

ENHANCED PERCEPTUAL IMAGE AUTHENTICATION WITH TAMPER LOCALIZATION AND SELF-RESTORATION

Fang Liu¹, Hui Wang², Lee-Ming Cheng¹, Anthony T.S. Ho², Shujun Li²

¹Department of Electronic Engineering, City University of Hong Kong

²Department of Computing, University of Surrey

ABSTRACT

In this paper, an enhanced perceptual image authentication approach is proposed with extra ability of tamper localization and image self-restoration by combining perceptual hashing and digital watermarking technologies. Compared with other perceptual hashing schemes, this proposed approach could locate the maliciously tampered regions and further recover these regions to some extent. Another advantage of this approach is its robustness to various non-malicious image processing operations. This approach could provide better robustness to most content-based image processing operations such as JPEG compression and additive Gaussian noises than most existing semi-fragile watermarking methods. Experimental results demonstrated the high authentication accuracy rate to non-malicious and malicious image processing operations. Moreover, maliciously tampered regions could be correctly localized and the original images can be recovered with good quality as well.

1. INTRODUCTION

With the rapid development of Internet and multimedia technologies, digital images are widely created and distributed. Due to the rapid growth of multimedia editing tools, digital forgery and unauthorized utilization have consequently become more and more convenient. How to guarantee the authenticity and the integrity of image content becomes an urgent problem to be solved. As a result, perceptual image hash functions and semi-fragile watermarking methods are introduced for content-based authentication. This kind of authentication differs from traditional hashing-based strict authentication, which considers an image as non-authentic even when just one bit of the image has been manipulated. However, in many applications images will normally undergo some benign processing operations, such as lossy compression, image filtering or enhancement, thus limiting the usefulness of strict authentication. Therefore, perceptual image hash functions and semi-fragile watermarking methods are required for perceptual authentication nowadays. These approaches could survive non-malicious processing operations and could also detect some malicious content modifications. The main difference between non-malicious and malicious processing operations is that the former do not change the image content perceptually while the latter aim at altering the image content.

Perceptual image hashing has drawn a lot of attention in recent years due to its outstanding robustness against a variety of non-malicious operations with the capability to detect malicious manipulations. The state of the art techniques could be generally classified into four categories: statistics based approaches [1,2], relation based approaches [3,4], low-level feature extraction based approaches [5,6], and approaches based on preservation of coarse image representation [7,8]. In addition, semi-fragile watermarking schemes are constructed for content-based authentication as well. Many good semi-fragile watermarking techniques have been proposed with good visual quality of watermarked images and high accuracy of localizing tampered regions [9-12]. Some of the schemes also have the ability of self-restoration. Lin and Chang [9] proposed a DCT based semi-fragile watermarking scheme robust to JPEG compression. The authentication watermark utilizes the invariant relationship between two coefficients in two blocks before and after JPEG compression which is embedded into each pair of blocks. The extra recovery watermark is generated from Huffman coding of highly quantized DCT coefficients of a down-scaled edition of the original image. However, there is noticeable quality degradation of watermarked image with both authentication and recovery watermarks embedded. Ho *et al.* [10] proposed a semi-fragile watermarking scheme in the Pinned Sine Transform (PST) domain for localizing tampered regions with a fragile watermarking for content restoration. To improve the security of the watermarking system, Lee *et al.* [11] proposed a watermarking authentication scheme based on SVD (singular vector decomposition) of 4×4 blocks, which can prevent the VQ (vector quantization) attack and the histogram analysis attack. Tsai and Chien [12] proposed a novel semi-fragile image authentication and self-restoration watermarking scheme working in discrete wavelet transform (DWT) domain. The watermark is generated from the low-frequency DWT band, and embedded into the high-frequency DWT band considering some features of human visual system (HVS). This algorithm could resist JPEG compression and additive white Gaussian noise (AWGN) while being able to localize and recover tampered regions. While many semi-fragile watermarking schemes have been proposed, they are normally designed to be robust to some specific non-malicious image processing operations such as JPEG compression and AWGN. Other non-malicious processing

operations such as image filtering or image enhancement are often considered as malicious attacks, and therefore these schemes cannot provide proper authentication.

The existing problems of perceptual hashing and authentication watermarking schemes have motivated us to propose a new approach combining the two methods to achieve high authentication accuracy with the ability of tamper localization and self-restoration. The proposed hybrid scheme employs the research on perceptual hashing in [13] and on semi-fragile watermarking for content authentication and self-restoration in [14]. The proposed perceptual hashing scheme in [13] has very good robustness against most of the common non-malicious processing operations, such as image filtering, lossy compression, image contrast enhancement, and fragility to malicious tampering. In addition, those features used to generate the perceptual hash are key-dependent, so they can provide defense against many other attacks. However, as most perceptual hashing algorithms, this perceptual hashing method does not have the ability of localizing and recovering tampered regions. The semi-fragile watermarking method proposed in [14] has good performance on tamper localization and self-restoration, however, it is not robust to some non-malicious processing operations. The scheme may treat non-maliciously processed images as maliciously manipulated ones and report wrong authentication results. While the two schemes both have obvious drawbacks, a combination and optimization of them could produce a better hybrid authentication scheme with more functionalities (tamper localization and self-restoration).

The rest of this paper is structured as follows. In Section 2, a general introduction of our proposed authentication algorithm is given, which includes detail of pre-processing and the authentication procedure with tamper localization and self-restoration. Next, the experimental results are reported and analyzed in Section 3. Finally, the conclusions are given in Section 4.

2. PROPOSED AUTHENTICATION ALGORITHM

Since one of the objectives of our proposed algorithm is to support image self-restoration of detected tampered regions, the localization watermark and the recovery information both have to be embedded into the original image in a pre-processing procedure. Thus, in this section we first introduce the pre-processing procedure in detail. The authentication procedure including the tamper localization and image recovery at the detector side is then given afterwards.

2.1. Pre-processing of original images

The detailed pre-processing procedure is illustrated in Fig. 1. Firstly, the hash code and the recovery information are extracted from the original image. Then, the recovery

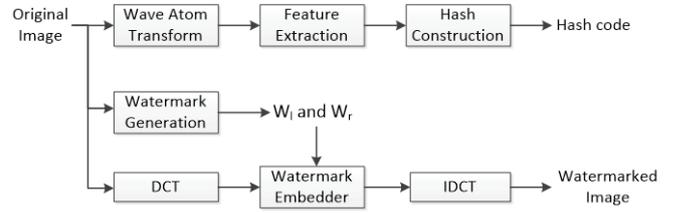


Fig. 1. Pre-processing procedure.

information and the localization watermark are embedded into the original image. In the following, we introduce the hash generation and the watermark embedding parts separately.

1). Hash generation: As discussed before, the proposed authentication scheme employs the perceptual hashing algorithm proposed in [13]. The hash extraction process can be briefly described as follows:

Wave atom transform is first applied to the original image and the coefficients in the third scale band are extracted as features to generate the hash code. Then, the summation of each block in the third scale band, which serves as the most important features in the perceptual hash, is calculated for quantization. To make the obtained features key-dependent, random permutations are applied based on chaotic Rényi map controlled by a secret key K_h , and the randomized summations are quantized using 4-bit gray code to generate the hash vector. Finally, the perceptual hash code H is constructed by concatenating all these quantized summation bits, which is further protected by XORing it with a pseudo-random secret bit sequence generated from the same chaotic Rényi map. For greater detail, please refer to [13].

2). Watermark embedding: There are two different watermarks W_l and W_r used for tamper localization and restoration, respectively.

To generate the localization watermark, the original image is firstly divided into 8×8 non-overlapping blocks. Then, the first 6 MSBs of the mean value of each block are extracted as a localization feature, and a secret key K_l is used to generate a 6-bit pseudo-random number for each block. According to this key, each block is linked with another block from where its localization feature is obtained. The pseudo-random number generator (PRNG) should be cryptographically strong, *e.g.* built on top of a strong cipher like AES. Thus, each block has a 6-bit localization feature of its corresponding block and a 6-bit random number. The localization watermark W_l is the result of XORing the two 6-bit numbers.

To generate the restoration watermark, each 8×8 block is further divided into four non-overlapping 4×4 sub-blocks. The mean pixel value of each sub-block is calculated and then normalized by multiplying a scaling factor c . The restoration watermark W_r is then generated by applying a

conditional-random mapping function with another key K_r for scrambling the order of all the image blocks with condition that the distance between the original block and the one after scrambling has to be longer than a certain pre-defined value. Here we set the distance limitation as a quarter of the image size. This is to make it harder for an attacker to remove restoration watermarks of manipulated blocks without significantly downgrading the overall visual quality of the image.

After both watermarks are generated, DCT is applied to each 8×8 block for embedding both localization and restoration watermarks. Like in [14], the 6 bits of W_l are embedded into 6 selected DCT coefficients in low frequency band by the normal binary QIM (Quantization Index Modulation) method [15] with step size T_l . Each normalized restoration watermark W_r is embedded using an improved edition of the one reported in [14] which can be described as follows. Let $r = \lfloor \frac{x}{T_2} + \frac{1}{2} \rfloor$ then

$$y = \begin{cases} W_r + rT_2 - \frac{T_2}{2}, & \text{if } r \text{ is odd,} \\ (T_2 - W_r) + rT_2 - \frac{T_2}{2}, & \text{if } r \text{ is even,} \end{cases} \quad (1)$$

where x is the selected DCT coefficient for embedding W_r , y is the watermarked DCT coefficient, and $T_2 > \max(W_r)$ is the quantization step size. After both watermarks are embedded, inverse DCT is applied to each block and the watermarked image is obtained.

2.2. Authentication with tamper localization and self-restoration

The proposed algorithm has the ability of authenticating an image based on its content, localizing the tampered region(s) if the image is identified as manipulated, and recovering the tampered region(s) to the level determined by the restoration watermark. Fig. 2 illustrates the detailed procedure of image authentication, tamper localization and image self-restoration. In the following, we explain the whole procedure as two main steps: image authentication, tamper localization and self-restoration.

1). Image authentication: As shown in Fig. 2, the verifier first computes the hash code H' from the tested image I' employing the same perceptual hashing method with the same secret key K_h as described in Section 2.1. Then, the hash code H' is compared with the original hash code H (which has been transmitted to the verifier via a different channel in advance) for the purpose of authentication. Here, the normalized Hamming distance (NHD) is used as the similarity metric. Assuming that the i -th values of H and H' are denoted as $H(i)$ and $H'(i)$, respectively and L is the length of the hash code, the NHD is defined as:

$$d(H, H') = 1/L \sum_{i=1}^L \delta(H(i), H'(i)), \quad (2)$$

where

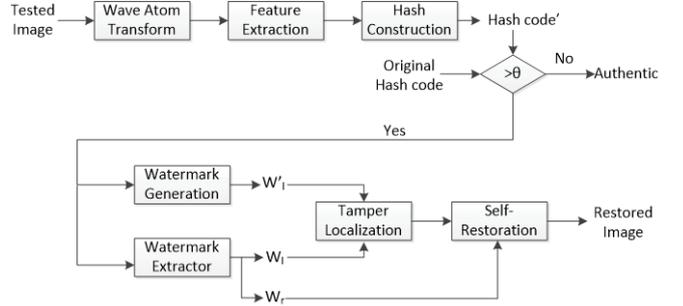


Fig. 2. Procedure for image authentication and self-restoration.

$$\delta(H(i), H'(i)) = \begin{cases} 0, & H(i) = H'(i), \\ 1, & H(i) \neq H'(i). \end{cases} \quad (3)$$

It should be noticed that this distance is expected to approach zero for images which have the same content and close to 0.5 for those images which have totally different content. Here, a threshold θ is presented to distinguish whether the two images I and I' have the same content. The following rules is employed to verify I' :

- (i) if $d(H, H') > \theta$, then $I \neq I'$,
- (ii) if $d(H, H') \leq \theta$, then $I \approx I'$.

The above rules mean that if the calculated NHD is larger than θ , the test image I' is considered as tampered; otherwise it will be considered as authentic. When an image is authenticated as tampered, the following procedure will be applied (as shown in Fig. 2).

2). Tamper localization and self-restoration: Firstly, the tested image is divided into 8×8 blocks. A reference watermark for localization is generated following the same method with same secret key K_l as described in Section 2.1. After that, DCT is applied to each block and the embedded watermark for localization is extracted by the QIM method. The extracted localization watermark is then compared with the reference localization watermark. If the difference between the two watermarks for one block is greater than a pre-defined threshold, then the block is indicated as possible-modified in advance. Because a modified block changes both the extracted watermark about the block itself and the reference watermark about its corresponding block, the block itself and its corresponding one will be both indicated as possible-modified. Thus, if the block is marked as possible-modified, we check whether its corresponding block is marked as well. If yes, then the block is finally indicated as modified. If not, the block is indicated as not modified. After all blocks in the whole image are processed, we apply the mathematical morphology erosion operation to remove isolated blocks which are considered as false positives and perform the morphological opening operation to remove isolated blocks which are considered as false negatives. Finally, the remaining blocks are marked and recorded to form a localization map.

Secondly, the same conditional-random mapping function is applied using the same key K_r following the same method as the one used in the embedding process to find the blocks with the restoration watermarks. If a block is marked as tampered, the restoration watermark will be then extracted from the four DCT coefficients in its corresponding block using the method defined as follow. Let $\tilde{r} = \lfloor \frac{\tilde{x}}{T_2} + \frac{1}{2} \rfloor$, then

$$\tilde{W}_r = \begin{cases} \text{mod}(\tilde{x} + \frac{T_2}{2}, T_2), & \text{if } \tilde{r} \text{ is odd,} \\ T_2 - \text{mod}(\tilde{x} + \frac{T_2}{2}, T_2), & \text{if } \tilde{r} \text{ is even,} \end{cases} \quad (4)$$

where \tilde{W}_r denotes the extracted watermark for restoration, \tilde{x} is the DCT coefficient for extraction \tilde{W}_r , and T_2 is the same quantization step size used in the embedding process. For all the tampered blocks, the four mean values of each 8×8 block's 4×4 sub-blocks are obtained after renormalizing the extracted watermarks.

Finally, the four mean values are used to estimate the DC coefficient and the first three AC coefficients of the tampered block following the method proposed in [14]. The recovered block then goes through inverse DCT to get pixel values. The whole image is obtained after all the tampered blocks are recovered.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, 20 grey-scale images of size 512×512 [16] were used as the original images, and the proposed approach was implemented with MATLAB. To balance the robustness and imperceptibility of our watermarking scheme, the locations for embedding localization watermark are chosen as (1 2), (2 1), (1 3), (2 2), (3 1), (3 2) by step size $T_1=12$. For restoration watermark, the scaling factor c equals to 0.13, its embedding locations are (1 4), (2 3), (3 3), (4 1), and the quantization step size is $T_2=12$. Since some information has to be embedded into original images, it is important that the embedding process does not lead to a failure of the perceptual hashing algorithm. To this end, we first tested the NHD between the watermarked images and the original images, and the mean NHD is around 0.020 for all the 20 tested images, which is low enough not to influence the hashing algorithm. The average PNSR of all watermarked images is 37.82 dB, which is considered acceptable by us.

3.1. Non-malicious processing operations

Some common non-malicious image processing operations were employed to test the robustness of our proposed approach. Table 1 shows the mean NHD values between the hashes of the 20 original images and those of their watermarked and non-maliciously processed images. As can be seen in the column "Non-malicious Attacks Only" in Table 1, the NHD values are very low for all non-malicious

Table 1. System performance under non-malicious attacks and combined attacks.

Non-malicious Processing Only				Combined 10% Malicious Attack with Non-malicious Attacks	
Non-malicious Attack	Parameter	Mean NHD	Authentication	Mean NHD	PSNR of Recovered Image (dB)
JPEG (QF)	85	0.021	Yes	0.315	31.338
	75	0.024	Yes	0.316	29.465
	65	0.029	Yes	0.316	27.640
	55	0.029	Yes	0.315	26.115
	45	0.034	Yes	0.320	16.209
	35	0.038	Yes	0.318	14.358
	25	0.052	Yes	0.317	14.207
Additive Gaussian White Noise (Variance)	15	0.086	Yes	0.315	11.895
	3	0.021	Yes	0.315	31.389
	6	0.023	Yes	0.315	29.628
	9	0.026	Yes	0.316	27.631
	12	0.027	Yes	0.314	23.563
Median Filtering (Size)	15	0.031	Yes	0.315	19.657
	18	0.034	Yes	0.316	17.422
	3	0.046	Yes	0.317	14.082
Contrast Change	5	0.086	Yes	0.324	12.051
	7	0.120	Yes	0.332	12.303
	+10%	0.020	Yes	0.317	19.707
	+20%	0.032	Yes	0.312	15.489
Low-pass Filtering (Variance, Window)	-10%	0.021	Yes	0.315	20.566
	-20%	0.020	Yes	0.317	16.204
	0.5,3	0.042	Yes	0.316	24.031
	0.5,5	0.042	Yes	0.316	23.994
	0.5,7	0.042	Yes	0.316	23.994
	1,3	0.093	Yes	0.322	14.723
Salt and Pepper Noise (Density)	1,5	0.112	Yes	0.323	13.167
	1,7	0.115	Yes	0.322	13.071
	0.001	0.031	Yes	0.315	28.280
Salt and Pepper Noise (Density)	0.002	0.037	Yes	0.313	22.605
	0.003	0.042	Yes	0.313	20.629
	0.05	0.137	Yes	0.324	9.249

processing operations. The threshold θ was set to 0.15 for the best performance of authentication accuracy. With this setting our approach was able to authenticate most of these processed images, achieving 97% accuracy rate for the non-malicious processing operations tested. Since these non-maliciously processed images can pass the authentication procedure, they do not need to go through the tamper localization and recovery procedure. However, if the semi-fragile watermarking scheme in [14] alone is used for authentication, the results will be much worse. For example, the authentication false rate (including both false positive and false negative) of method [14] is almost 100% under JPEG compression with QF=45.

3.2. Malicious attacks

"Copy and Paste" attacks with different tampered percentages were applied to watermarked images to test the

system’s performance against malicious attacks. Table 2 shows the mean NHD values under these malicious attacks, which are much larger than the mean NHD values under non-malicious processing operations shown in the previous subsection. Moreover, the authentication accuracy rate for these tampered images is 100%. After the tested images are identified as modified ones by the hashing based authentication procedure, the watermarking algorithm will be applied to localize and recover the tampered region(s). The mean localization false rates and the mean PSNR values of recovered images are also shown in Table 2. It can be observed that the false localization rate shows a rising tendency with the increase of the tampered percentage. In addition, the PSNR value of the recovered image declines with the increase of the localization false rate and the tampered percentage. If an image has been tampered with too much, the probability that the block stores the recovery watermark of any given tampered block will become large. In this case, the recovery information will be lost, and the affected tampered block cannot be restored any more, leading to severe distortion to the visual quality of the recovered image. In our experiments, the recovered image with a 30% “Copy and Paste” attack can still show the content of the original image with some level of visual detail, but there are approximately 5.78% falsely recovered blocks randomly distributed in the whole recovered image.

3.3. Malicious attacks combined with non-malicious processing

Since in common circumstances malicious attacks are always accompanied by some non-malicious processing, we investigated these situations as well. Here, 10% tampered images were used for the malicious attack part, and they were further processed by the non-malicious operations shown in Table 1. The NHD values after combined attacks are shown in the fifth column of Table 1. These attacked images all failed to pass the authentication phase due to the NHD values larger than the pre-set threshold $\theta=0.15$. The authentication accuracy for our proposed system is 100% for images gone through these combined attacks. However, the performance of self-restoration is not as good as that of authentication. The last column of Table 1 lists the mean PSNR values of recovered images under different combined attacks. As shown in Table 1, the proposed recovery scheme can achieve an acceptable level of recovered image quality when JPEG compression is up to QF=55, AWGN with variance up to 12, low-pass filtering with variance up to 0.5 and window up to 7, and salt and pepper noise with density up to 0.003.

Figure 3 gives an example of the entire test. It can be shown that the watermarked image in Fig. 3(b) is very similar to the original image Fig. 3(a). The localization map in Fig. 3(e) clearly reveals the location of the maliciously tampered region, which matches the region shown in Fig.

Table 2. System performance under malicious attacks, with examples of localization masks (black pixels show detected tampered regions) and recovered images.

Tampered Percentage	Mean NHD	Authentication	Localization False Rate	PSNR of Recovered Images(dB)
2%	0.209	No	0.00% 	36.733 
5%	0.278	No	0.00% 	35.746 
10%	0.311	No	0.07% 	33.659 
20%	0.353	No	0.67% 	27.518 
30%	0.365	No	5.78% 	21.487 
50%	0.413	No	27.08% 	14.963 

3(c). The recovered image in Fig. 3(g) shows good visual quality under a 10% malicious attack. Fig. 3(d) shows an image tampered with by a 10% “Copy and Paste” attack combined with JPEG compression with QF=55. Figures 3(f) and 3(h) show the localization map and the recovered image in this case, respectively. We observe that the tampered region is largely correctly marked with only a few false positive points. Although the visual quality of the recovered image is not as good as Fig. 3(g), the content in the tampered region is still discernible after the self-restoration process.

4. CONCLUSIONS

Content-based robust image authentication is very useful in today’s digital world. In this paper, an enhanced perceptual authentication method is proposed with the ability of tamper localization and image self-restoration. The proposed approach has better robustness against various kinds of non-malicious processing operations than common watermarking methods. It also has the additional ability of tampered localization and self-restoration comparing with common perceptual hashing schemes. Our experimental results have demonstrated that the proposed approach maintains very high authentication accuracy rate to non-malicious processing operations and malicious attacks and their combinations as well. Moreover, the tampered regions could be correctly localized and the recovered images show good visual quality as well.

5. REFERENCES

- [1] M. Schneider and S. F. Chang, "A robust content based digital signature for image authentication," in Proc. IEEE Conf. on Image Processing (ICIP '96), vol. 3, pp. 227-230, 1996.
- [2] R. Venkatesan, S. M. Koon, M. H. Jakubowski, and P. Moulin, "Robust image hashing," in Proc. IEEE Conf. on Image Processing (ICIP '00), pp. 664-666, 2000.
- [3] C. Y. Lin and S. F. Chang, "A robust image authentication system distinguishing JPEG compression from malicious manipulation," Proc. IEEE Trans. Circuits Syst. Video Tech., vol. 11, no. 2, pp. 153-168, 2001.
- [4] C. S. Lu and H. Y. M. Liao, "Structural digital signature for image authentication," IEEE Transactions on Multimedia, pp. 161-173, June 2003.
- [5] V. Monga and B. L. Evans, "Perceptual image hashing via feature points: performance evaluation and tradeoffs," IEEE Trans. Image Process., vol. 15, no. 11, pp. 3452-3465, 2006.
- [6] F. Khelifi and J. M. Jiang, "Perceptual image hashing based on virtual watermark detection," IEEE Trans. Image Process., vol. 19, no. 4, pp. 981-994, 2010.
- [7] A. Swaminathan, Y. Mao, and M. Wu, "Robust and secure image hashing," IEEE Trans. Inf. Forensics Secur., vol. 1, no. 2, pp. 215-230, 2006.
- [8] V. Monga and M.K. Mihcak, "Robust and secure image hashing via non-negative matrix factorizations," IEEE Trans. Inf. Forensics Secur., vol. 2, no. 3, pp. 376-390, 2007.
- [9] C. Y. Lin and S. F. Chang, "Semi-fragile Watermarking for Authenticating JPEG Visual Content," Proc. SPIE, Security and Watermarking of Multimedia Contents, pp.140-151, 2000.
- [10] A. T. S. Ho, X. Zhu, and Y. Guan, "Image content authentication using pinned sine transform," EURASIP J. Appl. Signal Process., vol. 2004, pp. 2174-2184, 2004.
- [11] S. Lee, D. Jang, and C. D. Yoo, "An SVD-based watermarking method for image content authentication with improved security," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '05), vol. 2, pp. 525-528, 2005.
- [12] M.-J. Tsai and C.-C. Chien, "Authentication and recovery for wavelet-based semifragile watermarking," Opt. Eng., vol. 47, no.6, art. no. 067005, 2008.
- [13] F. Liu, L. M. Cheng, H. Y. Leung, and Q. K. Fu, "Wave atom transform generated strong image hashing scheme," Opt. Comm.s, vol. 285, no. 24, pp. 5008-5018, 2012.
- [14] H. Wang, A. T. S. Ho, and X. Zhao, "A novel fast self-restoration semi-fragile watermarking algorithm for image content authentication resistant to JPEG compression," in Digital Forensics and Watermarking:

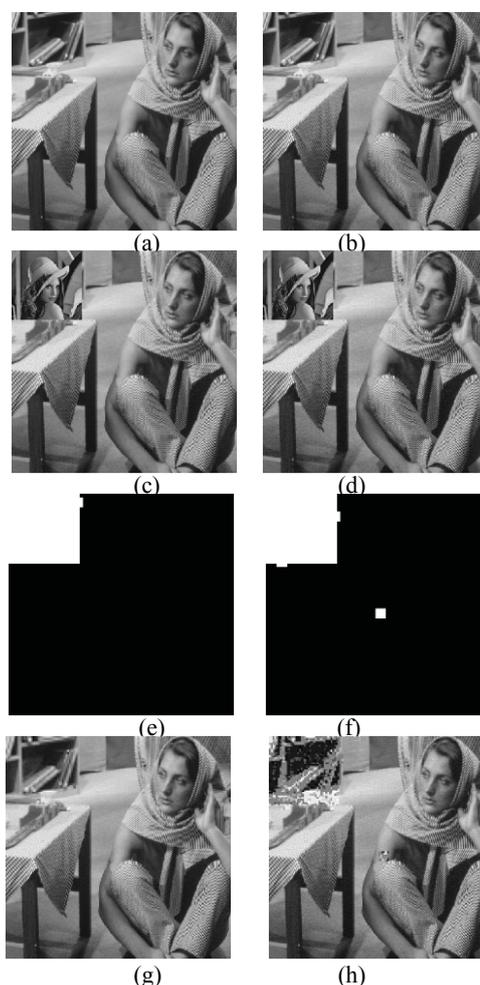


Fig. 3. (a) The original image, (b) the watermarked image, (c) an image tampered with by 10% "Copy and Paste" attack, (d) an image tampered with by 10% "Copy and Paste" attack followed by JPEG compression with QF=55, (e) the localization map under a 10% "Copy and Paste" attack, (f) the localization map under the combined attack, (g) the recovered image under a 10% "Copy and Paste" attack, (h) the recovered image under combined attack.

10th International Workshop, IWDW 2011, Atlantic City, NY, October 23-26, 2011, Revised Selected Papers, Lecture Notes in Computer Science, vol. 7128, pp. 72-85, 2011.

- [15] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," IEEE Trans. Inf. Theory, vol. 47, no. 4, pp. 1423-1443, 2001.
- [16] Computer Vision Group, University of Granada, Test Images: Miscellaneous gray level images (512×512), Available: <http://decsai.ugr.es/cvg/dbimagenes/g512.php>