33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Medical Image Integrity Control and Forensics Based on Watermarking - Approximating Local Modifications and Identifying Global Image Alterations

H. Huang, G. Coatrieux, Member, *IEEE*, H.Z. Shu, *Senior Member, IEEE*, L.M. Luo, *Senior Member, IEEE*, Ch. Roux, *Fellow, IEEE*

Abstract—**In this paper we present a medical image integrity verification system that not only allows detecting and approximating malevolent local image alterations (e.g. removal or addition of findings) but is also capable to identify the nature of global image processing applied to the image (e.g. lossy compression, filtering …). For that purpose, we propose an image signature derived from the geometric moments of pixel blocks. Such a signature is computed over regions of interest of the image and then watermarked in regions of non interest. Image integrity analysis is conducted by comparing embedded and recomputed signatures. If any, local modifications are approximated through the determination of the parameters of the nearest generalized 2D Gaussian. Image moments are taken as image features and serve as inputs to one classifier we learned to discriminate the type of global image processing. Experimental results with both local and global modifications illustrate the overall performances of our approach.**

I. Introduction

MEDICAL imaging plays a crucial role in the healthcare system. Images support diagnosis, treatment decision and serve also research purposes. As a result, any medical image seen and analyzed has to be trustworthy. Trustworthiness can be mapped into two security components [1]: "integrity", which ensures data has not been modified by non-authorized persons; and "authenticity" which asserts data origin and its attachment to one patient. If it is vital to keep images safe from any damage, it is also important being able to detect an image has been modified and in which manner when considering the medico-legal framework. These are the aspects we focus on in this paper.

Medical images can be modified accidentally, as for example during communication, or deliberately. In the later situation, images can be tampered malevolently with the introduction or removal of findings. If used, such an image will induce in error the medical staff. Some image processing allowed or authorized by the application framework may also lead to similar situations. As example, in telemedicine applications, lossy image compression is tolerated so as to reduce the amount of information to be transmitted. However, depending on its extent this process may induce inacceptable information loss. It may result in a misdiagnosis [2] involving at the same time the responsibility of the physician who, not informed, interprets the modified image.

In our view, three levels of integrity can be considered [3]:

-- Level 1 (L1): Modification Detection -- an alarm should be given under any kind of image modification;

-- Level 2 (L2): Modification Location -- untrustworthy parts of the image have to be indicated; either in a rough way, so as to designate areas still interpretable by the physician;

-- Level 3 (L3): Forensics/integrity analysis -- the nature of the modification over the whole image or within untrustworthy regions has to be identified to have an idea about its origin (accidental, authorized/non authorized).

Different strategies have been proposed in the literature to verify image integrity. These techniques include the use of image digests/signatures/hashes or perceptual hashes[4], watermarking [5] and blind forensics methods [6]. The first kind of methods verifies image integrity based on the comparison of hashes computed over the image under investigation or some parts of it with the hashes shared with the image. Such a hash can be computed in different ways. Cryptographic hash functions allow verifying the exact identity of the image under investigation with the original image, and can be used to achieve L1 [7]. They provide the best performances in terms of detection and are extremely difficult to counterfeit. To localize alterations (L2), one can compute hashes on independent image areas. However, because of cryptographic hash's length, they can be replaced by checksums based on error detection codes [7]. Checksums are less efficient in terms of detection than cryptographic hashes. Perceptual hashes are another kind of image digests [4]. They aim at detecting malevolent image content changes but are robust to global image processing such as JPEG, filtering … processes. Linear digests, as suggested in [8], can also be exploited to achieve L3. Their analysis can lead to the approximation of the modification by a pre-defined model giving thus an idea about its position, extent and amplitude. In this work, we make use of this strategy.

Watermarking is an effective tool for verifying image integrity and authenticity. One common approach consists in inserting a specific watermark [4]. The non-detection of this later informs about image integrity loss. In some cases watermarking is combined with image signatures. As example, in [7] a set of signatures is computed from one Region Of Interest (ROI) and then watermarked within Regions Of Non-Interest (RONI). Some watermarking

H. Huang, G. Coatrieux and Ch. Roux are with IT. Telecom Bretagne; Inserm U650, France, e-mail: {Hui.Huang, Gouenou.Coatrieux, Chiristian.Roux}@telecom-bretagne.eu.

H. Z. Shu, L. M. Luo are with LIST Southeast University, China, e-mail: {shu.list, luo.list}@seu.edu.cn.

schemes have been proposed to restore partly tampered parts of the image [9]. However, these ones only roughly recover the image, i.e. not the image details. Furthermore, they require high embedding capacity and spread the watermark over the whole image introducing risks of interferences with image interpretation.

The third strategy refers to blind forensic technique working with no a priori knowledge about the original image [6]. They involve the extraction of some image features [10] that reveal the statistical nature of image modifications. Computed on an image under investigation, these features are provided as input of a classifier that discriminates original images from others modified by global image processing.

The system we propose watermarks a ROI signature into RONI with the aim of approximating local modification and identifying the nature of global image modification. It is based on a set of image moments digests we proposed in [8], and which are computed in independent pixel blocks. If this signature allows approximating local modification by a generalized 2D Gaussian function model, it fails with global image processing such as JPEG compression. To overcome this issue, we propose in this work to use this signature as image feature for a classifier learned to discriminate the kind of the modification applied to the image.

The rest of this paper is organized as follows. Section II describes our image signature and system architecture. Before concluding in Section IV, Section III reports some experimental results for both local and global modifications.

II. METHOD DESCRIPTION

A. System Architecture

The system we proposed extracts digests from pixel blocks of image ROIs and embeds these digests into the rest of the image (i.e. RONIs). ROI and RONIs can be user defined or automatically detected. RONI may correspond to the black background of the image (Fig. 4b). Since only the non-relevant parts of the image are manipulated, invisibility requirement becomes less strict. It is thus possible to use highly robust watermarking strategies, making the extraction of the watermarked digests possible even after the image has been globally and severely altered. The reader may refer to [7] for more details about the method used.

As illustrated in Fig. 1b, differences between recomputed and extracted digests are analyzed in order to achieve the three integrity verification levels: L1, L2 and L3 (see section I). In our system, L1 is achieved using the well known cryptographic hash functions SHA-256 (Secure Hash Algorithm), whose output is a 160-bit long signature with a misdetection probability lower than $1/2^{160}$. To satisfy L2, the image is divided into non overlapping pixel blocks. In order to minimize the amount of data to be watermarked, we compute one checksum per block by mean of Hamming Codes [7]. Integrity is consequently controlled at the block level and one block is detected tampered if its recomputed signature differs from the embedded one. Based on L1 and L2, it is possible to know if the image has been altered (L1) and which parts cannot be used trustingly (L2). The next step is to achieve L3.

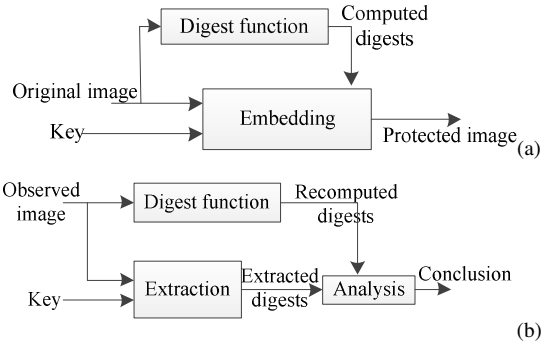


Fig. 1. Principle of our integrity control system. A secret watermarking key is used to guarantee that only the entitled users can verify the image integrity. (a) Integrity protection; (b) Integrity verification.

B. Image Moment for Local Tampering Approximation

Our focus is to have an idea about the shape of one malevolent tampering and to know if it is a finding addition or removal. We also aim at refining more precisely the tampering position within a pixel block (see section II.A) and its dimensions (i.e. amplitude and size). Herein, we defined the modification as the difference signal between the original image and its modified version. Working on pixel blocks, B_{org} and B_{mod} being the original and the modified block respectively, we approximate the modification $\Delta_B = B_{mod} - B_{org}$ by its nearest 2-D generalized Gaussian function model (G), as shown in Fig. 2 estimating its different parameters.

Parameters of the Gaussian function to be determined are: its center of mass (r_0, c_0), where r_0 and c_0 are the row and column of the center, the direction of the major axis (θ), the deviations (σ_1 and σ_2) along the major axis and minor axis (Fig. 2) and the amplitude (A).

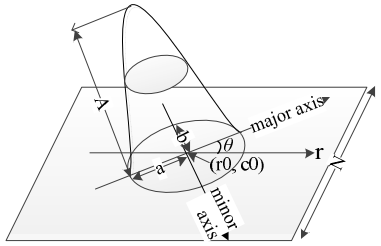


Fig. 2. Gaussian function model in an N×N pixel block

To estimate these parameters we suggested in [8] a pixel block digest which includes geometric image moments. Beyond the fact they are easy to compute and that they are linear - allowing us to gain easy access to the moments of the modification Δ_B (i.e. by subtracting moments of B_{org} and B_{mod}) – they also have interesting properties that can help us to determine the parameters of the Gaussian function (G).

Introduced by Hu [11], one general image moment M_{nm} is defined with a basis function $\varphi_{nm}(x,y)$ and an image intensity function $f(x,y)$

$$M_{nm} = \sum_x^N \sum_y^N \varphi_{nm}(x,y)(f(x,y)\,,\ n,m = 0,1,2,\ldots \quad (1)$$

Among existing moments, geometric moments are the simplest one defined with $\varphi_{nm}(x,y) = x^n y^m$. Different orders of these moments represent different spatial

characteristics of the image intensity distribution, which can give us access to parameter estimator of the Gaussian function *G:*

-- 0^{th} order moment M_{00}, represents the sum of gray value of the image;

-- The center of mass of *G*, thus the center of Δ_B, can be derived from the two first order moments (M_{01}, M_{10}): $c_0 = M_{10}/M_{00}, r_0 = M_{01}/M_{00}$;

-- The second order geometric moments (M_{02}, M_{11}, M_{20}) can determine several image features, such as the direction of major axis, the deviations (σ_1 and σ_2) and the length of axis of the Gaussian bottom [12].

Due to the limited paper length, we cannot detail the complete procedure of modification approximation. Readers may refer to [8] for more details. Nevertheless, this procedure requires to watermark for each pixel block the six following geometric image moments: $M_{00}, M_{10}, M_{01}, M_{11}, M_{20}, M_{02}$. Because, two bytes are needed to encode each of these moments, we achieve L3 with a 96-bit long digest per block.

As reported in [8], this approach suffers from the uni-polarity and linearity of the modification model. It fails to give the right approximation of modifications which include both positive and negative signal variations as those induced by common global image processing. The solution we propose to overcome this issue is exposed in the next section.

C. *Global Image Modification Identification*

Inspired by the strategy followed by blind forensic mechanisms (see section I), we propose to exploit the previous geometric moments $M_{00}, M_{10}, M_{01}, M_{11}, M_{20}, M_{02}$ computed on independent pixel blocks as image features within a classification process which purpose is to distinguish the kind of the global process applied to an image (lossy JPEG/JPEG2000 compression, filtering, scaling, …). Image features used by blind forensics solutions point out the inherent signal variations specifically attached to one kind of image modification. These variations are principally located within the image details and are partly independent of the image content (see [6]).

In our system, we watermark an a priori knowledge: the original pixel block image moments. We can thus remove the effect of the image content and focus on moments alteration caused by image processing. Suppose the modified version of the image $f(x,y)$ is $f(x,y) + \varepsilon(x,y)$, their respective $(n+m)^{th}$geometric moments (M_{nm} and M'_{nm}) are:

$$M_{nm} = \sum_x^N \sum_y^N x^n y^m f(x,y) \quad (2)$$

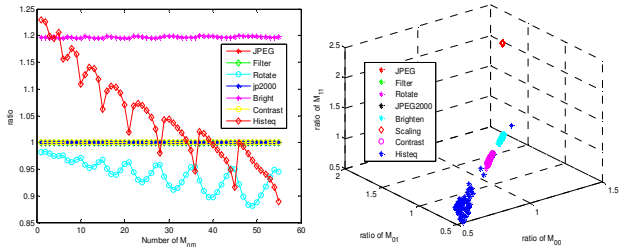
$$M'_{nm} = \sum_x^N \sum_y^N x^n y^m (f(x,y) + \varepsilon(x,y)) \quad (3)$$

As illustrated in Fig. 3, in the case of one retina image (see Fig. 4a), the ratio between M_{nm} and M'_{nm} (see eq. (4)) discriminate not too badly different image processing.

$$\nabla M_{nm} = M'_{nm}/M_{nm} \quad (4)$$

Fig. 3a shows the moment ratios variations up to 10^{th} orders (55 moment values) between one image and different modifications. Fig. 3b shows the scatter diagram of moment ratios -- $\nabla M_{00}, \nabla M_{01}, \nabla M_{11}$. Whence, it is possible to distinguish different image processing forms by learning classifiers with the moment ratios as image features.

The determination of the modification type is a multi-class decision problem. One of the solutions for constructing a multi-class classifier is to convert the multiclass classification problem into a set of binary classification problems that are efficiently solved using binary classifiers. For these experiments, Support Vector Machine (SVM) [13] was used as binary classifier. This choice stands on the fact that SVM provide superior classification performances in many applications. Thus, in our context, one SVM has been learned to differentiate two different modification types (e.g. JPEG vs. Filtering, Rotation vs. Scaling …). Based on the responses of these classifiers, a multi-class conclusion is drawn. Among the different strategies for combining decisions of binary classifier, Max-Wins Voting (MWV) is commonly used [14].



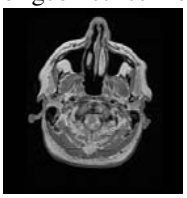
(a) (b)

Fig. 3. Ratios between geometric moments of the original image and its modified version in the case of retina images; (a) Ratios of moments of order up to 10^{th} (55 values) for modifications; (b) Scatter diagram of moments ratios for different modification (with the first 2^{nd} order moments).

III. Experimental Results

In these experiments, it is assumed that L1 and L2 procedures have been done. We have worked on blocks of two distinct sizes. Whatever the image modality, the elementary block size is of 64×64 pixels from which 6 geometrics moments are extracted and watermarked according to the above procedure. Local modification approximation is conducted at this resolution while global modification identification is applied on 128×128 blocks. The reason of this choice results from the fact that we have used the same system to build up SVM classifiers. Technically, this does not require the watermarking of extra-geometric moments. Indeed, geometric moments of 128×128 block can be linearly derived from those of its four 64×64 sub-blocks [12]. This is a precious advantage of geometrics moments.

(a)



(b)

Fig. 4. Image test samples a) Retina image, squared regions indicate position of local modifications (b) MRI of the head JPEG compressed with Q=75.

Experiments have been conducted on retina images of 8-bit depth (Fig. 4a) and on 12-bit depth MRI of the head of 256×256 pixels (Fig. 4 b). If in MRI the ROI corresponds to the "head" and is automatically detected [7], ROIs in retina are identified manually. For global modification, we use only

TABLE I - Detection rates of our multi-class classifier

Image	JPEG2000	JPEG	filtering	rotation	scaling	brighten	contrast	Hist. Eq.
Retina	68.06	97.50	80.00	79.17	80.00	100	74.31	100
MRI	67.03	99.17	77.66	80.00	80.00	98.79	80.52	100

6 geometric moment ratios from the 128×128 block centered at the ROI.

We have considered both local and global modifications. For local modifications (see Fig. 5), some findings have been removed. In order to evaluate the approximation performances of our scheme we use the Mean Square Error (MSE) to measure the distance between the real modification and its approximation. For global image processing, the detection rate is used as performance indicator. It corresponds to the number of modified and original images correctly detected versus the number of tested images.

Considering the two modifications shown in Fig. 5a and 5b, we get MSE values of 101.15 and 4.01 respectively. It is obvious that the more the real modification is similar to the model - in this experiment a 2D Gaussian function, the more the approximation is correct.

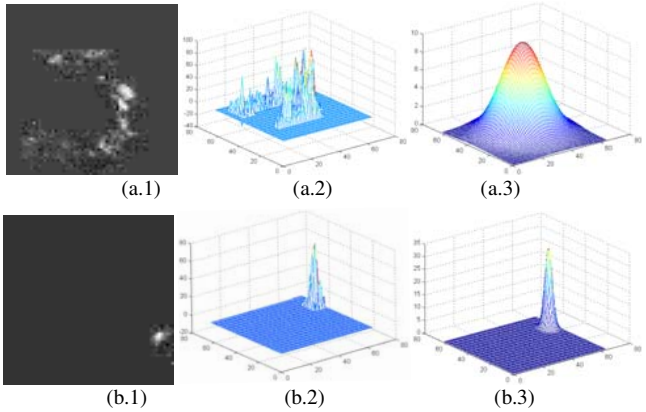


Fig. 5. Image modifications and their respective approximations: (a.1)(b.1) correspond to 2D view of the modifications and (a.2)(a.3)(b.2)(b.3) to 3D views of real modifications and their approximations respectively.

Global modifications we retained for these experiments as well as the parameters we have used are given in Table I. These modifications have been considered for both MRI (see Fig. 4b) and retina images. Binary classifiers (i.e. SVMs) have been learned on 72 retina images and 145 MRI images with the 6 geometric moment ratios as image features.

TABLE II IMAGE MANIPULATION AND THEIR PARAMETERS

Modification	Values of parameters				
Scaling up(γ_s %)	1	5	10	25	50
Rotation angle(θ)	1	5	15	30	45
Deviation of Gaussian filter (σ)	0.3	0.5	1.0	2.0	3.0
Contrast enhancement rate (γ_c %)	1	5	8	10	
Brighten rate (γ_b %)	2	5	8	10	
Quality factor(Q)	95	85	80	75	60
Compression rate JP2K (γ_j)	2:1	5:1	10:1	20:1	50:1
Histogram equalization					

In our system, one image is detected as globally modified if all independently protected pixel blocks are declared unauthentic at the output of level L2. In that case, this image is passed through all binary classifiers and a final decision about the kind of modification is taken making use of the MWV method. We give in Table II, the detection rates we achieved depending on the nature of the modification for retina and MRI images. From these results it can be seen that the type of the modification is estimated correctly with acceptable detection rate except for the lossy JPEG2000 compression. In fact, when JPEG and JPEG 2000 are considered as part of possible modification types, JPEG2000 classifier is easily confused with JPEG compressed images. Indeed, if JPEG compression is omitted, the detecting rate for JPEG2000 can reach 96%.

IV. CONCLUSION

In this paper, we have proposed a system for verifying the integrity of medical images. This system distinguishes three levels of integrity decision: detection, localization and approximation of the image alteration. For the latter level, we suggest approximating any malevolent local modifications by its nearest 2D generalized Gaussian function whose parameters are derived from the geometric image moments. In case the image is globally processed, these image moments can be used to identify the modification type. Our system can help to find out the motivation of the tampering, but it keeps limited to the detection of predefined kinds of image modification or tampering.

REFERENCES

[1] L. Kobayashi and S. Furuie, "Proposal for DICOM Multiframe Medical Image Integrity and Authenticity," *Journal of Digital Imaging*, vol. 22, no. 1, pp. 71-83, Feb. 2009.

[2] A. Giakoumaki, S. Pavlopoulos, and D. Koutouris, "A medical image watermarking scheme based on wavelet transform," in *Proc. of IEEE-EMBC Conf.*, 2003, pp. 856–859.

[3] H. Huang, G. Coatrieux, J. Montagner, H. Z. Shu, L. M. Luo, and C. Roux, "Medical image integrity control seeking into the detail of the tampering," in *Proc. of IEEE-EMBC Conf.*, *2008*, pp. 414-417.

[4] B. Schneier, *Applied cryptography: protocols, algorithms, and source code in C*. Wiley, 1996.

[5] X. Guo and T.-ge Zhuang, "Lossless Watermarking for Verifying the Integrity of Medical Images with Tamper Localization," *Journal of Digital Imaging*, vol. 22, no. 6, pp. 620-628, Dec. 2009.

[6] H. Farid, "Image forgery detection," *Signal Processing Magazine, IEEE*, vol. 26, no. 2, pp. 16-25, 2009.

[7] G. Coatrieux, H. Maitre, and B. Sankur, "Strict integrity control of biomedical images," in *Proceedings of SPIE*, 2001, vol. 4314, pp. 229 -240.

[8] H. Huang, G. Coatrieux, H. Z. Shu, L. M. Luo, and C. Roux, "Medical image tamper approximation based on an image moment signature," in in *Proc. of IEEE- Healthcom Conf.*, 2010, pp. 254-259.

[9] S. C. Liew and J. M. Zain, "Reversible Tamper Localization and Recovery Watermarking Scheme with Secure Hash," *European Journal of Scientific Research*, vol. 49, no. 2, pp. 249–264, 2011.

[10] S. Bayram, I. Avcibas, B. Sankur, and N. Memon, "Image manipulation detection," *Journal of Electronic Imaging*, vol. 15, no. 4, pp. 041102-17, Oct. 2006.

[11] Ming-Kuei Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179-187, 1962.

[12] D. R. Mukundan, K. R. Rao, and K. R. Ramakrishnan, *Moment Functions in Image Analysis*.

[13] L. Wang, *Support vector machines: theory and applications*. Springer, 2005.

[14] U. H. G. Kressel, "Pairwise classification and support vector machines," in *Advances in kernel methods*, 1999, pp. 255–268.