SEMI-FRAGILE WATERMARKING FOR STILL IMAGES AUTHENTICATION AND CONTENT RECOVERY

A. Piva[*], *R. Caldelli*[+], *F. Bartolini*[+], *M. Barni*[×]

[*] National Inter-university Consortium for Telecommunications (CNIT)
[+] Dept. of Electronics and Telecommunications, University of Florence
Via S. Marta 3, 50139 - Florence, Italy
[×] Dept. of Information Engineering, University of Siena
Via Roma 56, 53100 - Siena, Italy

ABSTRACT

A self recovery authentication algorithm is presented which hides an image digest into some of the DWT subbands of the to-be-authenticated image. The digest is computed through a properly modified version of JPEG coding operating at very high compression ratios. Particular care is given to ensure robustness against innocuous manipulations, and to prevent forgery attempts. Security aspects are also discussed in great detail. Experimental results demonstrate the good performance of the proposed system.

1. INTRODUCTION

During the last decade an increasing attention has been paid to data-hiding based authentication of still images and video. According to the data-hiding approach, the authenticating information is hidden within the to-be-protected digital content [1, 2, 3]. As opposed to conventional cryptographic tools, watermarking technology allow to associate the authenticating information to the host data, in such a way that this information can be recovered even if the host data have suffered some (moderate) transformations, e.g. high quality lossy compression. The basic idea of watermarking based authentication is then to compute a kind of digest of the digital document, and to hide it inside the document itself. For authenticity verification, it only needs to recover the embedded digest from the to be checked document, and to compare it with the digest computed for the to be checked document. Of course, the digest should be transparent at least to watermark embedding, i.e. it should produce the same result also after the watermark has been embedded. A simple way to achieve this transparency is to use, as a digest, a lossy compressed version of the document. This choice permits also to satisfy other important requirements of multimedia data authentication applications, in particular it allows to authenticate the content (by visual inspection), to be transparent to the manipulations that do not affect the content, and to localize the modifications.

The class of watermarking based authentication algorithms that use, as digest, a compressed version of the document itself are usually referred to as self recovery techniques, because they also allow to obtain an estimate of the original content.

In this paper a new, very simple, self recovery authentication technique that hides an image digest into some DWT subbands of the to be authenticated image is described. The new scheme is based on a previous work by Campisi et al. [4], where the color information of an image is hidden in the discrete wavelet domain (DWT), for improving compression efficiency. The proposed technique is especially designed for video surveillance and/or remote sensing applications, and aims at detecting possible malevolent object manipulations undergone by the image by means of a self recovery processing. This valuable characteristic has to be maintained also when an image is processed through an usual and friendly transformation like JPEG compression.

By reducing the compression ratio during the digest computing step, the proposed approach allows to achieve a graceful degradation of the digest with respect to the amount of global manipulations suffered by the authenticated image. This is obtained by avoiding the lossless entropy coding step of common image compression algorithms. Particular care is given to the discussion of security aspects, by showing that it is extremely difficult for an attacker to create a forged image that is judged as authentic by the verification process.

2. DIGEST EMBEDDING

The data embedding part of the proposed scheme is sketched in figure 1. Given a $N \times N$ image, after applying a 1-level DWT, the two horizontal and vertical details subbands are further DWT decomposed. The full-frame DCT (Discrete Cosine Transform) of the low-pass version of the original image, with size $N/2 \times N/2$, is computed. The full frame DCT coefficients are then scaled down to decrease their obtrusiveness when they will be hidden: to this aim the JPEG quantization matrix is used (in practice each scaling coefficient of the matrix is applied to a block of coefficients of the full frame DCT). The first M lowest frequency coefficients are selected and stored in a vector $\mathbf{c} = (c_1, c_2 \dots c_M)$ (we usually set $M = N^2/32$). The DC coefficient is discarded because it has a too high energy: as a matter of fact, we are not interested in authenticating the

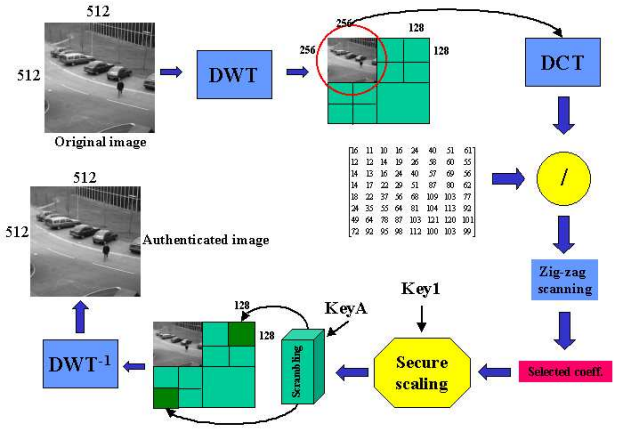

Fig. 1. Sketch of the embedding procedure.

mean grey level of the image.

Following ahead the scheme depicted in Figure 1, these coefficients are further scaled (the need for such step will be clarified in subsection 2.1 and 2.2). Each DCT coefficient can be hidden in each sub-band more than once, thus ensuring a certain degree of robustness. For a $N \times N$ original image, each one of these sub-bands has a size of $N/4 \times N/4$: since two subbands are selected for embedding, we have $N^2/8$ available positions, that is four times the number M of DCT coefficients to be casted. Due to this fact, the sequence of DCT coefficients is quadruplicated obtaining a new vector $\mathbf{p} = (c_1, c_2 \dots c_M, c_1, c_2 \dots c_M, c_1, c_2 \dots c_M, c_1, c_2 \dots c_M)$.

The DCT coefficients are substituted to the DWT coefficients in the two detail sub-bands highlighted in dark grey in Figure 1. Before the replacement, a scrambling process, depending on a secret key ($KeyA$), is applied to the vector $\mathbf{p}$ obtaining a new vector $\mathbf{p_{scrambled}}$, in such a way that statistically the four replicas of each DCT coefficient will occupy different locations in the two sub-bands. This is important because, if a manipulation occurs we can be quite confident that not all the replicas of a given coefficient will be removed by the attack.

The two chosen sub-bands have been selected because they grant a good trade-off between invisibility and robustness of the hidden information. Finally inverse DWT is applied and the authenticated image is obtained. The original image and the authenticated one appear very similar from a quality point of view and a PSNR of about 36 dB has been obtained with different test images.

2.1. Visibility issues

Before the original DWT coefficients are replaced by the watermark, these values are scrambled with a key-dependent rule to introduce a degree of robustness and security. This scrambling operation is basically an internal permutation that moves coefficients in different positions with respect to those they had previously. In this way coefficients of high amplitude may fall close to low amplitude coefficients. This can result in some wavelet values being much higher than the other belonging to the same neighborhood, causing an unpleasant quality degradation resulting in some small false contours (artifacts) all around the image. To avoid this undesired effect a further scaling operation has been introduced before scrambling. Each DCT coefficient is processed according to the following rule:

(a) (b)

Fig. 2. Visibility comparison: original image (a) and authenticated image without de-emphasis(b). The watermark has been artificially increased to better illustrate the artifacts.

$$c_{scaled}(i) = c(i) \cdot \alpha \cdot \ln(i + 2 + rand(i)) \qquad (1)$$

where $c(i)$ indicates the DCT coefficient in position i within the zig-zag scan and $c_{scaled}(i)$ is the corresponding scaled coefficient; α is a strength factor (usually slightly higher than 1) which is set on the basis of the image final quality, and $rand$ is a shift parameter (ranging between -0.5 and 0.5) generated pseudo-randomly by means of a PRNG (Pseudo-Random Number Generator) initialized with a secret key $Key1$ (the purpose of this parameter will be more deeply discussed in section 2.2). In practice a sort of emphasizing pre-process is applied, to enhance the high frequency part of the spectrum with respect to the low frequency components. This shrewdness allows to get rid of the problem pointed out before, because all DCT coefficients are now weighed with a logarithmic function that basically depends on their position in the zig-zag scanning.

2.2. Security issues

We now consider why a further security step, as hinted in Equation 1, has been inserted and why a scrambling is not sufficient to grant a complete safety against intentional attacks. Let us assume that a potential hacker perfectly knows how the algorithm works, and that he also knows the scrambling key ($KeyA$): under this circumstances, he can modify the authenticated image by inserting some wanted changes and create a seemingly authentic image by reintroducing in the right DWT sub-bands the informative data related to the forged image. Actually the hacker does not know the key, and thus he is inhibited from operating as above . However, if he is able to crack the scrambling rule he can pour his data in the correct way. Cracking the scrambling rule can be computationally intensive but not infeasible, it only needs that the attacker

compute the digest, i.e. the selected DCT coefficients of the low pass band of the 1-level DWT, and for each coefficient find where it has been placed in the two detail bands. The insertion of an additional secret key-dependent ($Key1$) random scaling (Equation 1) makes the estimation of the scrambling rule unfeasible. In fact he can not understand where the DCT coefficients are relocated after scrambling because the computed (and selected) DCT coefficients ($c(i)$) are different from those that are actually embedded ($c_{scaled}(i)$).

3. INTEGRITY VERIFICATION

In the integrity verification phase the DWT of the to-be-checked $N \times N$ image is computed and the two sub-bands, supposed to contain informative data, are selected (see Figure 3). These data are reversed into a vector $\mathbf{p'_{scrambled}}$,

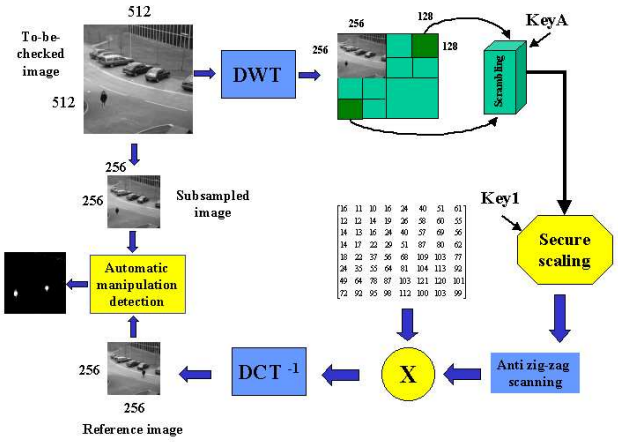

Fig. 3. Sketch of the integrity verification process.

which is inversely scrambled by means of the secret key $KeyA$, thus resulting in a sequence $\mathbf{p'}$. An estimate of the hidden DCT coefficients is then obtained by averaging the four copies of each extracted coefficient. After that a unique set of authentication data $\mathbf{c_{extracted}}$ (i.e. M coefficients) is obtained. By knowing the private scaling key ($Key1$) it is possible to correctly invert the scaling operation performed during the authentication phase (equation (1)), i.e.:

$$c_{reconstructed}(i) = c_{extracted}(i) \cdot \frac{1}{\alpha} \cdot \frac{1}{\ln(i + 2 + rand(i))} \quad (2)$$

The inverse scaled coefficients are then put in the correct positions, in such a way to obtain an estimate of the DCT of the reference image (missing elements are set to zero, and a DC coefficient with value 128 is reinserted). These values are weighed back with the JPEG quantization matrix, and then the inverse-DCT is applied to obtain an approximation of the original reference image. The quality of this extracted image (having size $N/2 \times N/2$) is very satisfactory and permits to make a good comparison with the checked for authenticity verification.

An automatic system for the detection of manipulations has also been implemented in the following way: the to-be-checked image is resized to $N/2 \times N/2$, and a pixel-wise absolute difference between this subsampled image and the extracted image digest is computed. The difference is thresholded to obtain a binary image, where the white pixels indicate a local difference between the to-be-checked and the extracted image. Such a system has to be intended as a primary step to alarm the user, in charge of the integrity verification, to pay attention that a possible manipulation could be occurred.

4. EXPERIMENTAL RESULTS

The proposed algorithm has been tested with various images for different types of use, in particular in this section experimental results related to two specific application fields like video surveillance and remote sensing are presented.

In Figure 4 the reference image extracted for authentication purpose is presented in three different situations. In the first case (Figure 4 (a)) the image digest that is recovered through the detection process, as explained in Section 3, when the authenticated image has not undergone attacks is presented. This image presents the same characteristics of the original one and its quality is perfectly sufficient to well distinguish scene objects and to understand through a comparison if the checked image is authentic or not. In the second case (b) the image recovered when the authenticated image has been JPEG compressed with a quality factor of 80% is reported. In this circumstance image sharpness is slightly poorer with respect to the case depicted in Figure 4 (a), and a sort of noise is superimposed to the image. Notwithstanding this undesired effect, the image is still good to satisfy application purposes, like to determine if something has been changed in the video scene and obviously to estimate which was the original image. This result is very important because it shows that the proposed authentication system is able to offer a degree of robustness against JPEG compression, that can not be considered an intentional modification invalidating image authenticity, but only an usual processing step adopted for data storage and/or data transmission. In Table 1 the values of PSNR (Peak Signal to Noise Ratio) of the image digest recovered after the authenticated image has been JPEG compressed with respect to the image digest extracted when the authenticated image has undergone no compression are given. Other experimental tests have been carried out to evaluate the performance of the proposed approach when modifications are brought to an image to alter its effective appearance and especially its content. Here a possible application of the proposed algorithm to remote sensing is presented. In Figure 5 (a) (the authenticated, and JPEG compressed image) a view of Turin is depicted, in particular attention is focused on the famous building *Mole Antonelliana* (in the white circle). After that a further intentional modification is applied to this image by erasing the *Mole Antonelliana* and substituting it with the circular building (see Figure 5 (b)). In Figure

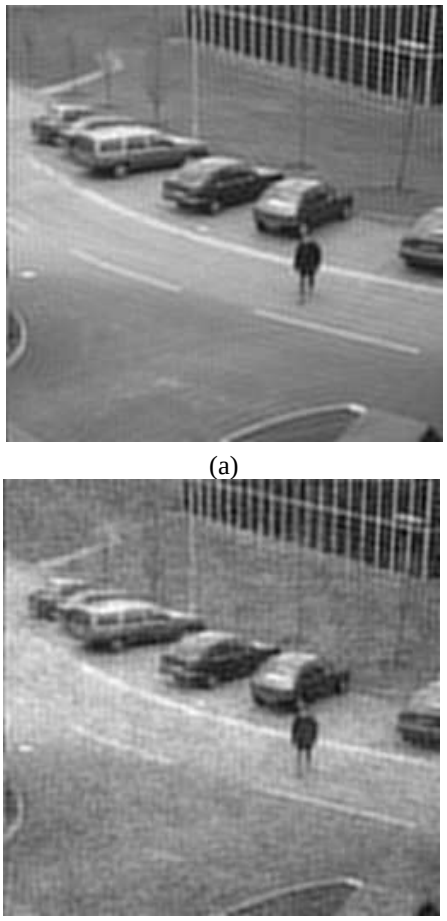

Fig. 4. Reference image extraction: no attacks (a) extraction after 80% JPEG compression (b).

PSNR vs JPEG quality factor		
PSNR	Quality factor	Compression ratio
49.3	100	1.8
30.3	80	7.1
25.7	60	11.3
24.2	50	13.4

Table 1. PSNR of the extracted digest image after JPEG compression, with respect to the digest extracted from a non-compressed image.

5 (c) the extracted reference image is shown; the *Mole Antonelliana* is still visible in its original position even though heavy changes occurred; finally, in Figure 5 (d) the automatic manipulations detection image is shown, where it is possible to clearly identify the manipulated area.

5. CONCLUSIONS

A secure self-recovery authentication scheme for digital images has been proposed. The scheme hides an image digest into some DWT sub-bands of the to be authenticated image. Experimental results showed that the proposed authentication scheme, although very simple, is able

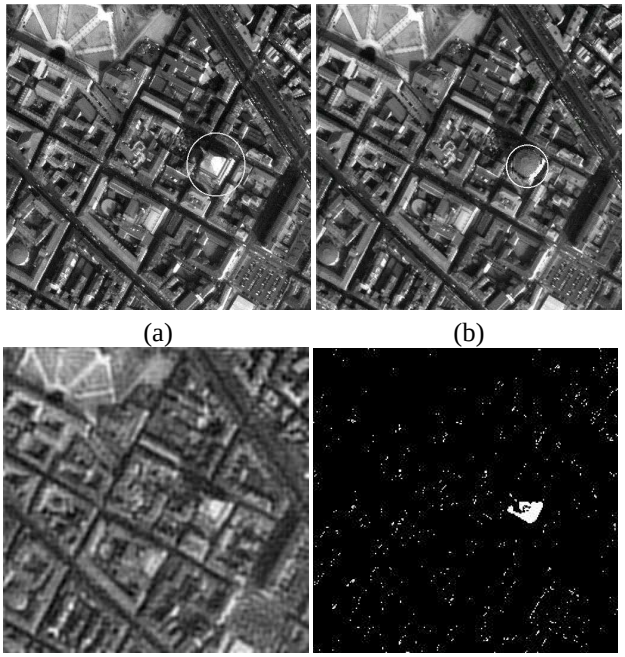

Fig. 5. Remote sensing: authenticated image after JPEG compression with a quality factor of 80% (a); manipulated image (b), extracted reference image (c) automatic manipulations detection (d).

to identify malicious content manipulations, and to produce a good estimate of the original (authentic) content. Moreover, an automatic system for the detection of manipulations has been implemented, giving as final result a binary image, where the white pixels indicate local differences between the to-be-checked and the extracted image.

6. REFERENCES

[1] C. Rey and J.L. Dugelay, "A Survey of Watermarking Algorithms for Image Authentication," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 6, p. 613–621, June 2002.

[2] M. Wu and B. Liu, "Multimedia Data Hiding," *Springer Verlag*, ISBN 0387954260, October 2002.

[3] Ee-Chien Chang, M.S. Kankanhalli, X. Guan, Z.Y. Huang and Y.H. Wu, "Robust Image Authentication Using Content-based Compression," *ACM Multimedia Systems Journal*, vol. 9, no. 2, p. 121–130, 2003.

[4] P. Campisi, D. Kundur, D. Hatzinakos, and A. Neri, "Compressive data hiding: An unconventional approach for improved color image coding," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 2, p. 152–163, February 2002.