IMAGE AUTHENTICATION AND TAMPER PROOFING USING MATHEMATICAL MORPHOLOGY

Anastasios Tefas and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 540 06, GREECE, {tefas,pitas}@zeus.csd.auth.gr

ABSTRACT

A novel method for image authentication and tamper proofing is proposed. A binary watermark is embedded in a grayscale or a color host image. The method succeeds in detecting alterations made in a watermarked image. The proposed method is robust against high quality lossy image compression. It provides the user not only with a measure for the authenticity of the test image but also with an image map that highlights the unaltered regions of the image when selective tampering has been made. Mathematical morphology techniques are also developed for the accurate detection of alterations in fine image details.

1. INTRODUCTION

In the last decades massive digitization of multimedia data such as photographs, paintings, speech, music, video, documents etc., became very popular. New techniques for the representation, storage and distribution of digital multimedia information have been developed. At the same time, the amount of digital data that is distributed through international communication networks has increased rapidly . In such an environment original digital products can be easily copied, tampered and transmitted back in the network. Consequently, the design of robust techniques for copyright protection and content verification of multimedia data became necessary.

When content verification of multimedia data is the objective, research is mainly focused in the development of *fragile watermarks*. By this term we mean watermarks that have the property of being distorted when the host media is somehow tampered. The design of algorithms that generate imperceptible fragile watermarks and detect any alterations of the host signal is the objective in content verification systems. In this paper, the problem of image authentication is treated.

Several approaches have been recently proposed to address the issue of tamper proofing. The development of a "trustworthy digital camera" is proposed in [1]. A digital image is captured from a camera and then it is passed through a hash function. The output of the hash function is encrypted by the photographer's private key and a separate authentication signal is created. In order to ensure authentication of the image the encrypted signal is decrypted by the photographer's public key and the hashed version of the original image is compared to that of the received image. In the same category belong the methods that require a separate header for image authentication [2].

A compression tolerant method for image authentication is proposed in [3]. The proposed scheme is based on the extraction of feature points that are almost unaffected by lossy compression. The major drawbacks of this method are the need of a separate header for storing the digital signature and the low accuracy in the detection of tampered regions. An authentication method that gives a distortion measurement instead of a binary decision on image authenticity is proposed in [4]. It does not need a separate signature file or header for image authentication but it can not detect the regions of the image that are authentic if selective modifications to fine details of the image have been made. A method for watermarking in the wavelet transform domain is proposed for authentication in [5]. The issue of detecting the tampered regions in an image is not addressed. A method for image authentication by changing the least significant bit (LSB) in an image is proposed in [6]. The method can detect alterations that are made in several image regions. However, it is not robust against lossy compression that does not reduce significantly the image quality.

In this paper a novel technique for image authentication is proposed. It is based on an established watermarking technique [7, 8]. The novelty of the method is based on the high accuracy in detecting tampered regions and alterations in fine image details. It provides the receiver with a measure and not only with a binary decision about the image authenticity. A new technique based on mathematical morphology for the detection of changes in small details of the image has also been developed.

2. WATERMARK GENERATION AND EMBEDDING

The watermark generation procedure aims at generating a three-valued watermark $w(\mathbf{x}) \in \{0, 1, 2\}$, from an image $f(\mathbf{x})$, given a digital key k. The watermark is a random sequence of three-valued data, thus, it is usually produced by a pseudorandom number generator. An alternative to random number generators is to use chaotic mixing systems that provide an additional level of security [9].

After the watermark generation we proceed to the watermark embedding by altering the pixels of the original (host) image according to the following formula:

$$f_w(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } w(\mathbf{x}) = 0 \\ g_1(f(\mathbf{x}), \mathcal{N}(\mathbf{x})) & \text{if } w(\mathbf{x}) = 1 \\ g_2(f(\mathbf{x}), \mathcal{N}(\mathbf{x})) & \text{if } w(\mathbf{x}) = 2 \end{cases} \quad (1)$$

where g_1, g_2 are suitably designed functions based on $\mathbf{x}$ and $\mathcal{N}(\mathbf{x})$ denotes a function that depends on the neighborhood of $\mathbf{x}$. The functions g_1, g_2 are called *embedding functions* and they are selected so as to define an inverse detection function $G(f_w(\mathbf{x}), \mathcal{N}(\mathbf{x}))$. The detection function, when applied to the watermarked image $f_w(\mathbf{x})$, gives the watermark $w(\mathbf{x})$:

$$G(f_w(\mathbf{x}), \mathcal{N}(\mathbf{x})) = w(\mathbf{x}) \quad (2)$$

Obviously several embedding functions and the appropriate detection function can be designed giving different watermarking schemes. The function that is used in our method is based on a superposition of real quantities in the pixels which are going to be signed:

$$g_1(f(\mathbf{x}), \mathcal{N}(\mathbf{x})) = \mathcal{N}(\mathbf{x}) \oplus \alpha_1 f(\mathbf{x}) \quad (3)$$

$$g_2(f(\mathbf{x}), \mathcal{N}(\mathbf{x})) = \mathcal{N}(\mathbf{x}) \oplus \alpha_2 f(\mathbf{x}) \quad (4)$$

where α_1, α_2 are suitably chosen constants and $\mathcal{N}(\mathbf{x})$ is a local neighborhood operation of the pixels around $\mathbf{x}$. The sign of α_1, α_2

is used for the detection function and its value determines the watermark power.

The size of the region around $\mathbf{x}$ used for the calculation of $\mathcal{N}(\mathbf{x})$ is important for the watermarking procedure. Moreover, the number of pixels used for the calculation of $\mathcal{N}(\mathbf{x})$ determines the upper bound of the number of watermarked pixels in an image. If a pixel to be signed is contained in the neighboring region of another signed pixel, the related watermark detection may be affected by the neighboring pixel alterations, thus resulting in a false detection. To avoid such problems we should use small watermark embedding neighborhoods (i.e., of size 3×3). The maximum number of pixels that can be signed in a host image of dimensions $N \times N$ by using blocks of $(2r+1) \times (2r+1)$ for calculating $\mathcal{N}(\mathbf{x})$ is given by:

$$l = \frac{N^2}{(r+1)^2} \tag{5}$$

3. WATERMARK DETECTION

In the detection procedure we generate first the watermark $w(\mathbf{x})$ according to the watermark key k. The detection function resulting from (3,4) is defined by:

$$G(f_w(\mathbf{x}), \mathcal{N}(\mathbf{x})) = \begin{cases} 1 & \text{if } f_w(\mathbf{x}) - \mathcal{N}(\mathbf{x}) > 0 \\ 2 & \text{if } f_w(\mathbf{x}) - \mathcal{N}(\mathbf{x}) < 0 \end{cases} \tag{6}$$

The detection function is valid if $\alpha_1 > 0$ and $\alpha_2 < 0$. This fact should be accounted for the design of the embedding functions. By employing the detection function in the watermarked image a bi-valued detection image $d(\mathbf{x})$ is produced:

$$d(\mathbf{x}) = G(f_w(\mathbf{x}), \mathcal{N}(\mathbf{x})) \tag{7}$$

Based on the watermark $w(\mathbf{x})$ and the detection image $d(\mathbf{x})$, we can decide on or against image authenticity. We can also detect changes made at certain image regions. The detection is based on the pixel to pixel comparison for the nonzero pixels in $w(\mathbf{x})$. By comparing the watermark $w(\mathbf{x})$ and the detection image $d(\mathbf{x})$ we form the false detection image:

$$e_w(\mathbf{x}) = \begin{cases} 1 & \text{if } w(\mathbf{x}) \neq 0 \text{ and } w(\mathbf{x}) \neq d(\mathbf{x}) \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The false detection image has value 1 (white pixels) if a watermarked pixel is falsely detected and 0 otherwise. The detection ratio is given by the ratio of the correctly detected pixels to the sum of the watermarked pixels in the image.

$$D_w = 1 - \frac{\text{card}\{e_w(\mathbf{x})\}}{\text{card}\{w(\mathbf{x})\}} \tag{9}$$

The embedding functions are designed in such way, so as the probability p of a pixel to be detected as signed with g_1 or g_2, for an unwatermarked image, to be 0.5. Thus, the detection ratio in an unwatermaked image forms a binomial distribution. The cumulative distribution function (*cdf*) of the detection ratio is given by:

$$P_n = p^m \sum_{i=0}^{n} \frac{m!}{i!(m-i)!} \tag{10}$$

where m is the total number of the watermarked pixels and n is the number of correctly detected watermarked pixels.

The decision about the image authenticity is taken by comparing the watermark detection ratio of the image with a predefined threshold T. The value of the threshold determines the minimum acceptable level of watermark detection and is user defined. That is, if the receiver of the image demands high authenticity the threshold should be very high. For complete authenticity (test image identical to the reference one) the watermark detection ratio should have value 1. If the user considers that high quality image compression does not destroy image authenticity, the threshold for the watermark detection ratio can be lower than 1. In that case, the watermark detection ratio can be above the threshold although the host image is tampered somehow and a second level examination has to be done. The second authentication test aims at detecting if the drop of the detection ratio is caused due to high quality lossy image compression or due to alterations in fine image details. A region of a tampered watermarked image that has watermark detection ratio 0.97 is illustrated in Figure 2b. If only the watermark detection ratio was used for image authenticity, this image would be considered authentic.

4. CONTENT TAMPERING DETECTION

A usual way for tampering an image is to manipulate certain image regions in order to hide something or to create a new image with different content. It is important for the user that receives this image to know if the entire image or some of its regions have been tampered. The method presented in the previous section succeeds in giving a measure for the image authenticity and can also provide the user with a binary false detection image $e(\mathbf{x})$. This image is produced by the watermark detection at each watermarked pixel according to (8). If an image region (e.g., an object or a person) is tampered then we expect this region to produce many false detection errors. Thus, the watermark detection ratio in this region should follow (10) and the false detection image should have dense white pixels (i.e., $e_w(\mathbf{x}) = 1$). According to (10) approximately half of the watermarked pixels will be detected falsely.

However, if the image is compressed with high quality, then the watermark detection ratio is high and the detection errors (i.e., white pixels in the false detection image) are spread all over the image domain. Automatic discrimination between the regions that are tampered (untrustworthy) and the regions that are compressed with high quality (trustworthy) is addressed in the section. The objective is the development of an automatic method that creates compact objects (regions) that correspond to the tampered regions while clears all the other falsely detected watermarked pixels that are produced due to high quality image compression.

To do so, we propose the use of mathematical morphology operations. That is because morphological operations with certain structuring elements are adequate for removing noise or for creating compact objects. Given a binary image $f(\mathbf{x}) : \mathcal{D} \subseteq \mathcal{Z}^2 \rightarrow \mathcal{U} = \{0, 1\}$ and a structuring element $g(\mathbf{x}) : \mathcal{G} \subseteq \mathcal{Z}^2 \rightarrow \mathcal{U}$, the *binary dilation* of the image $f(\mathbf{x})$ by $g(\mathbf{x})$ is defined as $f \oplus g$. The complementary operation, the *binary erosion*, is defined as $f \ominus g$ [10].

The size of the structuring element used in this operations determines the size of the objects that will be removed from the image or the size of the gaps that will be filled. We propose the use of a sequence of morphological operations in the false detection image $e(\mathbf{x})$ in order to form the image $t(\mathbf{x})$ which highlights the not authentic image regions.

$$t(\mathbf{x}) = (e \oplus v_1 \ominus v_2 \oplus v_3)(\mathbf{x}) \tag{11}$$

The first dilation aims at creating a compact object in the tampered regions. The erosion aims at removing the falsely detected watermarked pixels that are produced due to high compression and the final dilation restores the original size of the tampered regions. In order to find the size of the structuring elements used in these operations, we consider that after a dilation we want the region among the falsely detected pixels ($e_w(\mathbf{x}) = 1$) that come from a tampered region to be filled completely. Thus, the size of the structuring element should be at least the size of the minimum distance between two neighboring white pixels in the false detection image. To do so, we can decompose the two dimensional structuring elements in two one dimensional structuring elements. Thus, the objective is to find the minimum distance between two watermarked pixels in one dimension. We suppose that n watermarked pixels $W_1, W_2, \ldots, W_n$ are randomly and independently selected on the interval $[0, N)$. We define the random variable:

$$d = \min_j \mid W_i - W_j \mid \quad \text{for some fixed} \quad i \tag{12}$$

Then the objective is to estimate the mean value of d [11]. To do so we calculate the cdf of d. The cdf of d is given by:

$$F(d) = \begin{cases} 1 - \frac{2}{n}\left(1 - A\right)^n + \frac{2-n}{n}\left(1 - 2A\right)^n & 0 \le d \le \frac{N}{2} \\ 1 - \frac{2}{n}\left(1 - A\right)^n & \frac{N}{2} \le d \le N \end{cases} \tag{13}$$

where $A = \frac{d}{N}$ and the mean value of d is given by:

$$\mu_d = \frac{n+2}{2n(n+1)} N \tag{14}$$

Having calculate the mean value of the minimum distance between two watermarked pixels in one dimension we first apply a dilation with the structuring element v_1 of size $s \times s$ in order to fill the regions that are not considered authentic. In these areas we have many falsely detected pixels. The number of falsely detected pixels follows (10), i.e. we expect that half of the watermarked pixels are falsely detected. The dilation of the image by the structuring element v_1 aims at creating a connected object in these areas. Thus, each dimension s of the structuring element v_1 should follow (14) for $n = L/2$, where L is the average number of watermarked pixels per dimension:

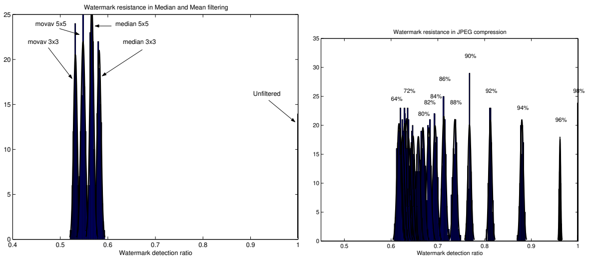
$$s = \frac{L+4}{L(L+2)} N \tag{15}$$

Once we have detected the tampered regions we want to remove the objects that are produced from only one falsely detected watermarked pixel. These isolated falsely detected pixels are generated due to high quality compression or minor image processing operations that do not reduce the image quality. To do so we apply an erosion by a structuring element with size $(s + 2) \times (s + 2)$ since the size of that objects is now $s \times s$. Finally a dilation by a structuring element of size 3×3 is applied in order to bring the tampered regions back to their original size.

5. EXPERIMENTAL RESULTS

As it was described in the introduction, the objective of an image authentication algorithm is to provide the user with all the information needed about the image authenticity. If we consider image data as untouchable messages, then the data for authentication have to be exactly the same as the original one. In that case, any alteration of the image is prohibited and the threshold of the acceptable image authenticity should be very high. Thus, the use of the authenticity measure is enough to assure authenticity. This type of authentication is called *complete authentication*. However, if the authentication of the image content is the objective, then high quality image compression is acceptable. The image authentication should fail if the image is tampered. Thus, only the use of the authenticity measure is not adequate for deciding about the image authenticity. In that case the image authentication algorithm should have the ability of tamper proofing. That is, to detect the tampered image regions that change the image content. This type of authentication is also called *content authentication*.

The proposed algorithm is adequate both for complete authentication as well as for content authentication. By varying the threshold of the watermark detection ratio we can pass from complete authentication to content authentication. Moreover, image content authenticity is assured by using the method presented in the previous section. Experiments were made for the proposed methods performance when image filtering, image compression, image editing and combined attacks are made to the host image.



(a) (b)

Figure 1: The estimated pdfs of the watermark detection ratio for several filtering types (a), and compression ratios (b).

Image filtering: If the host image is filtered then the output of the authentication algorithm should be negative, since filtering is considered to destroy authenticity. We have tested the performance of the proposed method for two types of image filtering with different window sizes. These filters are the moving average filter of size 3×3 and 5×5 and the median filter of size 3×3 and 5×5.

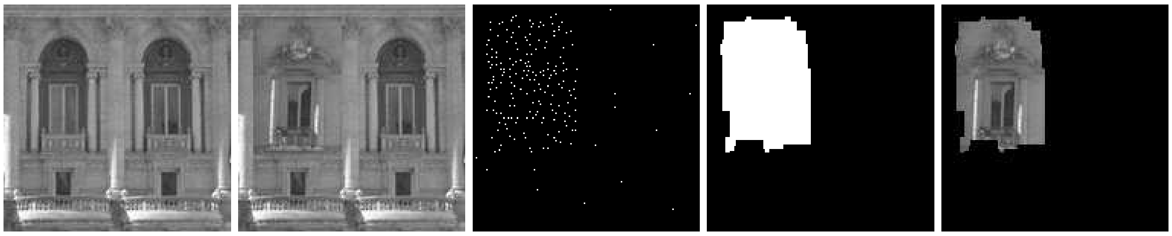
Ideally the pdf of the watermark detection ratio after filtering the original image should follow (10). In order to estimate the pdf of the detection ratio the reference (host) image was watermarked with 100 watermarks generated by randomly chosen keys and was filtered using the aforementioned filters. The estimated pdfs of the detection ratio are depicted in Figure 1a. It is obvious from the plots that the detection ratio is reduced significantly when the image is filtered and the output of the detection algorithm is that the image is not authentic.

(a) (b) (c)

Figure 2: (a) Original watermarked image. (b) Altered image. (c) False detection image.

Image editing: The most frequent type of tampering attack in an image is the selective editing of certain image regions or the tampering of fine image details. This type of attack aims at changing the content of an image and not necessary at reducing the quality of the original image. Thus, it is very difficult to detect such alterations. They succeed to change the content of the image without changing the watermark detection ratio significantly. The smaller are the tampered regions that can be detected by an image authentication algorithm, the better the algorithm performance is.

The algorithm described in Section 4 is used for tampered region detection. As has already been described, the watermark detection ratio should follow (10) in the tampered regions. The average number of watermarked pixels per row for an image of size 512×512 is $L = 34$ and the size of the structuring element that should be used for the detection of the tampered regions is calculated by (15) for the corresponding L. In this example, the size of the structuring element is found to be 16×16. We choose it to be of size 17×17 for having structuring element of odd size.



(a) (b) (c) (d) (e)

Figure 3: (a) Original watermarked image. (b) Altered compressed image with compression ratio 1:2. (c) False detection image. (d) Image map that highlights the non authentic regions. (e) Non authentic image regions

The original image is edited in several regions in such way as the alterations are imperceptible. The tampered image is illustrated in Figure 2a. The false detection image is shown in Figure 2b. By applying the procedure described in Section 4 the non authentic regions are highlighted.

Image compression: Although lossy image compression is not desirable for high quality multimedia products (complete authentication) we would prefer that the image authentication method is robust for high quality image compression (e.g., compression ratio 1:2, 1:3). Many methods that have been developed for image authentication have the major drawback that they are not robust for high quality lossy image compression. The method proposed in this paper has the advantage that the watermark embedding functions (1) are designed as to be robust in high quality compression. Accordingly, we could design embedding functions that are robust to image filtering as well, if robustness to image filtering was necessary [7].

In order to estimate the pdf of the detection ratio after lossy image compression with several compression ratios, the reference host image was watermarked with 100 watermarks generated by randomly chosen keys and was compressed at several compression qualities. The estimated pdfs of the watermark detection ratio for several compression qualities are depicted in Figure 1b. It is obvious from the plots that the detection ratio is significantly reduced when the image is heavily compressed. However, the drop of the detection ratio is smooth and for high quality image compression, the detection ratio is more than 90%. In that case if the authenticity measure threshold is 90% and the falsely detected watermarked pixels are spread all over the image domain, the image is considered authentic. To examine whether the falsely detected pixels are spread all over the image (high quality lossy compression) or are concentrated in a small area (image editing) the method proposed in Section 4 is used.

Combined distortions: When a watermarked image is compressed with high quality and then is edited somehow, the detection ratio will be decreased as it was described previously. The false detection image (8) will contain falsely detected watermarked pixels by the lossy compression and by the image editing. In that case the objective is to discriminate the regions that contain falsely detected pixels caused by lossy compression and the tampered regions.

The method proposed in Section 4 is used for discriminating between these two ways of tampering (lossy compression, image editing). The method, as it was described, has the ability to create a compact object in the regions that the density of falsely detected watermarked pixels is high while it clears all the other falsely detected watermarked pixels. A watermarked image that is compressed with compression ratio 1:2 and then edited is depicted in Figure 3b. The false detection image is shown in Figure 3c where the different density in falsely detected pixels, for the edited regions of the image, can be observed. The detection image that results after applying the method proposed in Section 4 is illustrated in Figure 3d. The tampered regions that are not considered authentic are highlighted by the use of the proposed algorithm while the high quality compressed regions are considered to be authentic.

6. CONCLUSIONS

A novel method for image authentication and tamper proofing has been proposed. It succeeds in detecting any alteration made in a watermarked image and decide for its authenticity. The proposed algorithm provides the user not only with a measure for the authenticity of the test image but also with an image map that highlights the unaltered regions of the image when selective tampering has been made.

7. REFERENCES

[1] G.L. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Transactions on Consumer Electronics*, vol. 39, no. 4, pp. 905–910, November 1993.

[2] M. Schneider and S.F. Chang, "A robust content-based digital signature for image authentication," in *Proc. of ICIP'96*, Lausanne, Switzerland, September 1996, vol. III, pp. 227–230.

[3] S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *Proc. of ICIP'98*, Chicago, USA, 4-7 October 1998, vol. I, pp. 425–429.

[4] B. Zhu, M.D. Swanson, and A.H. Tewfik, "Transparent robust authentication and distortion measurement technique for images," in *Proc. of DSP'96*, Loen, Norway, September 1996, pp. 45–48.

[5] L. Xie and G.R. Arce, "Joint wavelet compression and authentication watermarking," in *Proc. of ICIP'98*, Chicago, Illinois, USA, 4-7 October 1998, vol. II, pp. 427–431.

[6] P.W. Wong, "A public key watermark for image verification and authentication," in *Proc. of ICIP'98*, Chicago, USA, 4-7 October 1998, vol. I, pp. 425–429.

[7] G. Voyatzis and I. Pitas, "Digital image watermarking using mixing systems," *Computer & Graphics*, vol. 22, no. 3, 1998.

[8] G. Voyatzis and I. Pitas, "The use of watermarks in the protection of digital multimedia products," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1197–1207, July 1999.

[9] J. Fridrich, A.C. Baldoza, and R.J. Simard, "Symmetric ciphers based on 2d maps," in *Proc. IEEE Conf. on Systems, Man, and Cybernetics*, Octomber 1997, pp. 1105–1110.

[10] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, London, 1982.

[11] H.A. David, *Order Statistics*, John Wiley & Sons, New York, 1981.