{A}$.

3.3 Unforgeability

We shall define the notion of unforgeability. Informally, an adversary should not be able to create a valid watermarked image for a "new" image-message pair. In a forgery attempt, an adversary tries to create a legitimate watermarked image through unauthorized watermark embedding. Given oracle access to the watermark generation and embedding functions, the adversary chooses an image i and a message m and outputs a new watermarked image $\bar{i}$. The adversary wins if he can successfully forge a new watermarked image with any chosen image and embedded message.

Definition 8. *Watermark-forgery game*

Let O be a watermark encoding oracle that has 2 subroutines `WtmkGen` and `Emb`. The oracle O takes in a pair of image and message (i, m) as input and outputs a valid watermarked tuple $(i, m, \bar{i})$.

***Watermark − Forge*$_{\mathcal{A},\Pi}$**

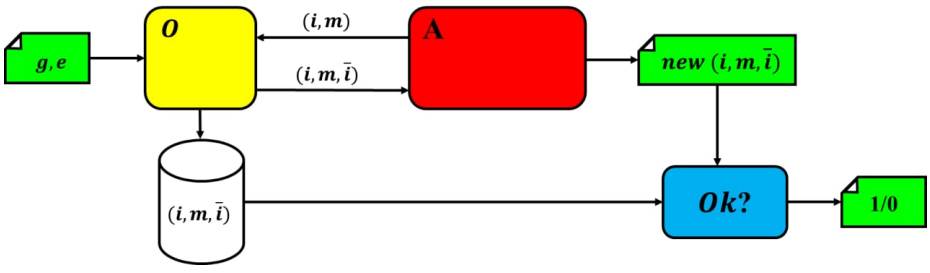

Fig. 4. High-level view of the $Watermark - Forge_{\mathcal{A},\Pi}$ game

The Watermark-forgery game $Watermark-Forge_{\mathcal{A},\Pi}$ has the following procedures:

1. Challenger uses `Keygen` to generate watermark generation key g and embedding key e.
2. Adversary $\mathcal{A}$ has oracle acess. Adversary outputs the tuple $(i, m, \bar{i})$. Let $\mathcal{Q}$ be all the image-message pairs the adversary has sent to the oracle in this game.
3. Adversary wins (output of the game is 1) iff $\texttt{Vrfy}_{g,e} = 1$, and $(i, m, \bar{i}) \notin Q$. In words, the newly watermarked image produced by the adversary $\bar{i}$ correspond to the new image-message pair (i, m).

Definition 9. *Watermark Unforgeability*

A digital watermark Π = `(KeyGen, WtmkGen, Emb, Ext, Vrfy)` is unforgeable iff for any PPT adversary A, there is a negligible function $negl$, such that:

$$Pr[Watermark - Forge_{\mathcal{A},\Pi} = 1] \leq negl(n) \tag{5}$$

This property is crucial, as this implies that only the only entity who can legitimately embed some message into the image is the entity who has knowledge of the generator and embedding key.

4 Construction

With the definition of the scheme and the associating security properties, we will now give a simple, unforgable construction of a digital watermarking scheme.

4.1 Relevant definitions

Before introducing the formal definition of the construction, we will introduce the crucial design decisions and defintions related to the construction. Let $\mathcal{H}$ be a collision-resistant, fixed length, unkeyed hash function, and MAC be a fixed-length, strong MAC.

Denote the length of the output of $\mathcal{H}$ as H, the length of the tag produced by MAC as T, the length of the input image i as I, the length of the embedding message m as M. We want to design a scheme that is imperceptible to the `Bit-Changes` similarity measurement.

Definition 10. *`Bit-Change` similarity measure*

The measure, denoted by $f(i_1, i_2)$, takes in two images $i_1, i_2 \in \{0,1\}^I$, and outputs the difference in bits of the two images. It does so by comparing every j^{th} $(1 \leq j \leq n)$ bit of i_1 and i_2, and stores the number of bit difference between the two images.

With this similarity measure, we will need to specify the threshold function.

Definition 11. *`Bit-Change` threshold function*

Given the size of the image I and parameter α, the threshold function $t(I, \alpha)$ is defined as:

$$t(I, \alpha) = \alpha I \tag{6}$$

Hence, two images $i_1, i_2 \in \{0,1\}^I$ is `Bit-Change` imperceptible iff the following holds:

$$f(i_1, i_2) \leq t(I, \alpha) = \alpha I \tag{7}$$

Our idea for the construction is to replace a portion of the original image with the watermark (embedded with the message), hence the length of the watermark embedded into the image must satisfy:

$$|w| \leq \alpha I \tag{8}$$

The construction leaves α as a parameter as we want the party in charge of watermarking to have more control with how covert they wish to have for their watermark. The smaller the α, the more convert (harder to detect using the `Bit-Change` similarity measure) the watermark gets.

For ease of extracting the message, we want to append the message with the keyed hash of the image concatenated with the message itself. The mathematical form of the watermark is $w = m||MAC_g(\mathcal{H}(i||m))$. Hence, to satisfy the $Bit-Change$ imperceptibility property, the following must hold:

$$M + T \leq \alpha I \tag{9}$$

The embedding process will replace $|w|$ bits in the original image with the watermark.

4.2 Construction of the Symmetric Digital Watermarking scheme

Construction: *Simple Digital Image Watermarking*

Choose some collision-resistant, fixed-length unkeyed hash $\mathcal{H}$, a strong, fixed length MAC that takes the input of length H, and message to the image i of size I and parameter α such that $M + T = \alpha I$.

Construct *Simple Digital Image Watermarking* as follows:

1. $\texttt{KeyGen}(1^m)$: on input 1^n, output a unformly chosen watermark generation key $g \in \{0,1\}^n$ and uniformly chosen embedding key $e \in \{0, I - \alpha I\}$.
2. $\texttt{WtmkGen}_g(i, m)$: on input of the watermark generation key g, image $i \in 0, 1^I$ and message $m \in 0, 1^M$, outputs $w := m||MAC_g(\mathcal{H}(i||m))$. The length of the watermark $|w| = \alpha I$
3. $\texttt{Emb}_k(i, w)$: on input of the embedding key e, the watermark $w \in 0, 1^{\alpha I}$ and the original image i, outputs the watermarked image $\bar{i}$. $\bar{i}$ is generated by first duplicate the original image i, that is, for $j \in \{1, n\}$, set $\bar{i}_j = i_j$. Afterwards, for $k \in \{0, \alpha I\}$, set $\bar{i}_{e+k} = w_k$.

4. $\texttt{Ext}_{e,g}(\bar{i})$: on input of generation and embedding keys (g, e), and the watermarked image $\bar{i}$, output the estimated watermark $\widetilde{w} := \bar{i}_e, \bar{i}_{e+1}, ..., \bar{i}_{e+\alpha I}$. The estimated message is $\widetilde{m} := \widetilde{w}_0, \widetilde{w}_1, ...\widetilde{w}_{\alpha I - K}$.
5. $\texttt{Vrfy}_{e,g}(i, \bar{i}, m)$: on input of generation and embedding keys (g, e) and the tuple $(i, \bar{i}, m)$, computes $(\widetilde{w}, \widetilde{m}) = Ext_{g,e}(\bar{i})$ and the original watermark $w = WtmkGen_g(i, m)$. Output 1 iff $\widetilde{m} = m$ and $\widetilde{w} = w$. Otherwise, output 0.

This construction clearly satisfies the inequality given in (9), hence it has the `Bit-Change` imperceptibility property.

Also, we will show that this construction has the Unforgeability property.

Proof. In the digital watermarking scheme, the security of the construction rests on 2 underlying security. 1: the watermark must be generated properly, 2: the watermark must be embedded correctly.

For the 2^{nd} condition, the above construction requires the adversary to know the starting bit for embedding. If the watermark is embedded at the wrong starting location, the extracted hash will appear random. This means that the probability for the adversary to guess the correct starting bit is $\frac{1}{I-|w|}$.

Suppose that the adversary $\mathcal{A}$ is able to correctly forge a watermarked image. This means that $\mathcal{A}$ is able to create a new valid watermarked image $\bar{i}$ for a new image-message pair (i', m').

We will grant the adversary more power. Indeed, suppose that the adversary, using some probabilistic polynomial time algorithm, can correctly predict the embedding location of the watermark inside the image. Now, in order to win the game, the adversary $\mathcal{A}$ only needs to create a valid watermark to perform the embedding process on the image and win the game.

Denote t as the tag concatenated to the original message, the above implies that the adversary $\mathcal{A}$ is able to generate a valid watermark $w' = m'||t$ for a newly chosen image i' from the adversary $\mathcal{A}$. This implies that the adversary has the ability to forge a MAC tag $t = MAC_g(\mathcal{H}(i||m))$. Again, note that the formulation of the tag that the construction employs is one instance of the hash-and-MAC paradigm. From Theorem 5.6 given in the textbook, since MAC is a secure MAC for fixed-length messages and $\mathcal{H}$ is collision resistant (both are required in the image watermarking construction above), the hash-and-MAC portion of the watermark, or the tag t following the message must be secure. Theorem 5.6 also implies that the adversary only have negligible probability of generating a valid MAC in the hash-and-MAC construction.

Hence, any PPT adversary $\mathcal{A}$, even with an polynomial time algorithm to correctly predict the location of the watermark in the image, the adversary still can not forge a valid watermark in polynomial time. Or, formally, the probability that the adversary wins the $Watermark - Forge$ game is:

$$Pr[Watermark - Forge_{\mathcal{A},\Pi} = 1] = negl(n) \tag{10}$$

This satisfies equation (5), hence by definition, the construction has the Unforgeability property.

5 Conclusions

We have presented the need for why digital watermarking is essential in the current context. Afterwards, a formal Symmetric Digital Watermarking scheme is introduced. Due to the high application variant properties of watermarking, we have focused on the image applications. We have also defined properties and games, which highlight the security-related characteristics that a digital watermarking scheme should have. Finally, we have provided a simple construction that satisfies the Imperceptability, as well as Unforgeability property.

References

1. Jonathan, K., Yehuda, L.: Introduction to Modern Cryptography. 2nd edn. CRC Press, New York (2015)
2. Hussain N., Wageeh B., Colin, B.: Digital image watermarking: its formal model, fundamental properties and possible attacks. EURASIP Journal on Advances in Signal Processing (2014)
3. Stefan, K., Fabien P.: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Britain (2000)
4. Ingemar C.: Digital watermarking and steganography. Morgan Kaufmann, USA (2008)
5. Frank S.: Digital watermarking and steganography: fundamentals and techniques. Taylor & Francis, USA (2008)
6. Weiqi Y., Jonathan W.: Fundamentals of Media Security. Ventus Publishing (2010)