



Watermarking-based image authentication with recovery capability using halftoning technique

Luis Rosales-Roldan^a, Manuel Cedillo-Hernandez^b, Mariko Nakano-Miyatake^{a,*}, Hector Perez-Meana^a, Brian Kurkoski^c

^a Postgraduate Section, Mechanical Electrical Engineering School, National Polytechnic Institute of Mexico, , Mexico

^b Electrical Engineering Division, Engineering Faculty, National Autonomous University of Mexico, Mexico

^c Japan Advanced Institute of Science and Technology, Japan

ARTICLE INFO

Article history:

Received 13 November 2011

Accepted 15 November 2012

Available online 23 November 2012

Keywords:

Watermarking

Content authentication

Recovery capability

Halftoning

IWT

DCT

SSIM criterion

ABSTRACT

In this paper two watermarking algorithms for image content authentication with localization and recovery capability of the tampered regions are proposed. In both algorithms, a halftone version of the original gray-scale image is used as an approximated version of the host image (image digest) which is then embedded as a watermark sequence into given transform domains of the host image. In the first algorithm, the Integer Wavelet Transform (IWT) is used for watermark embedding which is denominated WIA-IWT (Watermarking-based Image Authentication using IWT), while in the second one, the Discrete Cosine Transform (DCT) domain is used for this purpose, we call this algorithm WIA-DCT (Watermarking-based Image Authentication using DCT). In the authentication stage the tampered regions are detected using the Structural Similarity index (SSIM) criterion, which are then recovered using the extracted halftone image. In the recovery stage, a Multilayer Perceptron (MLP) neural network is used to carry out an inverse halftoning process to improve the recovered image quality. The experimental results demonstrate the robustness of both algorithms against content preserved modifications, such as JPEG compression, as well as an effective authentication and recovery capability. Also the proposed algorithms are compared with some previously proposed content authentication algorithms with recovery capability to show the better performance of the proposed algorithms.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Nowadays the developments based on digital technology have a strong impact on the people's life, for example it is quite common to take pictures everywhere and every time using his/her cellular phones with digital cameras, giving as a results that about 700,000 pictures per hour are uploaded to any computer network to be shared among different users. This kind of images can be easily

manipulated, realizing for example copy-and-paste operations, in which some part of the image is replaced by a part of another or the same image. These manipulations can be done using simple tools available in any PC, such as Photoshop and Corel Draw, etc., without almost any perceptual distortions. However, sometimes these alterations cause economical and/or moral damages to involved persons. This is one of the main reasons why digital image authentication has become one of the most important issues in the information security fields.

Cryptographic approaches, such as cryptographic hashing [1] and fragile watermarking [2–10] are effective digital material authentication methods. Almost all fragile

* Corresponding author.

E-mail address: mariko@infinitum.com.mx (M. Nakano-Miyatake).

watermarking methods embed the image digest into some LSB planes to recover the tampered regions [3–10]. In [9], two copies of digest image and permutation with push-aside operation are used to decrease loss of digest image caused by tampering/missing coincidence, obtaining high recovery capability. Zhang et al. employed a compressive sensing and compositive reconstruction techniques to recover the tampered region with high tampering percentage [10]. In both algorithms the digest image is embedded into three LSB planes of the original image, which is vulnerable against common image processing operation, including content preserving operations. Therefore generally fragile watermarking schemes are not adequate for practical image authentication, because they are bit-sensitive methods which mean that the method asserts its authenticity only if the integrity of whole bits of the digital contents is confirmed. In digital images, however, there are many file formats with/without compression options, and it is quite common to save compressed version of the image in the different file format. Then the goal of image authentication is to accept content preserving modification and reject the malicious manipulations, which is classified as content authentication scheme.

As effective content authentication for digital images, principally two approaches are proposed in literature, which are: image hashing or perceptual hashing [11–13] and semi-fragile watermarking [14–27]. In the image hashing-based authentication, perceptually robust bits sequences are extracted from the image and the extracted hash sequence is transmitted or saved together with the image for authentication purpose. While in the semi-fragile watermarking-based image authentication, the extracted bit sequence is embedded into the image as watermark sequence. Then the authenticity of the image is evaluated extracting the watermark sequence from the image under analysis. Both approaches present each one some advantage and disadvantage over their counterpart. The principal disadvantage of the watermarking-based technique is the distortion of the image quality caused by watermark embedding, on the other hand in the image hashing-based techniques, an additional file or data must be transmitted or hold in a safety manner.

Almost all content image authentication methods have localization capability of the tampered regions [14–27] and some of them can even recover these regions [20–27]. Considering that the recovery capability of the tampered regions is very useful for many applications, such as assessment insurance in automobile accidents and acquisition of important information from images, reliable recovery schemes with better recovered image quality are required. To recover the tampered regions, digest of the image must be extracted or generated from the same image and used as watermark sequence. In the embedding stage of [20], the original image is decomposed by the first level decomposition of Discrete Wavelet Transform (DWT), and the lowest frequency DWT sub-band is furthermore transformed by full-frame Discrete Cosine Transform (DCT). The first m lowest DCT coefficients are used as the watermark sequence, which is redundantly embedded into the middle frequency DWT sub-bands

of the original image. In the authentication and recovery stage, the extracted watermark is transformed by the inverse DCT and compared with the lowest frequency DWT sub-band of the image using the mean absolute error (MAE) criterion. If a tamper is detected, the image generated from the extracted watermark becomes the recovered image. In the watermarking scheme proposed by Ref. [21], two watermark sequences: authentication watermark and recovery watermark are embedded for authentication and recovery purposes, respectively. Both watermarks are embedded into different sub-bands in the Integer Wavelet Transform (IWT) domain. The authentication watermark is binary pseudo-random sequence, which is embedded using a quantization-based method and the recovery watermark is embedded in similar manner as in Ref. [20]. In the authentication stage, the authentication watermark is extracted to detect the tampered region, which is then recovered using the extracted recovery watermark. The algorithm proposed in Ref. [22] is an improved version of the one in Ref. [21], in the recovery watermark embedding stage, the Integer DCT, the Huffman coding and the BCH error control coding are introduced to reduce the watermark length and increase robustness. The principal disadvantage of this algorithm for an efficient reduction of watermark length is that the lookup table for Huffman coding must be generated and saved for each image. Authors of Ref. [23] proposed a semi-fragile watermarking method based on QIM dither modulation in the IWT domain, in which authentication watermark and recovery watermark are used. The authentication watermark is a predefined binary pattern and the recovery watermark is halftone version of the lowest sub-band generated by second level of IWT decomposition. This algorithm shows watermark robustness against several content preserving modifications using an adequate threshold value; however this threshold value causes high false negative error rates. A low quality of the recovered image is another disadvantage of this scheme because the extracted recovery watermark must be scaled four times.

In the image authentication scheme proposed in Ref. [24], firstly the original image is segmented, manually, into the ROI (Region of Interest) blocks and ROE (Region of Embedding) blocks. Then the watermark sequence is generated from the DCT coefficients of the ROI blocks and embedded into the DCT domain of the ROE blocks. The principal disadvantage of this scheme is that the number of ROI blocks is limited to about the 30% of the whole image, and then the whole image cannot be protected. In Ref. [25], a hybrid block-based watermarking technique composed by a robust watermarking scheme for self-correction and a fragile watermarking scheme for sensitive authentication is proposed. In this algorithm all types of alterations, including the content preserving modification, are detected as tamper and the recovery mechanism is triggered. Therefore, in general, the quality of the recovery image can be affected. In Ref. [26], a SLT-based (Slant Transform) semi-fragile watermarking scheme for image authentication and self-restoration is proposed, where two watermarks, a pseudo-random watermark sequence and the compressed version of the image generated from the SLT coefficient are used. Since the second

watermark (the compressed version) is embedded in the LSB plane, after content preserving modification such as JPEG compression, the altered region cannot be recovered. The importance of reliable tampering detection algorithms with recovery capability has given as a result the development of other efficient algorithms using both fragile and semi-fragile watermarking-based authentication algorithms.

In this paper, two watermarking-based image authentications algorithms with localization and recovery capability of the tampered region are proposed. In both algorithms, a halftone version of the original gray-scale image, which is used as an approximated version (image digest) of the original image, is embedded as a watermark sequence into given transform domains of the host image using a quantization-based watermarking algorithm. The first algorithm embeds the watermark sequence into the LL sub-band of the IWT domain, which is called *WIA-IWT* (*Watermarking-based Image Authentication using IWT*). In the second algorithm, which is called *WIA-DCT* (*Watermarking-based Image Authentication using DCT*); the DCT domain is used for this purpose. In the authentication stage of both algorithms, the watermark sequence (halftone image) is firstly extracted and converted in a gray-scale image, and then it is compared with the input image in a block-wise manner to determine the authenticity of the image or detect the tampered blocks. The similarity of both images (extracted and input images) is measured using the Structure Similarity (SSIM) index proposed in Ref. [28]. The SSIM index is an image quality assessment based on Human Visual System (HVS); therefore to employ efficiently this index in the proposed algorithms, the SSIM parameters are adjusted. Once the image is considered as tampered, the recovery process is started to estimate the original gray-scale image of the tampered block from the extracted halftone image. In the recovery stage, a Multilayer Perceptron (MLP) neural network is used to perform the inverse halftoning process to improve the quality of the recovered image. Two principal contributions of the proposed algorithms are the higher quality of recovered image compared with the quality of other semi-fragile recovery algorithms [20–23], and their high tamper detection capability while keeping sufficient robustness against content preserving modification. The first contribution is obtained by inverse halftoning process based on MLP neural network and a permutation method with push-aside operation [9], which reduces image digest loss caused by tampering/missing coincidence. The second contribution is obtained by the adjusted SSIM criterion used to determine if the input block is tampered or not.

There are some proposals in which the halftone image is used as a watermark sequence to recover the tampered regions [6,7,23,27]. In Refs. [6,7], the halftone image generated from the original image is replaced by the LSB plane of the original image. These methods can be categorized as fragile watermarking technique, because the embedded watermark sequence is vulnerable to common content preserving attacks, such as image compression and image format conversion, due to their LSB embedding domain. In Ref. [27], the authors proposed an image authentication algorithm based on the DCT, although it provides a high watermark imperceptibility and robustness, it is difficult to keep low false alarm and false negative error rates at the

same time, due to the Mean Square Error (MSE) criterion used to take decision if the block is tampered or not. A brief description of Ref. [23] and comparison results with the proposed algorithm are done in Section 3.

The rest of this paper is organized as follows. Section 2 describes the proposed algorithms and experimental results including comparison with previous reported image authentication algorithms with recovery capability [20–23], which are presented in Section 3. Finally Section 4 concludes this work.

2. Proposed algorithms

Both proposed authentication algorithms (*WIA-IWT* and *WIA-DCT*) consists of three stages: self-embedding, authentication and recovery stages as shown in Fig. 1. In the self-embedding stage, the image digest is generated from the original image and embedded into the image as a watermark sequence to get a watermarked image. Once the watermarked image is obtained, generally it could be sent through an open communication channel, where intentional and non-intentional alterations over the image may occur. In the authentication stage, the suspicious image, which is the watermarked and possibly tampered image, is received and then the watermark sequence is extracted from this image. The authenticity of the suspicious image is analyzed using the extracted watermark sequence. If the image is considered as tampered, then the recovery stage is triggered to get the information required to recover the tampered region. Each stage of the proposed algorithm is described in the following sub-sections.

2.1. Self-embedding stage

The self-embedding stage has two different processes, the first one is the watermark sequence generation and the second one is the embedding process. The watermark sequence generation is the same for both algorithms

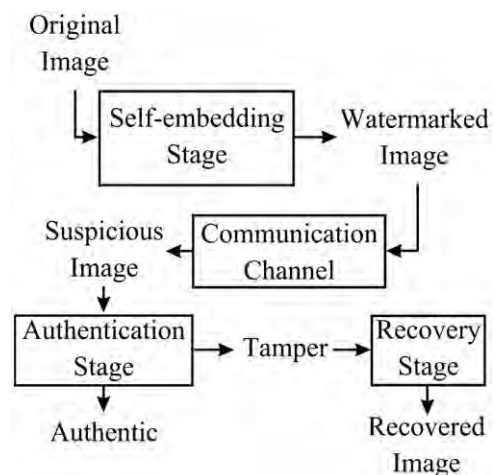


Fig. 1. General scheme of the proposed algorithms.

(WIA-IWT and WIA-DCT) and the embedding process is slightly different between two algorithms due to the difference of the embedding domains. In the watermark generation, the watermark sequence is generated from the down-sampled original image with half size in height and width. This operation is carried out to reduce watermark bit length, because an adequate embedding space in transform domain (IWT and DCT) is reduced compared with LSB domain in order to keep the watermark imperceptibility. We found that a suitable down-sampled size is equal to half size in width and height of the original image size, considering a tradeoff between the watermark imperceptibility and the recovered image quality. Once the original image is down-sampled, the error diffusion halftoning method proposed by Floyd–Steinberg [29] is applied to the down-sampled image to get the halftone image. Fig. 2(a) shows the Floyd–Steinberg error diffusion method. In this figure, Q means the quantization process that converts a gray-scale pixel value into a binary one, $b(i, j)$, using a threshold value $T_b = 128$ which is half value of 8-bits gray-scale dynamic range, as shown in Fig. 2(b). Then the error sequence $e(i, j)$ which is the difference between $u(i, j)$ and $b(i, j)$ is introduced to the 2D-filter H , whose coefficients are shown in Fig. 2(c), here black circle means the current pixel. The coefficients of H represent the diffusion ratio of the error produced in Q to the neighbor pixels (only four future neighbors in raster scan manner are considered). Here we used Floyd–Steinberg coefficients to generate halftone image, although Jarvis or Stucki coefficients can also be used [29] obtaining similar quality of halftone image.

The generated halftone image is scrambled by a permutation algorithm with push-aside operation [9] using a user's secret key, which is done to reduce watermark bit loss caused by tampering/missing coincidence and to increase the security assessment. The term tampering/missing coincidence is introduced in Ref. [10], which means that a recovery watermark (image digest) is embedded into a region where tampering occurs and as consequence of this, the embedded recovery watermark is lost. To reduce this coincidence, the recovery watermark must be embedded far from its corresponding region.

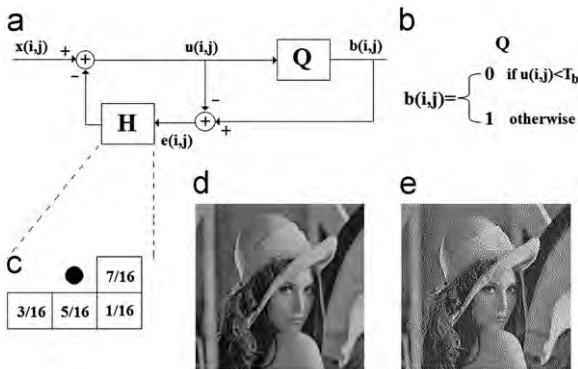


Fig. 2. (a) Floyd–Steinberg error diffusion halftoning method, (b) quantization operation Q (c) Floyd–Steinberg error diffusion filter, (d) original gray-scale image and (e) halftone version of (d).

The permutation method proposed in Ref. [9] guarantees that the recovery watermark for top-left (top-right) quarter part of the image is embedded into bottom-right (bottom-left) quarter part of the image, and vice versa. The permuted halftone data in a vector form W is the watermark sequence of the proposed algorithms.

In the embedding process for the WIA-IWT, the original image is decomposed using the 2D-IWT to obtain four sub-bands: LL, LH, HL and HH. The watermark sequence is embedded into the sub-band LL using the quantization-based watermarking method. On the other hand, in the WIA-DCT, the original image is segmented into 8×8 pixels blocks, which are transformed by using the 2D-DCT. The watermark sequence W is divided into blocks of 16 bits, and each block of W is embedded into the middle frequency range (shadowed coefficients in Fig. 3) of each block using the quantization-based watermarking method. The quantization embedding formula used for both algorithms is given by

$$\tilde{c}_n = \begin{cases} v_1 & \text{if } |c_n - v_1| \leq |c_n - v_2| \\ v_2 & \text{otherwise} \end{cases} \quad (1)$$

$$v_1 = \begin{cases} \text{sgn}(c_n) \lfloor \frac{|c_n|}{2S} \rfloor 2S, & \text{if } w_n = 0 \\ \text{sgn}(c_n) (\lfloor \frac{|c_n|}{2S} \rfloor 2S + S), & \text{if } w_n = 1 \end{cases}, n = 1 \dots N$$

$$v_2 = v_1 + \text{sgn}(c_n) 2S \quad (2)$$

where w_n is the n -th watermark bit, N is the watermark bits length, c_n and \tilde{c}_n are the n -th original and the watermarked IWT/DCT coefficients arranged in vector form, respectively, and S is the quantization step-size. Finally we obtained the watermarked image applying the inverse transform: inverse 2D-IWT or inverse 2D-DCT, to the watermarked coefficients. This stage is shown in Fig. 3.

2.2. Authentication stage

The block diagram of the authentication stages of WIA-IWT and WIA-DCT are shown in Fig. 4. These stages are composed by two different processes, the first one is the watermark extraction, and the second one is the authentication process using the extracted watermark sequence.

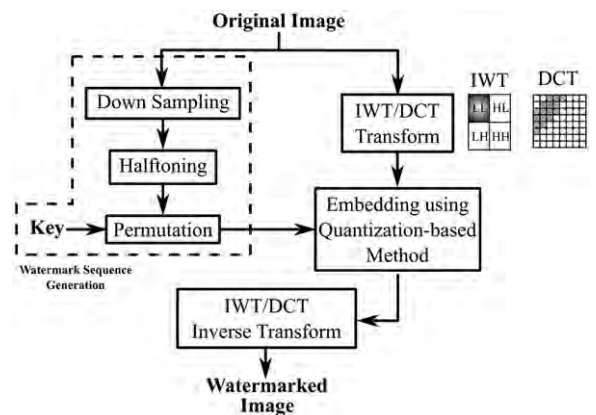


Fig. 3. Self-embedding stage of proposed algorithms WIA-IWT and WIA-DCT.

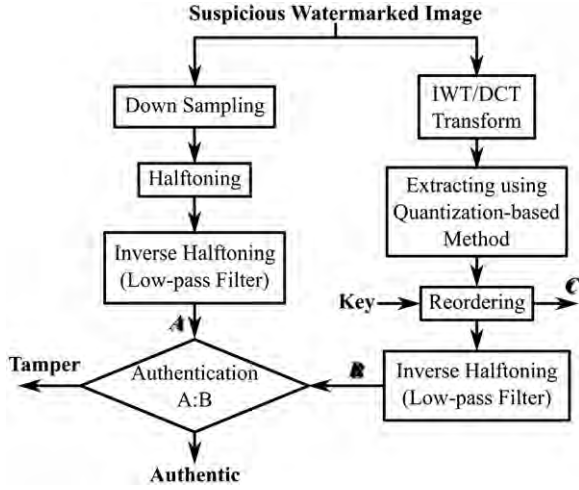


Fig. 4. Authentication stage of proposed WIA-IWT and WIA-DCT algorithms.

The watermark bits \tilde{w}_n of the suspicious image are extracted from the sub-band LL in case of the WIA-IWT or from all blocks in DCT domain in the case of the WIA-DCT. The watermark extraction is given by:

$$\tilde{w}_n = \begin{cases} 0 & \text{if } \text{round}(\hat{c}_n/S) = \text{even} \\ 1 & \text{if } \text{round}(\hat{c}_n/S) = \text{odd} \end{cases}, \quad n = 1 \dots N \quad (3)$$

where \tilde{w}_n is the n -th extracted watermark bit, and \hat{c}_n is n -th coefficient of LL sub-band of the IWT or n -th DCT coefficient of each block of the suspicious image. S is the same quantization step-size used in embedding stage for both algorithms. The extracted watermark bits \tilde{w}_n are reordered using the user's secret key given in the embedding stage, which is the halftone version of the original image and then it is converted to a gray-scale image, 'B' in Fig. 4, using Gaussian low-pass filter F_G given by the following equation:

$$F_G = \frac{1}{11.566} \begin{bmatrix} 0.1628 & 0.3215 & 0.4035 & 0.3215 & 0.1628 \\ 0.3215 & 0.6352 & 0.7970 & 0.6352 & 0.3215 \\ 0.4035 & 0.7970 & 1 & 0.7970 & 0.4035 \\ 0.3215 & 0.6352 & 0.7970 & 0.6352 & 0.3215 \\ 0.1628 & 0.3215 & 0.4035 & 0.3215 & 0.1628 \end{bmatrix} \quad (4)$$

Next, we also generate a halftone image from the down-sampled suspicious watermarked image, which is re-converted into a gray-scale image, 'A' in Fig. 4, using the same Gaussian low-pass filter F_G . The halftoning and inverse halftoning processes are quite important and essential, because the generated gray-scale image 'B' is blurred due to low-pass filter, therefore the direct comparison between the image 'B' and the down-sampled watermarked image results in large errors. These processes give the same level of blurring to the received image 'B' and to the image 'A', and then the difference between the image 'A' and the image 'B' reflects the real alteration occurred. The inverse halftoning used here is the simplest method with lower computational cost, even though it produces low quality gray-scale image, because

in this stage, a low computational complexity and an efficient detection of the modified areas is more important than a high quality of the gray-scale image. The comparison between the gray-scale images 'A' and 'B' is carried out using a block-wise strategy of size $L \times L$ pixels and then SSIM index [28] with adapted parameters of both images 'A' and 'B' is calculated, which is given by

$$SSIM_k(A_k, B_k) = [l(A_k, B_k)]^\alpha [c(A_k, B_k)]^\beta [s(A_k, B_k)]^\gamma \quad k = 1 \dots K \quad (5)$$

where $l(A, B)$, $c(A, B)$ and $s(A, B)$ are luminance, contrast and structure similarities between images 'A' and 'B', which are given by Eqs. (6)–(8), respectively, and α , β and γ are parameters that control the importance level of each similarity [28]. Sub-index k is block index, A_k and B_k are k -th block of $L \times L$ pixels and K is number of total blocks, such that $L \times L \times K$ is equivalent to the watermark length N .

$$l(A, B) = \frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1} \quad (6)$$

where μ_A , μ_B are the mean values of images 'A' and 'B', and C_1 is a small constant which avoids division by zero when both mean values are zeros. The range of this similarity is $[0, 1]$, if both images are identical then $l(A, B) = 1$. The contrast similarity is given by

$$c(A, B) = \frac{2\sigma_A\sigma_B + C_2}{\sigma_A^2 + \sigma_B^2 + C_2} \quad (7)$$

where σ_A , σ_B are standard deviations of images 'A' and 'B', and C_2 is a small constant with the same purpose as C_1 . Also this similarity is within $[0, 1]$, and when both images are identical, $c(A, B) = 1$. The structure similarity is the correlation value between the normalized images 'A' and 'B' by luminance and contrast, which is given by

$$s(A, B) = \frac{\sigma_{AB} + C_3}{\sigma_A\sigma_B + C_3} \quad (8)$$

where σ_{AB} is a covariance between both images, C_3 is a small constant and $|s(A, B)| \leq 1$.

Content preserving modifications, such as JPEG compression, scaling, filtering and rotation, produce in general a smoothing effect; as a consequence the contrast similarity given by Eq. (7) may be varied, while the variations of luminance similarity and structure similarity are small. On the other hand, the tampering produces significant variation in luminance and structure similarity. Therefore the smaller β makes SSIM more robust to the content preserving modification, while setting other two parameters α , γ equal to 1 to keep tamper detection capability. Because in this situation, the contrast similarity becomes closed to 1 and then the SSIM criterion strongly depend on only luminance and structure similarity. To compute the SSIM value in more accurate manner, the block size $L \times L$ must be sufficiently large (e.g. $L = 64$), however to detect tampered region with small extension, the block size $L \times L$ must be small (e.g. $L = 8$), actually we set the block size L is equal to 8 in all experiments. To solve this problem, image blocks 'A_k' and 'B_k' are scaled-up by eight times. The SSIM index of each block $SSIM_k(A_k, B_k)$, which is $|SSIM_k(A_k, B_k)| \leq 1$, is compared with a predefined threshold value Th , and if $SSIM_k(A_k, B_k) > Th$ then image blocks

' A_k ' and ' B_k ' are visually similar, therefore k -th block is considered as authentic, otherwise k -th block is considered as tampered. This operation is given by

$$\begin{cases} SSIM_k(A_k, B_k) \leq Th & k\text{-th block is tampered} \\ SSIM_k(A_k, B_k) > Th & k\text{-th block is authentic} \end{cases} \quad (9)$$

The threshold value Th is very important because it controls the false alarm error rates f_a and false negative error rates f_n , therefore we discuss about an adequate threshold value in Section 3. Although using an adequate threshold value Th determined as mentioned above, some isolated blocks may be detected as tampered one, however an isolated block with size of 16×16 pixels compared with the whole image of 512×512 pixels is visually insignificant, and then using the connected component labeling algorithm with eight adjacency [30], the isolated blocks are eliminated.

2.3. Recovery stage

If the authentication stage determines that some blocks of the suspicious image are tampered, then the recovery process will be triggered. The down-sampled suspicious watermarked image, its halftone version, the indexes of the altered blocks and the extracted halftone image (image 'C' in Fig. 4) are introduced as input data of this stage. Unlike the authentication stage, in this stage the quality of the recovered gray-scale image generated from the halftone image is very important, therefore a MLP-based inverse halftoning method is proposed. The unaltered blocks of the down-sampled suspicious image and its halftone version are used to train the MLP using the Backpropagation (BP) algorithm [31]. This recovery stage is shown in Fig. 5(a) and the MLP used to estimate the gray-scale image is shown in Fig. 5(b).

In the MLP training stage, the 4×4 neighborhood template of pixel 'X' generated from the unaltered blocks of the halftone image, shown in bottom-left part of the Fig. 5(b), is used as an input training pattern of BP algorithm. The desired output data is a gray-scale value of the corresponding pixel 'X' with the same neighborhood. Using those input and reference data, the MLP with structure 16–4–1 (16 input data, 4 neurons in a hidden layer and 1 output) is trained by using the BP algorithm in order that the connection weights of the MLP become the optimal values associating a given neighborhood of the halftone image with a gray-scale value 'X'. Once the MLP is trained, the tampered area of the halftone image is introduced to the trained MLP to get a better quality

of the gray-scale recovered region. In the general case of inverse halftoning, the gray-scale image may not be available, however in the proposed scheme; we can use the gray-scale image of the corresponding halftone image of the non-altered area to train a MLP, taking advantage of generalization capability of the MLP. This fact allows the generation of a high quality gray-scale image of the tampered regions. Fig. 6 shows an example of this process. The authentication stage of the proposed algorithm detects the tampered region indicated by black box, the extracted watermark sequence corresponding to the detected tampered region (the black box) is halftone image of this region as shown by 'Extracted Halftone Image' of Fig. 6, which corresponds to the signal 'C' in Figs. 4 and 5. The down-sampled image and its halftone version are the desired reference and input data, respectively, to train the MLP neural network as shown by Fig. 6. The 4×4 neighborhood template is scanned from left-to-top to bottom-right of the down-sampled halftone image, excluding tampered region indicated by the black box. In each scan position, 16 binary data corresponding to the 4×4 neighborhood template are introduced to MLP neural networks and the BP algorithm updates the connection weights using the desired data indicated by 'X' in the template. This operation is repeated until the BP algorithm converges. Once the MLP training stage is finished, the MLP has acquired a capacity to associate 16 binary data in neighborhood template to an equivalent gray-scale data located by 'X'. The gray-scale image of tampered region is generated introducing halftone image 'C' to the trained MLP. Fig. 7 shows a comparison between gray-scale images obtained using inverse halftone based on Gaussian low-pass filter given by Eq. (4), and the corresponding images using the proposed MLP-based inverse halftoning approach.

The PSNR of both images with respect to the original one are 23 dB and 32 dB, respectively, which indicates that the gray-scale image generated by the MLP-based inverse halftoning method has much better quality than the low-pass filter based method. Considering the perceptual quality, the images generated using the MLP-based method can conserve more details of the original images than those obtained using the Gaussian low-pass filter-based method.

3. Experimental results and analysis

In the proposed algorithms, an adequate setting of two parameters: quantization step-size S and threshold Th , is

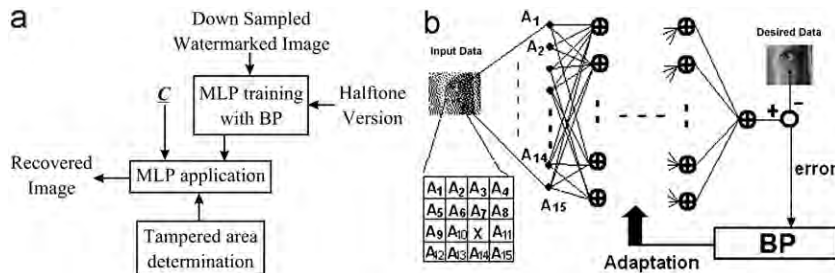


Fig. 5. (a) Recovery process, and (b) proposed MLP-based structure for estimating the gray-scale image.

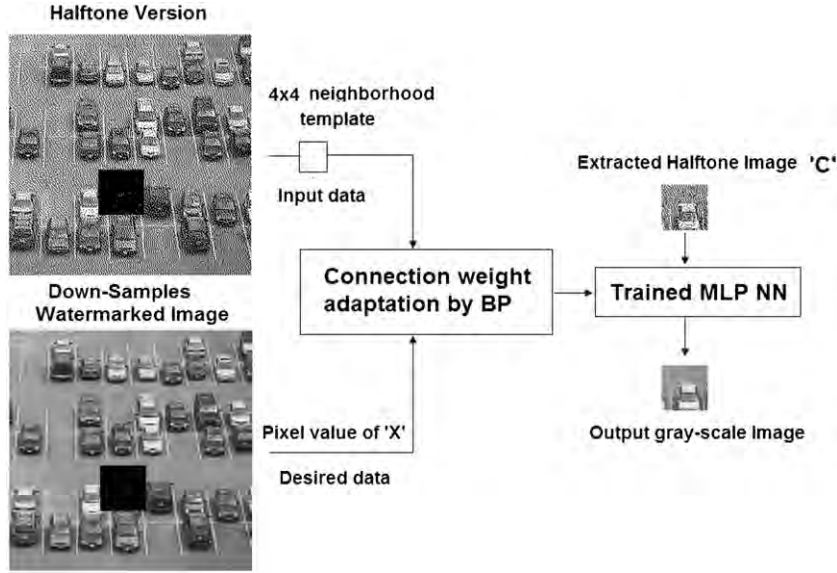


Fig. 6. An example of recovery process shown in Fig. 5(a).

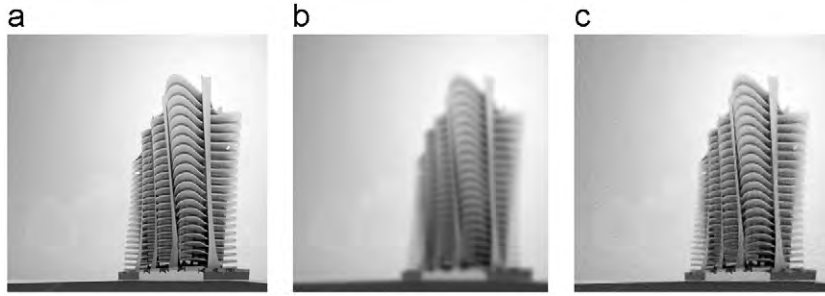


Fig. 7. Image quality comparison: (a) original Image, (b) gray-scale image generated using Gaussian low-pass filter (PSNR 23 dB), and (c) gray-scale image generated using the proposed MLP-based method (PSNR 32 dB).

considerably important. The quantization step-size S is used in the watermark embedding and extraction process given by Eqs. (1) and (2), which controls the watermark imperceptibility and robustness, while the threshold value Th controls the false alarm and false negative error rates which is directly related to robustness and tamper detection capability of the proposed algorithm. Considering this, firstly adequate values for step-size S and threshold Th must be analyzed. In all cases three parameters α , β and γ of SSIM criterion are set equal to 1.0, 0.5 and 1.0, whose values are determined in trial and error, considering common effects caused by almost all content preserving modification as mentioned in Section 2.2.

3.1. Watermark imperceptibility

As mentioned above, the watermark imperceptibility depends on the quantization step-size S , when S is increased, the robustness of the schemes is increased while the watermark imperceptibility is decreased. Fig. 8 shows the relationship between S and the average PSNR values of the watermarked image respect to its original version using 100 images and 10 different secret

keys for the permutation. The 100 images used in the evaluation of the proposed algorithms are typical test images, such as Lena, Baboon, Pepper, etc. and images captured by authors using digital camera. The size of all images is 512×512 . Fig. 8 shows the relationship between S and average PSNRs in WIA-IWT and WIA-DCT algorithms. Table 1 provides the average PSNR of the watermarked image generated by WIA-IWT and WIA-DCT using different step-size S , and Fig. 9 shows an example of the watermarked images generated by both algorithms together with its original one, in which quantization step-size S is set to 10 for WIA-IWT and 12 for WIA-DCT.

From Figs. 8 and 9, and Table 1, the watermark imperceptibility is guaranteed when the selected quantization step-size S is smaller than 10 for WIA-IWT and 12 for WIA-DCT algorithms.

3.2. Robustness of the proposed algorithms

The robustness in any content authentication schemes is that the schemes declare authenticity of the input image while the content of image is not modified, although it

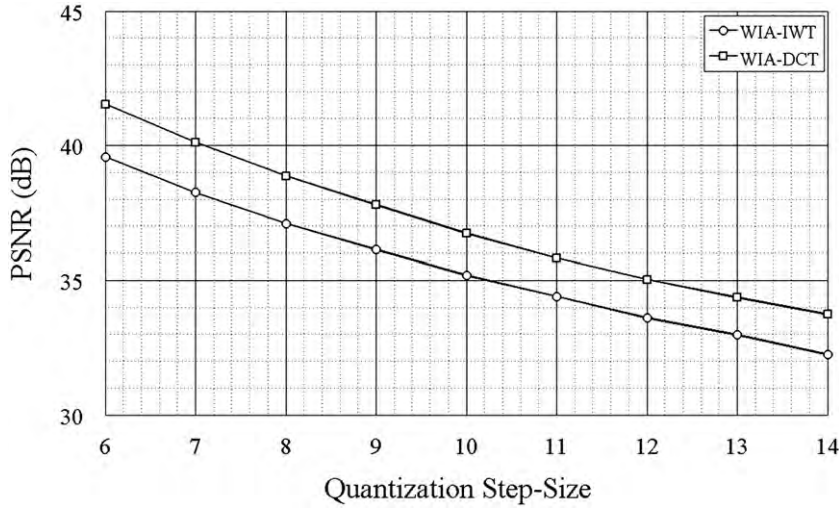


Fig. 8. Relationship between step-size S and watermark imperceptibility in the proposed WIA-IWT and WIA-DCT.

Table 1

Watermark imperceptibility of the proposed WIA-IWT and WIA-DCT with different step-size S .

Algorithms/ S	6	7	8	9	10	11	12	13	14
WIA-IWT	39.58	38.25	37.10	36.13	35.18	34.40	33.60	32.98	32.23
WIA-DCT	41.54	40.12	38.87	37.80	36.73	35.81	35.03	34.36	33.74

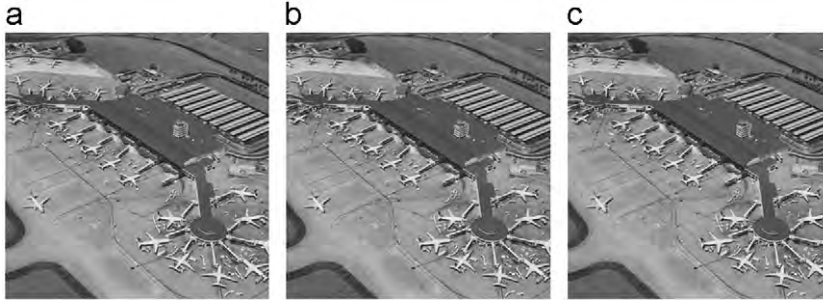


Fig. 9. Example of watermarked images together with their original version: (a) original image, (b) watermarked image (35.28 dB) generated by WIA-IWT using $S=10$ and (c) watermarked image (35.79 dB) generated by WIA-DCT using $S=12$.

receives several content preserving modifications, such as JPEG compression, filtering, scaling and rotation, etc. In both proposed algorithms, the SSIM given by Eq. (5) is calculated in each block with $L \times L$ pixels and then it is compared with a predefined threshold value Th to determine if each block of the input image is tampered or not, as given by Eq. (9). Therefore the robustness of our scheme must be evaluated using the SSIM values of each block together with the threshold Th under several quantization step-size S . Firstly to determine an adequate threshold value Th , the distribution of SSIM values must be analyzed under two conditions: the input image receives content alterations of 1.56% of the whole pixels and the input image receives JPEG compression with several quality factors Q_f . The 1.56% of the whole pixels is equivalent to 16 blocks of 16×16 pixels in an image of 512×512 pixels, which corresponds to the smallest visually recognizable object.

Fig. 10 shows the probability density function (PDF) of SSIM values and MSE values, which is used in [27]. The step-sizes S was selected to generate watermarked image with the same PSNR using the WIA-IWT and WIA-DCT. The WIA-IWT with $S=10$ and WIA-DCT with $S=12$ generate the watermarked images with approximately PSNR=35 dB. Fig. 11 shows Receiver Operating Characteristic (ROC) curves of the SSIM criterion and the MSE criterion with two quality factors Q_f of JPEG compression are equal to 100 and 70, where the quality of the watermarked image is equal to 35 dB. In this figure, the probability of false negative error rates f_n with respect to the probability of false alarm error rates f_a is shown.

Both figures indicate that the performance of the SSIM criterion is much better than that of the MSE criterion, because using the SSIM criterion, the lower f_a can be obtained without sacrificing the f_n and vice versa,

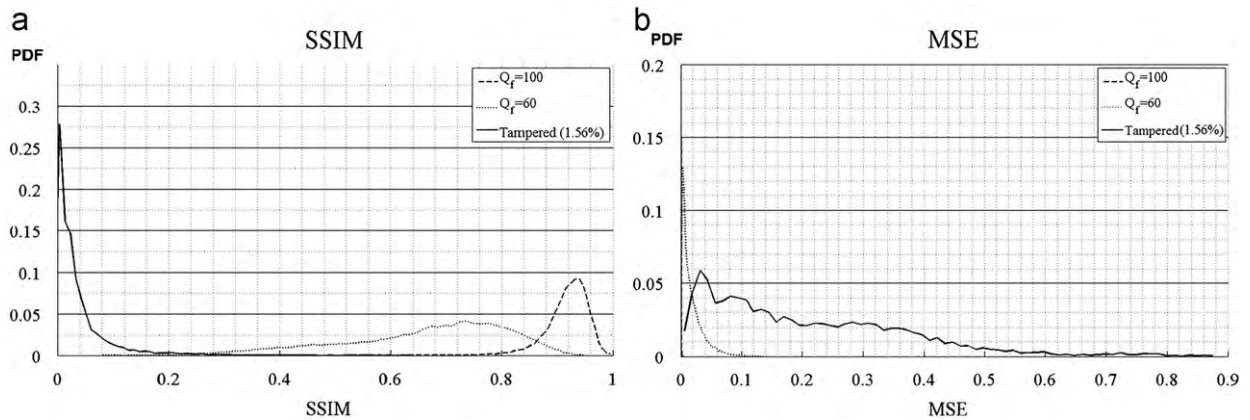


Fig. 10. Probability density function (PDF) of the SSIM and MSE values under the tampering and JPEG compression with $Q_f=100$ and 60: (a) PDF of SSIM values, (b) PDF of MSE values, in both cases the quality of the watermarked image is PSNR=35 dB.

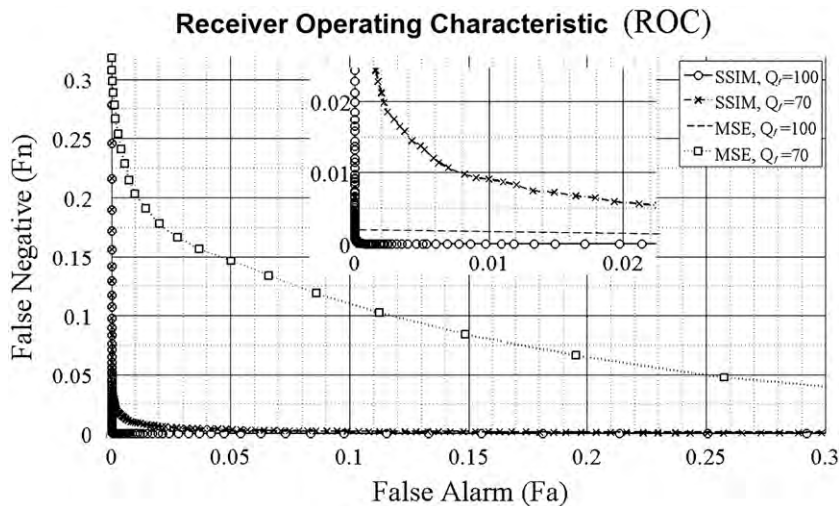


Fig. 11. Receiver Operating Characteristic curves of SSIM and MSE criteria together with zoomed image of bottom-left part.

Table 2

Relationship between threshold value and false alarm error rates in % under JPEG compression.

	Without attack			80			70			65			60		
Th	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5	0.3	0.4	0.5
f_a (%)	0	0	0	0	0	0	0	0	0	0	0.02	0.09	0.03	0.4	4.6

however using the MSE criterion to obtain lower f_n , the f_a must be sacrificed. Table 2 shows false alarm error rates f_a using three different threshold values (0.3, 0.4 and 0.5) under JPEG attacks with different quality factors, while Table 3 shows false negative error rate f_n using these three threshold values.

In the proposed algorithms, we consider that the false negative error rate f_n must be as small as possible, while keeping the false alarm error rate f_a reasonably small, because the false alarm errors do not cause the damage in the objective of the proposed algorithms. Generally in watermarking field, especially in copyright protection and copy control applications, the threshold value

Table 3

Relationship between threshold value and false negative error rates.

Th	0.3	0.4	0.5
f_n	0.0026	4.76×10^{-5}	9.52×10^{-6}

is calculated controlling false alarm error rate f_a as small as possible, because the false alarm error may accuse erroneously an innocent person of copyright violation, and in copy protection task it prohibits copying even if the digital material is copy free [32]. On the other hand, in our proposed schemes, the false alarm error triggers the

recovery process, however, in which the same contents as the original version can be recovered, and then the recovered image shows the same contents as the original one. On the other hand, the false negative error causes that the recovered image cannot present the whole contents of the original image. Under above mentioned philosophy, the adequate threshold value can be obtained under the condition given by

$$f_n = \int_{Th}^1 Nh_{(1.56\%)}(k) \leq 10^{-4} \quad (10)$$

where $Nh_{(1.56\%)}$ is the estimated probability when 1.56% of the pixels of the images are tampered and Th is the adequate threshold value. Taking in account the f_n in Table 3, we determined an adequate threshold value Th , which is $Th=0.4$. Using the selected threshold value Th , the robustness of the proposed algorithms is evaluated. Table 4 shows f_a when the input watermarked images receive several content preserving attacks. In case of down-up sampling attack, the watermarked images are scaled using the indicated scaling factor and then re-scaled to obtain the images with same size, and also in the rotation attack, firstly the images are rotated by the given angle and then these are rotated in the inverse direction by the same angle to obtain the same orientated image.

Table 4

False alarm error rates in % under several content preserving attacks. A: Dynamic range change [50–200], B: histogram equalization, C: median filter (3×3), D: Gaussian Noise ($\sigma^2=0.009$), E: down-Up sampling (scaling factor 0.9), F: rotation (3degree), G: impulsive noise ($d=1\%$), and H: Speckle noise ($d=0.9$).

Attacks	A	B	C	D	E	F	G	H
f_a (%)	0.4	0.09	7.2	1.2	2.0	2.3	1.4	1.5

Table 5

Main characteristics of previously proposed semi-fragile algorithms [20–23] and the proposed one.

Algorithm	W1 (digital signature)	W2 (image digest)	Embedding domain for W1	Embedding domain for W2	Embedding algorithm
Piva [20]	No use	Lower DCT coefficients of LL1 sub-band of DWT	–	HH2,VV3 of DWT (2nd decomposition)	Replacement
Chamlawi1[21]	BinP	Lower DCT coefficients of LL1 sub-band of IWT	LL2,VV2, DD2,LL3, HH3,DD3, DD1 of IWT	HH2,VV3 of IWT (2nd decomposition)	W1: quantization W2: replacement
Chamlawi2[22]	BinP	Huffman encoded sequence of LL1 sub-band of IWT	LL2,VV2, DD2,LL3, HH3,DD3, DD1 of IWT	HH2,VV3 of IWT (2nd decomposition)	W1: quantization W2: replacement with some suitable coefficients
Phadikar [23]	BinP	Halftone image of LL2 sub-band of IWT	LH2, HL2 of IWT	LH1,HL1 of IWT	QIM-dither modulation
Proposed WIA-IWT Proposed WIA-DCT	No use	Halftone image of down-sampled image	–	LL1 of IWT 16 lowest coefficients except DC and 2 lowest AC	Quantization

The robustness of the proposed algorithms is compared with that of previously reported algorithms [20–23]. All of these algorithms are semi-fragile watermarking with tamper detection and recovery capability. The brief description of each algorithm is provided in Table 5.

The second column indicates the type of authentication watermark (digital signature) which is used to determine the tamper region, where ‘BinP’ means a binary pattern and ‘No use’ means that this type of watermark is not necessary, because W2 (image digest) is also used for this purpose. In Chamlawi’s works [21,22], a pseudo-random binary pattern related with LL subband is used, while in Phadikar’s work [23], the user’s predetermined binary pattern is used. In Piva’s work and the proposed algorithms, W2 is used for both purposes: tamper detection and recovery. In forth column, LL1, VV1 and DD1 ($l=2$ or 3) are sub-bands generated by wavelet decomposition of HL1 or LH1, respectively, and DD1 is highest sub-band of first level decomposition.

The tamper detection method and a criterion together with a recommended threshold value used to determine if the input image is tampered or not, are quite different in each algorithm [22–23], therefore the Receiver Operating Characteristic (ROC) curve is considered as a reliable assessment to compare robustness and tamper detection capability of each algorithm in a fairly manner. The authentication watermark (digital signature) W1 and its embedding algorithm of Refs. [21] and [22] are the same, therefore these two algorithms are considered as same algorithm for computation of f_a and f_n . It is worth noting that the watermarking embedding energy and some parameters used in each algorithm are adjusted to generate the watermarked images with same quality whose PSNR is 35 dB. Fig. 12 shows the ROC of three algorithms [20–23] and the proposed one, here the watermarked images are compressed by JPEG compression with quality

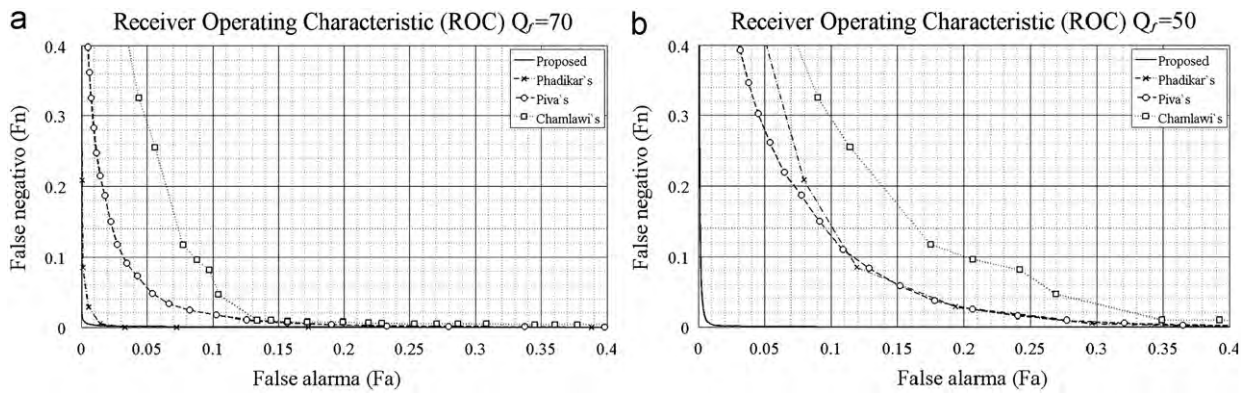


Fig. 12. Receiver operating characteristics of three previous works and the proposed algorithm when watermarked image received JPEG compression attack: (a) quality factor $Q_f=70$ and (b) quality factor $Q_f=50$.

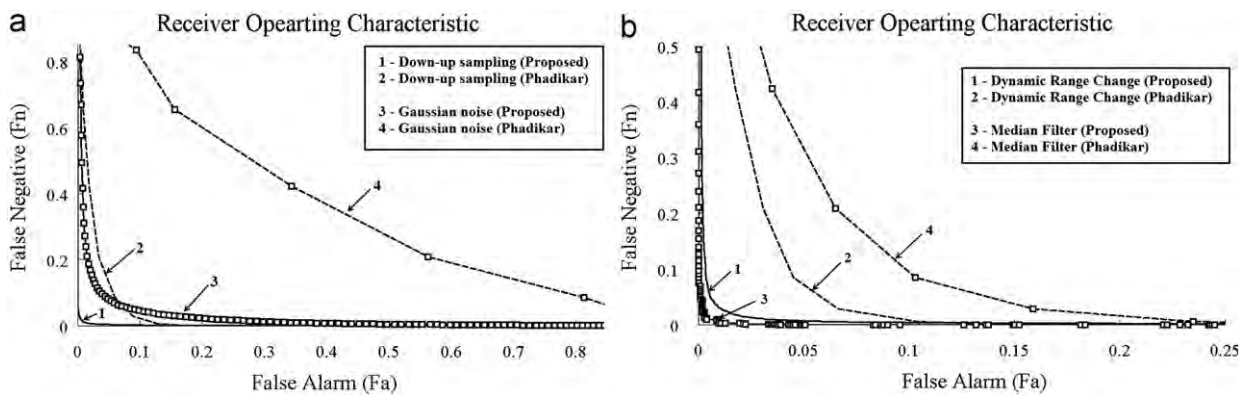


Fig. 13. ROC curves of the proposed algorithm and Phadikar's algorithm: (a) down-up sampling with scaling factor 0.9 and Gaussian noise with variance 0.009, and (b) dynamic range change into [50–200] and Median filter with window of 3×3 .

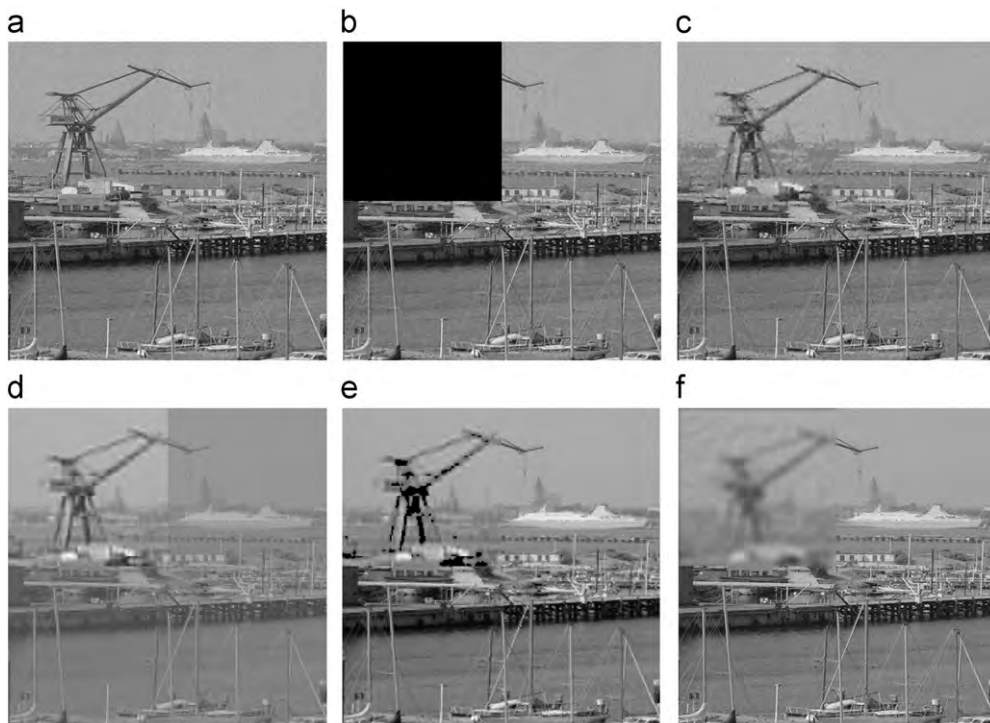


Fig. 14. Comparison of quality of recovered image: (a) watermarked image (b) tamper detection result which indicates tampered region by black box, (c) recovered image using the proposed algorithm with 30.31 dB, (d) recovered image using Piva's method with 26.07 dB, (e) recovered image using Chamlawi's method with 22.80 dB, and (f) recovered image using Phadikar's method 23.46 dB.

factor $Q_f=70$ and $Q_f=50$. Fig. 13 shows comparisons of robustness between the proposed algorithm and Phadikar's one [23].

From these figures, we can conclude that the proposed algorithm show higher robustness to several content preserving modifications than the previous works. The ROC curves provided in Figs. 12 and 13 show that the proposed algorithm provides lower false alarm error rates

and false negative error rates at the same time under several content preserving modifications. It is worth noting that all evaluation results are obtained using the same 100 images.

3.3. Recovery capability

As mentioned above, the principal contribution of the proposed algorithm is the generation of a high quality recovered image, even the presence of false alarm errors. This desirable performance of the proposed algorithm is obtained by MLP-based inverse halftoning processing and permutation with push-aside operation [9]. Fig. 14 shows a comparison of the proposed algorithms in the quality of recovered images with above mentioned four algorithms [20–23]. Actually several factors, such as false alarm errors, false negative errors, the extension of tampered region, tampering/missing coincidence rate and characteristics of input image (plane image or detailed image)

Table 6

Relationship between tampering extension rates regarding to the original image size and recovered image quality.

Tampering extension (%)	PSNR (dB)
23	32
28	27
35	24
40	23

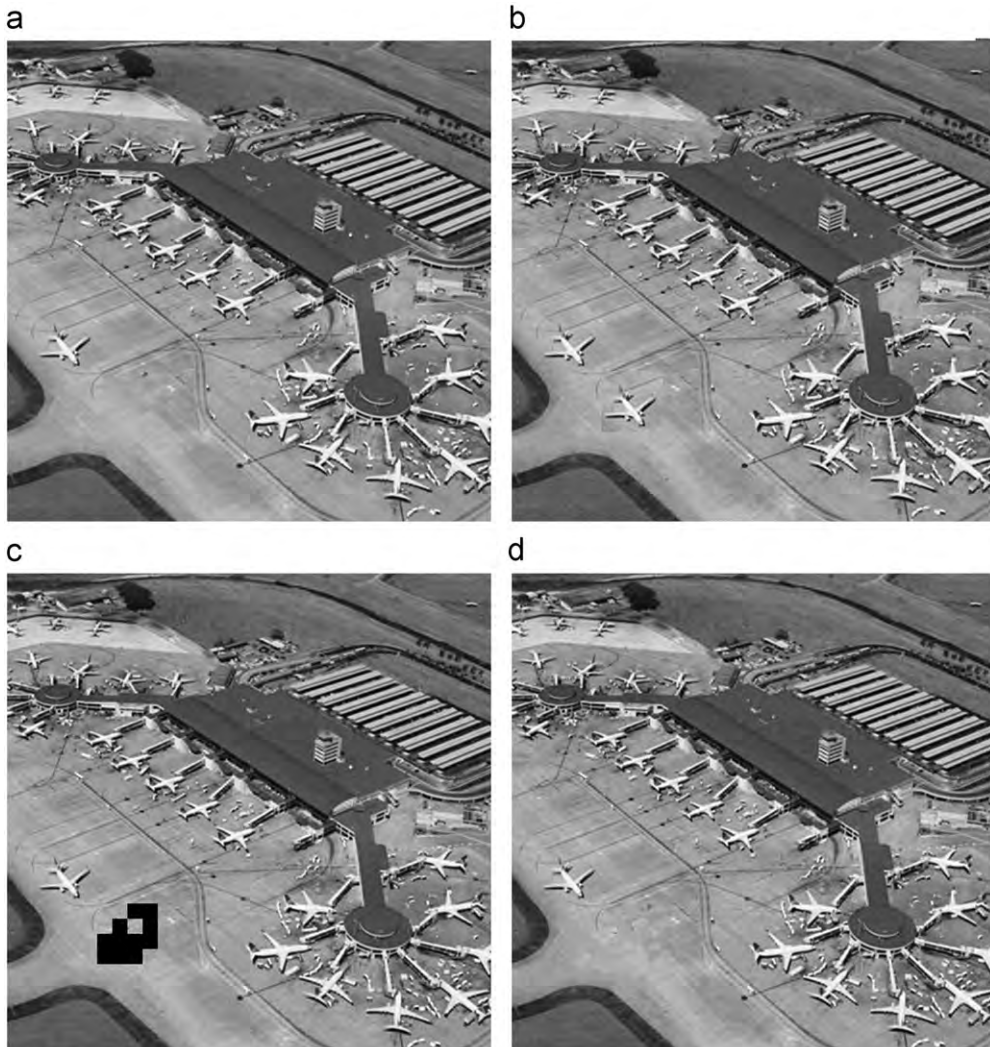


Fig. 15. Tamper detection and recovery capability: (a) original image, (b) tampered image with 1.35% tampering pixels, (c) tampered detection and (d) recovered image with a PSNR equal to 37.81 dB.

affect considerably the quality of recovered image in each algorithm, therefore to perform a fairly comparison, the recovered image of each algorithm is generated in an ideal condition, in other words neither false alarm nor false negative errors occur and tampering/missing coincidence rate is null. In all algorithms, their embedding parameters are adjusted to generate the watermarked image with PSNR 35 dB respect to the original one. Here expansion rate of tampered region is 25% which is located in top-left part of the image as shown in Fig. 14(b).

From Fig. 14, the quality of the recovered image of the proposed algorithm is better than that provided by the previous works [20–23] in terms of PSNR and also HVS. Use of halftone image as image digest and inverse halftoning method based on MLP neural networks provides a good quality of the recovered image of the proposed method. Phadikar's method [23] also uses halftone image as image digest, however their halftone image is generated from LL2 sub-band ($N/4 \times N/4$) of IWT decomposition

while our halftone image is generated from the down-sampled image with ($N/2 \times N/2$ pixels) of the original image of $N \times N$, reducing distortion introduced by interpolation operation in up-sampling.

Table 6 provides the PSNR of the recovered image respect to the watermarked one under different extension of the tampered regions. In the proposed scheme, until 25% of the tampered region can be recovered without tampering/missing coincidence. When the expansion rate exceeds 25%, the quality of the recovery image decreases due to loss of recovery watermark (image digest) caused by tampering/missing coincidence and false alarm errors.

Some examples of the tamper detection and recovery process of the proposed algorithms are shown by Figs. 15 and 16. In all cases, tampered versions are generated using simple copy-and-paste using common drawing tools, such as Photoshop and Coral Draw, etc.

In these figures, (a) shows the original un-tampered image, its tampered version is shown in (b), the tampered



Fig. 16. Tamper detection and recovery capability: (a) original image, (b) tampered image with 2.05% of tampering pixels (c) tampered detection and (d) recovered image with a PSNR equal to 34.73 dB.

image with the detected tampered blocks (indicated by black blocks) are shown by (c) and (d) shows the recovered images. In Fig. 15(b), an airplane is added into the image tampering the 1.35% of pixels of the image, in Fig. 16(b) a handgun is deleted tampering the 2.05% of pixels. From the Figs. 15(c) and 16(c), the proposed algorithms can detect correctly the tampered blocks and provide fairly good quality of the recovered images as shown in Figs. 15(d) and 16(d). The qualities of the recovered image respect to their watermarked versions are 37.81 dB and 34.73 dB, respectively. In some cases, the tamper detection process may detect erroneously some blocks without tamper (false alarm error); however these erroneous blocks do not affect the contents of the recovered image. These blocks only cause slight degradation of the recovered image respect to their original versions.

Because the proposed algorithms *WIA-IWT* and *WIA-DCT* differ only in the embedding domain, both of them provide the same tamper detection and recovery capability if the threshold is properly selected as mentioned above.

4. Conclusions

In this paper we proposed two image authentication algorithms based on watermarking technique. In the first algorithm, the watermark embedding is carried out in the Integer Wavelet Transform (IWT) domain and in the second one this process is performed in the Discrete Cosine Transform (DCT) domain. We called the first and second algorithms as *WIA-IWT* (*Watermarking-based Image Authentication using IWT*) and *WIA-DCT* (*Watermarking-based Image Authentication using IWT*), respectively. In both algorithms an image digest, which is used for tamper detection and recovery of the tampered region, is generated using error diffusion halftoning technique and then it is embedded into the image as watermark sequence using quantization-based watermarking algorithm. To obtain lower false alarm error rates when images receive several content preserving attacks, such as JPEG compression, filtering, rotation, scaling and so on, while keeping the false negative error rates lower than 10^{-4} , the Structural Similarity (SSIM) as a criterion is introduced to determine the tamper regions in a block-wise sense. The desirable performance of the proposed algorithms is shown in terms of false alarm error rates, false negative error rates and receiver operating characteristics (ROC) curve. To improve the quality of the recovered image, Multilayer Perceptron (MLP) neural network is introduced as inverse halftoning in the recovery process and also permutation with push-aside operation is introduced to reduce tampering/missing coincidence problem. The robustness and quality of recovered image of proposed algorithms are compared with previously reported semi-fragile watermarking algorithms [20–23] under the same condition (watermark imperceptibility of each watermarked image, same input images, etc.), showing better performance of the proposed algorithms. The proposed algorithm can recover the tampered region even with 25% of pixels in perfectly manner, which means that both false alarm and false negative errors are almost null, providing a good quality with more than 30 dB. When tampered extension exceeds 25%, the quality of recovered image may reduce due to loss

of digest image caused by tampering/missing coincidence problem, and false alarm errors. The proposed algorithms *WIA-IWT* and *WIA-DCT*, differ only in the embedding domain, then both of them provide the same tamper detection and recovery capability if the threshold is properly selected.

Acknowledgments

We thank the National Science and Technology Council of Mexico, National Polytechnic Institute of Mexico and the National Autonomous University of Mexico by support provided during the realization of this research. Special thanks for the reviewers who contributed to improve this paper with helpful suggestions.

References

- [1] A. Menezes, V. Oorschot, S. Vanstone, Handbook of Applied Cryptography, second ed. CRC, Boca Raton, FL, 1998.
- [2] P.W. Wong, N. Memon, Secret and public key image watermarking schemes for image authentication and ownership verification, IEEE Transactions on Image Processing 10 (10) (2001) 1593–1601.
- [3] M.U. Celik, G. Sharma, E. Saber, A.M. Tekalp, Hierarchical watermarking for secure image authentication with localization, IEEE Transactions on Image Processing 11 (6) (2002) 585–595.
- [4] S.-S. Wang, S.-L. Tsai, Automatic image authentication and recovery using fractal code embedding and image inpainting, Pattern Recognition 41 (2008) 701–712.
- [5] P.-L. Lin, C.-K. Hsieh, P.-W. Huang, A hierarchical digital watermarking method for image tamper detection and recovery, Pattern Recognition 38 (2005) 2519–2529.
- [6] H. Luo, S.-C. Chu, Z.-M. Lu, Self embedding watermarking using halftoning technique, Circuit Systems and Signal Processing 27 (2008) 155–170.
- [7] H. Luo, Z.-M. Lu, S.-C. Chu, J.-S. Pan, Self embedding watermarking scheme using halftone image, IEICE Transactions on Information and Systems E91-D (1) (2008) 148–152.
- [8] J. Fridrich, M. Goljan, Image with self-correcting capabilities, International Conference on Image Processing (3) (1999) 792–796.
- [9] T.-Y. Lee, S.-D. Lin, Dual watermarking for image tamper detection and recovery, Pattern Recognition 41 (2008) 3497–3506.
- [10] X. Zhang, Z. Qian, Y. Ren, G. Feng, Watermarking with flexible self-recovery quality based on compressive sensing and compositional reconstruction, IEEE Transactions on Information Forensics and Security 6 (4) (2011) 1223–1232.
- [11] V. Monga, B.L. Evans, Perceptual image hashing via feature points: performance evaluation and tradeoffs, IEEE Transactions on Image Processing 15 (11) (2006) 3453–3466.
- [12] A. Swaminathan, Y. Mao, M. Wu, Robust and secure image hashing, IEEE Transactions on Information Forensics and Security 1 (2) (2006) 215–230.
- [13] J.S. Seo, J. Haitisma, T. Kalker, C.D. Yeo, A robust image fingerprinting system using the Radon transform, Signal Processing: Image Communication 19 (2004) 325–339.
- [14] C.-Y. Lin, S.-F. Chang, A robust image authentication method distinguishing JPEG compression from malicious manipulation, IEEE Transactions on Circuits and Systems for Video Technology 11 (2) (2001) 153–168.
- [15] K. Maeno, Q. Sun, S.F. Chang, M. Suto, New Semi-Fragile Image, Authentication watermarking techniques using random bias and nonuniform quantization, IEEE Transactions on Multimedia 8 (1) (2006) 32–45.
- [16] D. Kundur, D. Hatzinakos, Digital watermarking for telltale tamper proofing and authentication, Proceedings of IEEE 87 (7) (1999) 1167–1180.
- [17] N. Ishihara, K. Abe, A semi-fragile watermarking scheme using weighted vote with sieve and emphasis for image authentication, IEICE Transactions on Fundamentals E90-A (5) (2007) 1045–1054.
- [18] K. Ding, C. He, L. Jiang, H. Wang, Wavelet-based semi-fragile watermarking with tamper detection, IEICE Transactions on Fundamentals E88-A (3) (2005) 787–790.

- [19] X. Qi, X. Xin, A quantization-based semi-fragile watermarking scheme for image content authentication, *J. Journal of Visual Communication and Image Representation* 22 (2011) 187–200.
- [20] A. Piva, F. Bartolini, R. Caldelli, Self recovery authentication of images in the DWT domain, *International Journal of Image and Graphics* 5 (1) (2005) 149–165.
- [21] R. Chamlawi, A. Khan, I. Usman, Authentication and recovery of images using multiple watermarks, *Computers and Electrical Engineering* 36 (2010) 578–584.
- [22] R. Chamlawi, A. Khan, Digital image authentication and recovery: employing integer transform based information embedding and extraction, *Information Sciences* 180 (2010) 4909–4928.
- [23] A. Phadikar, S.P. Maity, M. Mandal, Novel wavelet-based QIM data hiding technique for tamper detection and correction of digital images, *Journal of Visual Communication and Image Representation* 23 (2012) 454–466.
- [24] C. Cruz, R. Reyes, M. Nakano, H. Perez, Semi-fragile watermarking based content image authentication scheme, *Revista Facultad de Ingeniería* (56) (2010) 161–170.
- [25] Y. Hasan, A. Hassan, Tamper detection with self correction on hybrid spatial-DCT domain image authentication technique, *IEEE International Symposium on Signal Processing and Information Technology* (2007) 608–613.
- [26] Z. Zhao, A. Ho, H. Trehame, V. Pankajakshan, C. Culnane, W. Jiang, A novel semi-fragile image watermarking authentication and self-restoration technique using Slant transform, *International Conference on Intelligent Information Hiding and Multimedia Signal Processing* 1 (2007) 283–286.
- [27] J. Mendoza-Noriega, B. Kurkoski, M. Nakano-Miyatake, H. Perez-Meana, Halftoning-based self-embedding watermarking for image authentication and recovery, *International Midwest Symposium on Circuit and Systems* 1 (2010) 612–615.
- [28] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error measurement to structural similarity, *IEEE Transactions on Image Processing* 13 (2004) 1–14.
- [29] M. Mese, P. Vaidyanathan, Recent advances in digital halftoning and inverse halftoning methods, *IEEE Transactions on Circuits and Systems-I* 49 (6) (2002) 790–805.
- [30] R.C. Gonzalez, R.E. Woods, *Digital Image Processing*, 3rd edition, Addison-Wesley Publishing Company, Reading, 2006.
- [31] R. Nielsen, *Neurocomputing*, Addison-Wesley Publishing, Reading, 1991.
- [32] J. Bloom, I. Cox, T. Kalker, J. Linnartz, M. Miller, C. Traw, Copy protection for DVD video, *Proceedings of the IEEE* 87 (7) (1999) 1267–1276.