BLIND DETECTION OF MALICIOUS ALTERATIONS ON STILL IMAGES USING ROBUST WATERMARKS

Christian Rey and Jean-Luc Dugelay*

Abstract

Digital image manipulation software is now readily available on personal computers. It is therefore very simple to tamper with any image and make it available to others. Insuring digital image integrity becomes a major issue. In this paper, we propose an original method to protect image authenticity using an invisible and robust watermark. Our scheme is independent of the signer, however the latter must have a high capacity and be able to extract the watermark in full blind detection mode. Our approach is based on the extraction of features from the image. These features are chosen so as to be unaffected by non malicious alterations such as lossy compression. They are embedded in the image using an iterative process so that watermarked image features and information contained in the watermark coincide perfectly. The authenticity is verified by comparing the features of the tested image, with those of the original image recovered from the watermark.

1. Introduction

Digital image manipulation software is now available on personal computers. It is therefore very simple to both tamper with any image and to make it available to others. Furthermore, new standards, such as MPEG 4, open new possibilities in terms of interactive multimedia applications [1]. Users will have the possibility to easily modify the content of a scene. For instance, they will be able to move, modify, delete or add objects to a scene. In this context, it is important to protect both the scene and its different components in terms of copyright as well as to preserve its content. A minimal requirement is to at least detect these types of digital manipulations.

In the security community, an integrity service is unambiguously defined as one which insures that the sent and received data are identical. Of course, this binary definition is also applicable to image, however it is too strict and not well adapted to this type of digital document. Indeed, in real life situations, images will be transformed their pixel values will therefore be modified but not the actual semantic meaning. In other words, the problem of image authentication is released on the image content, for example: when modifications of the document may change its meaning or visually degrade it. In order to provide an authentication service for still images, it is important to distinguish between malicious manipulations, which consist of changing the content of the original image (captions, faces, etc.) and manipulations related to the usage of an image such as format conversion, compression, filtering, etc.

Unfortunately this distinction is not always clear, it partly depends on the type of image and its usage. Indeed the integrity criteria of an artistic master piece and a medical image will not be the same. In the first case, a Jpeg compression will not affect the perception of the image, whereas in the second case it may discard some of the fine details which would render the image totally useless. In the latter case, the strict definition of integrity is required.

* Institut Eurécom, Multimedia Department – Sophia Antipolis, FRANCE

2. Which method for image authentication

Until now, the majority of publications in the field of watermarking mainly address the copyright of still images. Other security services, such as image content authentication, are still marginal, and many fundamental questions remain open. One may wonder for example if it is preferable to use a fragile watermark, a robust watermark or even use a completely different technique.

Most methods currently proposed [2, 3, 4, 5] for providing image authentication are based on a fragile or semi-fragile watermark. The basic idea underlying those techniques is to insert a specific watermark so that any attempt to alter the content of an image will also alter the watermark itself (figure 1). The authentication process therefore consists of locating watermark distortions in order to locate the regions of the image that have been tampered with. The major drawback of these approaches is that it is difficult to distinguish between malicious and non-malicious attacks.

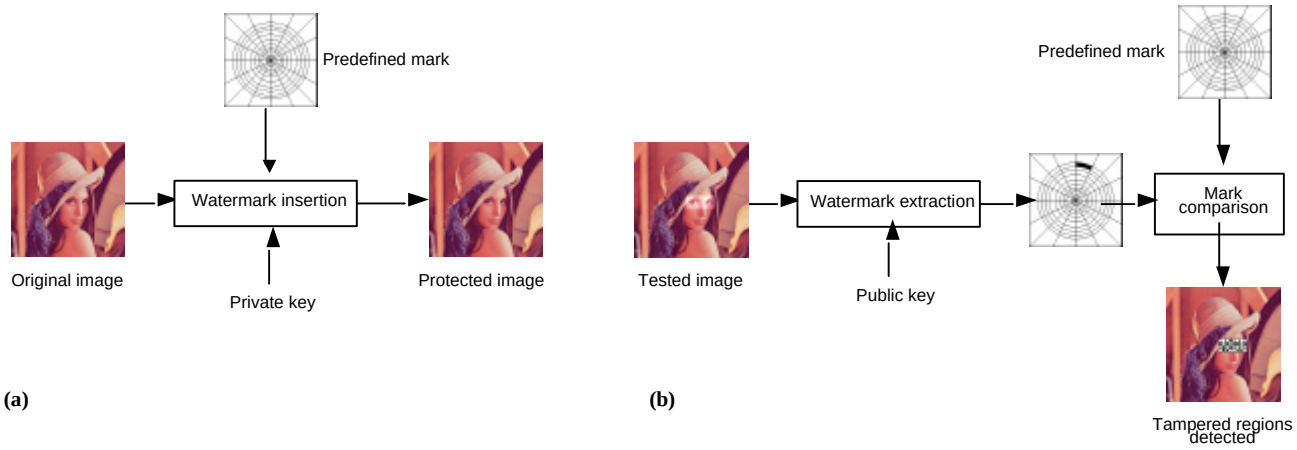

Figure 1 – Generic fragile watermark scheme: (a) Image security, (b) Authenticity verification

There are alternative techniques to classical watermarking approaches, which provide image authentication. A number of these consist, as described by Queluz [7], and Lin and Chang [8, 9], in using an external digital signature which is generated from image features. There are some analogies between those methods and image indexing where many techniques use this type of digital signature in order to retrieve images with respect to their content [6]. These signatures are generated from significant features that summarize the semantic content of the image such as colour, shape or texture.

In the context of image authentication, the signature is encrypted and then transmitted along with the image data. The authentication is operated via the decoding of the signature followed by a comparison between the decoded signature and characteristics extracted from the image (figure 2). Additionally, Bhattacharjee et Kutter [10] propose a technique which uses an external signature based on feature-point. These feature-points are chosen so that they remain relatively invariant after a lossy compression.

The main advantage of those methods is that they allow for a more robust and precise detection of tampered regions. However, in this case the image is not self-sufficient. Therefore, the benefits of watermarking are reduced and it becomes necessary to be able to guarantee the authenticity of the image/signature pair.

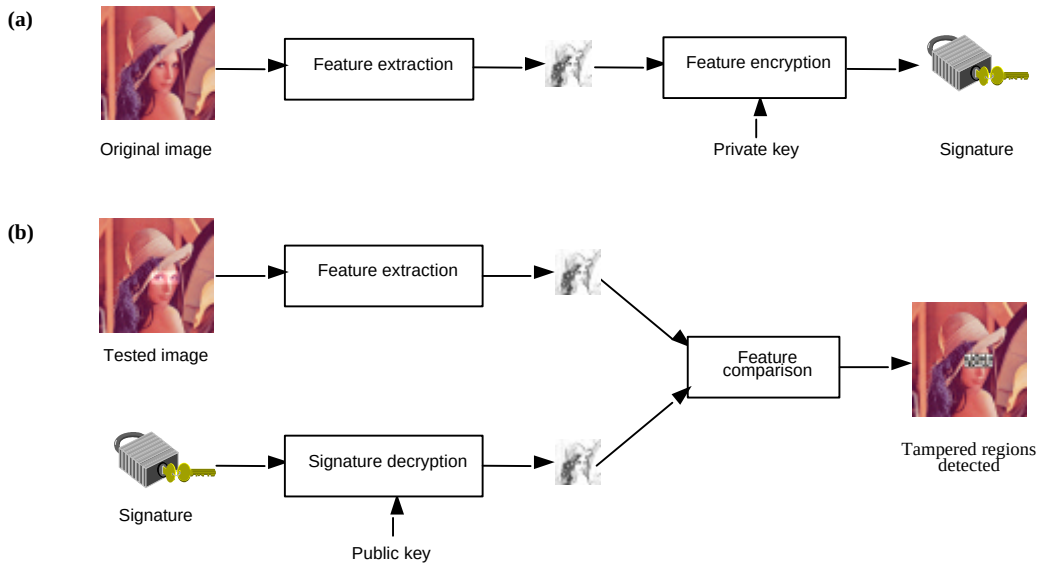

Figure 2 – Generic digital signature scheme: *(a)* Signature generation, *(b) Authenticity verification*

From an early definition proposed by Wu and Liu [4], we can express the key points that have to be taken into account when designing an image authentication system:

- Determine whether an image has been tampered or not;
- Locate precisely any malicious alteration made on the image;
- Embed authentication data in the image rather than in a separate file;
- The watermark must be invisible;
- Authentication system must tolerate some loss of information (originating from compression algorithm) and more generally non-malicious manipulations;
- Finally allow to restore altered regions (even partially).

3. Proposed scheme

The proposed method relies on a robust watermark, and has the advantage of being independent of the watermarking algorithm. However, the signer must have a high capacity and be able to extract the signature in full blind mode.
We have used our own fractal-inspired signer originally designed for copyright [11, 12]. Its performances are fairly good in terms of robustness (e.g. robust to *Stirmark 3.1* [13, 14]), but it has been originally calibrated for hidden message of size 64 bits, for copyright purpose.

3.1. Basic principle

As opposed to classical techniques, which employ a fragile watermark, the watermark we use is not fixed. Instead, the signature depends on the image itself. The basic idea consists of first extracting features from the original image, and hiding them within a robust and invisible watermark (fig. 3.a.). Then, in order to check whether an image has been altered, we simply compare its features with those of the original image recovered from the watermark (fig. 3.b.).

A number of constraints are imposed by this method, mainly in terms of robustness and storage capacity of the signature. Robustness is required in order to allow lossless extraction. Accurate detection is directly linked to the amount of information inserted into the image. It is necessary to find a good compromise for the size of the signature so that both robustness and accurate detection can be achieved.

Figure 3 – Generic robust watermark scheme: (a) Image security, (b) Authenticity verification

3.2. Choice of the image features

The choice of image features used will directly affect the type of image alterations that we wish to be able to detect. Additionally, those features will depend on the type of image under consideration (paintings, satellite images, medical images, etc.). The features are typically selected so that invariant properties are maintained under weak image alterations (lossy compression) and broken for malicious manipulations. These features could be also used to restore partially the tampered regions of the image.
Typical features used to provide image authentication are: edges, colours, gradient, luminance; or combination of them.

3.3. Iterative embedding process

One of the problems faced by our method is that the image is slightly modified while inserting the watermark. Even small image variations may affect the image properties. Thus since the features of the original image and the watermarked image are not exactly the same, there are risks of false positive detections. This risk may be more or less important according to the choice of selected features.

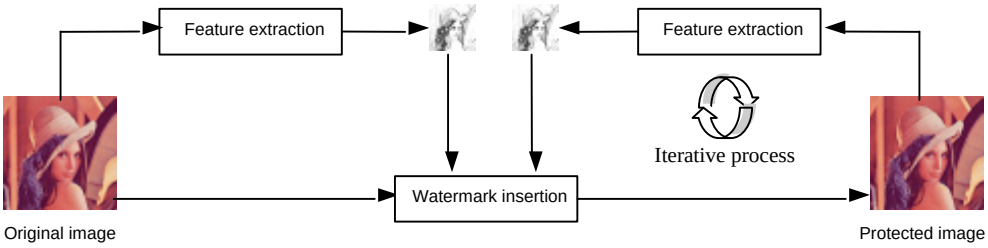

Figure 4 - Iterative embedding process

In order to solve this problem we have implemented an iterative watermarking algorithm (fig. 4). The idea here is to sign the image, extract features from the newly obtained image, and then repeat the watermarking process on the original image (in order to avoid cumulating distortions) using the newly computed features. Thanks to this iterative process, hidden features will perfectly coincide with the protected image features. In practice, three iterations are enough (fig. 5).

(a) *(b)* *(c)*

Figure5 – False alteration detections (the protected image has not undergone any attack) (a) first iteration, (b) second iteration, (c) third iteration.

4. Preliminary results

Figure 6 shows our first results using the previously described technique. In this example, the original image has been protected using the block mean luminance (fig. 6.a.). Using *Paint Shop Pro* we have replaced the kiwi fruit, in the bottom-left hand corner of the image, by a lemon (fig. 6.b). Figure 6.c. shows the regions that have been identified by our system as altered regions. Thanks to such a system, the viewer is able to know that the image has been tampered with.

(a) *(b)* *(c)*

Figure 6 - (a) original image (protected), (b) tampered image, (c) detection of tampered regions.

5. Concluding remarks

In this paper we have presented a new method for digital image authentication. Our initial results are very encouraging. With this method we detect and precisely locate malicious image manipulations such as removal or addition. We are currently working on an automatic method to identify optimal image features so that the best compromise can be achieved. One may also think about extending the technique so that image/object alterations can be identified (shift, zoom, etc.). Additionally, the origin of the object added to the scene could be recovered from its own local watermark if object had been extracted from another protected image or video and re-used.

This work is partly supported by the National French Telecom project: RNRT Aquamars [15].

6. References

[1] Thomas Sikora, 'MPEG Digital Video-Coding Standards', *IEEE Signal Processing Magazine*, pp. 82-100, Sep. 1997.

[2] M. M. Yeung & F. Mintzer, 'An Invisible Watermarking Technique for Image Verification', *IEEE International Conf. on Image Processing (ICIP'97)*, Vol. 2, pp. 680-683, Santa Barbara California US, Oct. 1997.

[3] R. B. Wolfgang & E. J. Delp, 'Fragile Watermarking Using the VW2D Watermark', *SPIE International Conf. on Security and Watermarking of Multimedia Contents*, vol. 3657, No. 22, EI '99, San Jose, USA, Jan. 1999.

[4] M. Wu and B. Liu, 'Watermarking for Image Authentication', *IEEE International Conf. on Image Processing*, Chicago, USA, Oct. 1998.

[5] D. Kundur and D. Hatzinakos, 'Towards a Telltale Watermarking Technique for Tamper-Proofing', *IEEE International Conf. on Image Processing (ICIP'98)*, Chicago, USA, Oct. 1998.

[6] P. Aigrain, H. Zhang and D. Petkovic, 'Content-based Representation and Retrieval of Visual Media: A State-of-the-Art Review', *Multimedia Tools and Applications*, 3(3):179-202, Nov. 1996..

[7] M. P. Queluz, 'Towards Robust, Content Based Techniques for Image Authentication', *IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, Dec. 1998.

[8] C.-Y. Lin and S.-F. Chang, 'A Robust Image Authentication Method Surviving JPEG Lossy Compression', *SPIE Storage and Retrieval of Image/Video Database*, San Jose, Jan. 1998.

[9] C.-Y. Lin and S.-F. Chang, 'Generating Robust Digital Signature for Image/Video Authentication', *Multimedia and Security Workshop at ACM Multimedia 98*, Bristol, UK, Sep. 1998.

[10] S. Bhattacharjee and M. Kutter, 'Compression Tolerant Image Authentication', *IEEE International Conf. on Image Processing (ICIP'98)*, Chicago, USA, Oct 1998.

[11] J.-L. Dugelay and S. Roche, 'Fractal Transform based Large Digital Watermark Embedding and Robust Full Blind Extraction', *in proceedings of IEEE-ICMCS'99*, Florence, Italy, June 1999.

[12] J.-L. Dugelay, 'Procédé de dissimulation d'informations binaires dans une image numérique'. French patent FR2775812, Institut Eurécom, 10 Sep. 1999. Available from http://www.inpi.fr. Also available as WOFR9900485.

[13] Stirmark: http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/

[14] J.-L. Dugelay & F. A. Petitcolas, 'Image watermarking: possible counter attacks against random geometric distortions', *Conference Electronic Imaging 3971, Proceedings of SPIE Vol. 3971, Security and Watermarking of Multimedia Contents II*, San Jose, CA, January 24-26, 2000.

[15] RNRT Aquamars: http://www.telecom.gouv.fr/rnrt/projets/paquamars.htm