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

Reversible Watermarking based on Invariant Image Classification and Dynamical Error Histogram Shifting

W. Pan, Ph.D. Student, G. Coatrieux, N. Cuppens, F. Cuppens, *IEEE Members*, and Ch. Roux, *IEEE Fellow*

Abstract—**In this article, we present a novel reversible watermarking scheme. Its originality stands in identifying parts of the image that can be watermarked additively with the most adapted lossless modulation between: Pixel Histogram Shifting (PHS) or Dynamical Error Histogram Shifting (DEHS). This classification process makes use of a reference image derived from the image itself, a prediction of it, which has the property to be invariant to the watermark addition. In that way, watermark embedded and reader remain synchronized through this image of reference. DEHS is also an original contribution of this work. It shifts predict-errors between the image and its reference image taking care of the local specificities of the image, thus dynamically. Conducted experiments, on different medical image test sets issued from different modalities and some natural images, show that our method can insert more data with lower distortion than the most recent and efficient methods of the literature.**

I. Introduction

For about ten years, several reversible watermarking schemes have been proposed for protecting images of sensitive content, like medical images for which any modification may impact their interpretation. These methods allow the user to restore exactly the original image from its watermarked version by removing the watermark. It becomes thus possible to update the watermark content, as for example security attributes (e.g. one digital signature or some authenticity codes), at any time without adding new image distortions. However, if the reversibility property relaxes invisibility constraints, it may also introduce discontinuity in data protection like for data encryption. In fact, the image is no more protected once the watermark removed. So, even though watermark removal is possible, its imperceptibility has to be guaranteed as most applications have a high interest to keep the watermark in the image as long as possible, thus continuously protecting the information in its storage, transmission and also processing [1]. This is the reason why, there are still needs for reversible techniques that introduce the lowest distortion as possible with high embedding capacity.

The concept of reversible watermarking has been introduced by Mintzer *et al.* [2] in 1997. Basically, the watermark signal is added to the image taking care not introducing gray level value under-flows (negative) and over-flows (greater than 2^d-1 for a d bit depth image). To satisfy this constraint, Honsinger *et al.* [3] add the watermark w to the image I using arithmetic modulo: $I_w = (I + w) \bmod V_{max}$, where V_{max} corresponds to the maximum value of the signal dynamic range and I_w corresponds to the watermarked image. Though this method avoids over/underflows it may introduce a salt and pepper noise due to jumps between congruent values of the image histogram. Since, several other methods have been proposed.

In [4], Ni *et al.* introduced the famous Histogram Shifting (HS) modulation. HS adds gray values to some pixels in order to shift a range of the image histogram and create a 'gap' near the histogram maxima. Pixels which belong to the class of the histogram maxima are then shifted to the gap or kept unchanged to encode one bit of the message '0' or '1'. Instead of working in the spatial domain, several schemes apply HS to some transformed coefficients or pixel predict-errors which histograms are concentrated around one single maxima located on zero. This maximizes HS capacity [5-7] and also simplifies maxima location within the histogram. In [5], Thodi *et al.* applied HS to the difference of two adjacent pixels for data embedding. In [6], we extended Ni *et al.* scheme to Haar wavelet coefficients which have a "Laplacian" distribution. Recently in [7], Sachnev *et al.* come back into the spatial domain and propose to predict pixel value through its four nearest neighboring pixels. They apply HS to the predict-error and achieve better performances than any existing schemes. In fact, it appears that the shape of their predict-error distribution has a smaller variance than pixel difference distribution.

Even though a better predictor can improve HS capacity, this one is actually uniquely defined for the entire image in the above strategies. In the sequel, we show up how local specificities of the image can be used to improve embedding capacity. These specificities are identified from the image predict-errors. We name our method Dynamical Error Histogram Shifting (DEHS) and introduce it in section II.

Moreover, depending on the image content, Pixel Histogram Shifting (PHS) may be more efficient than Predict-Error Histogram Shifting (PEHS). That is the case for medical images which usually contain a lot of black background (i.e. pixels of null gray value). In this region where PEHS makes more difficult the management of over/underflows, PHS will provide better results with less complexity (the black background histogram maxima

W. Pan, G. Coatrieux and Ch. Roux are with the Institut Telecom; Telecom Bretagne; Unite INSERM 650 Latim, Technopole Brest-Iroise, CS 83818, 29238 Brest Cedex 3 France (e-mail: {wei.pan, gouenou.coatrieux }@telecom-bretagne.eu).

N. Cuppens and F. Cuppens are with the Institut Telecom; Telecom Bretagne; UMR CNRS 3192 Labsticc.

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

corresponds to the null gray value). Capacity is maximized and underflows are simply avoided by shifting pixel value to the left, i.e. by adding a positive gray value. Consequently, considering the local content of the image in order to select the most locally adapted lossless modulation should allows optimizing the compromise capacity/image distortion. However, one question remains about how watermark embedder and reader remain synchronize. In fact, the reader needs to know which modulation it has to use for message extraction.

The method we propose in this paper is derived from the one we proposed in [6]. This latter makes use of an image classification process for the purpose of identifying parts of the image that can be additively watermark without introducing under/overflows. This classification process is conducted on a reference image derived from the image itself, a prediction of it, which has the property to be invariant to the watermark addition. Thus watermark embedder and reader remain synchronized; the reader will retrieve the same image of reference. Herein, we propose to adapt this process so as to identify image parts where PHS and DEHS will be the more efficient. We will show in this paper that for medical images PHS is applied to the black background of the image and DEHS to the rest of the image, i.e. regions with non-null image signal.

The rest of the paper is organized as follows. The proposed scheme and its main functionalities (HS based modulations and image classification) are presented in section II. Section III sums up the performance analysis of our scheme in terms of imperceptibility and capacity and in comparison with one very recent and efficient approach proposed in [7]. Experiments have been conducted on different sets of medical images from different modalities and also on some common natural test images for a fair comparison with [7]. Conclusions are made in section IV

II. Proposed reversible watermarking scheme

As mentioned above, our scheme is additive and leans on two main steps. The first corresponds to a signal classification for the purpose of identifying two sets of regions in the image, sets that will be independently watermarked by PHS or DEHS. The classification output is invariant to the second step: message embedding. At the detection stage, the classification process remains the same and the reader just has to extract the message and restore the image with the convenient HS modulation. In this section, we summarize the key points of PHS and DEHS modulations before presenting the classification process and our complete scheme.

A. *Basic HS Modulation principles*

1) Pixel Histogram Shifting (PHS)

As said before and shown in figure 1a, PHS or the original HS modulation introduced by Ni *et al.* consists in shifting a range of the image histogram by adding or subtracting one gray level to pixels from the histogram maxima (h_{max}) toward its minima (h_{min}), so as to leave one gray level empty (a "gap") near h_{max}. Depending on the bit of the message to embed, pixels that belong to class h_{max} are shifted to the gap to encode '1' or left unchanged to encode '0'. As defined and as mentioned before, PHS is more appropriate on the black background of the medical image where h_{max} correspond to the zero gray level value.

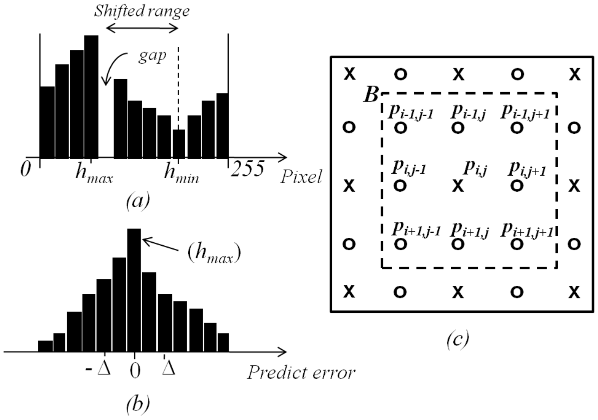

Fig.1. (a) PHS of a grayscale image - h_{max} and h_{min} correspond to the histogram maxima and minima respectively; (b) Predict-Error HS, Δ is the magnitude of the shift; (c) Pixel neighborhood for prediction – in a 3x3 pixels block B, $p_{i,j}$ is estimated through its four nearest neighbors $p_{i-1,j}$, $p_{i,j+1}$, $p_{i+1,j}$ and $p_{i,j-1}$.

2) Predict Error Histogram shifting (PEHS)

Working on predict-error simplifies histogram maxima location because the histogram is centered on zero (see fig. 1b). The predict-error $e_{i,j}$ of one pixel $p_{i,j}$ is given by $e_{i,j} = p_{i,j} - \hat{p}_{i,j}$, where $\hat{p}_{i,j}$ is the predicted value of $p_{i,j}$ (see fig. 1c). In the sequel and as in [7], we consider $\hat{p}_{i,j} = \left(p_{i-1,j} + p_{i,j+1} + p_{i+1,j} + p_{i,j-1}\right)/4$ where $p_{i-1,j}$, $p_{i,j+1}$, $p_{i+1,j}$, $p_{i,j-1}$ are the four nearest neighbor pixels of $p_{i,j}$ (see fig. 1c). This predict-error has a Laplacian distribution.

The predict-error can thus be shifted by +/-Δ, where Δ is the magnitude of the shift. One predict-error which does not belong to reference range or class C_r = [-Δ; Δ[is considered as a "*non-carrier*". PEHS shift it by -Δ if it is negative or by Δ if it is positive. On the contrary, predict-errors which belong to reference class C_r = [-Δ*;* Δ[are considered as "*carriers*". Such a predict error is used for embedding one bit of the message. Left unchanged it encodes '0', shifted to the range [-2Δ, -Δ[or [Δ, 2Δ[, it encodes '1'.

B. *Dynamical Error Histogram Shifting (DEHS):*

With the PEHS modulation, predict-errors that encode the message belong to the reference class C_r = [-Δ, Δ[, other predict-errors are non carriers. This predicate is static for the whole image and does not take care of the local specificities of the image signal.

Moreover, because prediction acts as a low pass filter, most predict error carriers are located within smooth image regions. Highly textured regions contain non-carriers. The basic idea of our proposal is thus to gain carriers in such a region by adapting the reference class C_r depending on the local context of the pixel or predict-error to be watermarked.

Let us consider the pixel block in figure 1c. Let us also assume that we aim only at watermarking $e_{i,j}/p_{i,j}$ leaving intact $p_{i,j}$ neighborhood. For each of the eight neighbors of $p_{i,j}$: $\{p_{i-k,j-l}\}_{k,l\ =-1\ \ldots 1}$, we can get a predict-error $e_{i-k,j-l}$. However, because $p_{i,j}$ will be modified for insertion, we use $\hat{p}_{i,j}$ instead of $p_{i,j}$ in their calculation (e.g. $e_{i-1,j} = p_{i-1,j} - \hat{p}_{i-1,j}$ with $\hat{p}_{i-1,j} = \left(\hat{p}_{i-2,j} + p_{i-1,j+1} + \hat{p}_{i,j} + p_{i-1,j-1}\right)/4$). Because of the local stationnarity of the image signal we can assume without too much risk that contiguous predict-errors have the same behavior. As a consequence, we suggest to consider as class of reference C_r the histogram range to which belong the absolute values of predict-errors: $\{|e_{i-k,j-l}|\}_{k,l\ =-1\ \ldots 1}$. In the sequel, we propose to use as reference class $C_r = [m_e-\Delta/2, m_e+\Delta/2[$, where m_e is the mean value of $\{|e_{i-k,j-l}|\}_{k,l\ =-1\ \ldots 1}$. As it can be noticed by the reader, DEHS is applied one predict-error absolute value and not on the predict-error directly. The reason stands in the fact that contiguous predict-errors are distributed around the zero value. Thus, their mean results in placing C_r centered on zero. This gives no advantages compared with PEHS. Anyway, as defined, our reference class C_r is determined dynamically for each predict-error of the image. C_r location is also refined independently of $p_{i,j}$, it will be retrieved by the reader. DEHS is more appropriate for regions where the image signal is non-null.

C. Invariant image classification

The purpose of this classification is to identify parts of the image where to apply PHS and DEHS in order to gain in terms of capacity while minimizing image distortion. For medical images this is somewhat equivalent to distinguish the image black background from the anatomical object.

As stated before, our classification process exploits a reference image $\hat{I}$ derived from the image I itself. $\hat{I}$ is a predicted version of I so as to keep image signal properties. The originality of our approach resides in the fact that $\hat{I}$ remains unchanged after I has been watermarked.

Before presenting this process, it is important to notice that in our concern the embedding process is conducted in several pass. In fact, in each pass we consider only a quarter of the pixels. Pixels watermarked in one pass are not re-watermark and at each time a classification process is done.

Let us consider one pass and the set of pixels indentified by a "cross" in figure 1c. In this block, only $p_{i,j}$ will be modified. We recall that shifting $e_{i,j}$ by +/- Δ results in adding/subtracting Δ to $p_{i,j}$.

In order to decide which HS modulation to apply on $p_{i,j}$, we propose to consider its predicted value $\hat{p}_{i,j}$. $\hat{p}_{i,j}$ does not depend on $p_{i,j}$ or on its watermarked version. Consequently, it remains invariant to the insertion process. Thus, in one pass, the reference image $\hat{I}$ contains predict-pixels. Our classification process is then rather simple. Considering a d bit depth image, if $\hat{p}_{i,j} < \Delta$ or $\hat{p}_{i,j} > (2^d-1) - \Delta$, then $p_{i,j}$ belongs to the PHS region otherwise to the DEHS region.

D. PHS and DEHS under/overflow management

Even though PHS and DEHS regions are identified, we still have to face the overflow and underflow issue.

- PHS underflows/overflows

According to the previous classification PHS is applied to two distinct parts of the signal dynamic identified by $\hat{p}_{i,j} < \Delta$ (*low-part*) and $\hat{p}_{i,j} > (2^d-1) - \Delta$ (*high-part*). Because in the low-part (high-part resp.) PHS shifts pixels by adding (subtracting resp.) one gray value; there is no risk of underflow (overflow resp.). This is not the case for the DEHS.

- DEHS underflows/overflows

By definition, DEHS results in adding/subtracting Δ to pixels. Whence, some pixels may lead to an under/overflow if watermarked. To distinguish "*watermarkable*" pixels, i.e. pixels which can be modified, from the others, we propose a second classification process (similar to the one depicted in details in [8]). Let us consider again one block B^k of the image I (see fig. 1c) and one insertion pass. Beside $p^k_{i,j}$, none of its eight nearest neighbors are modified. It is then possible to characterize B^k from its invariant reference block: $\hat{B}^k = \left[\hat{p}^k_{i,j}, p^k_{i-1,j-1}, \ldots, p^k_{i+1,j+1}\right]$ through two characteristics defined as $\hat{B}^k_{\min}$ and $\hat{B}^k_{\max}$ which correspond to the minimum and maximum values of $\hat{B}^k$ respectively. Then, considering the N_o and N_u blocks that if watermarked lead to an overflow or and underflow respectively, we can identify two thresholds T_{min} and T_{max}:

$$T_{min} = max_{\,n=0\ldots Nu}(\hat{B}^n_{\min});\ T_{max} = min_{\,m=0\ldots No}(\hat{B}^m_{\max}) \qquad (1)$$

A block B^k or its corresponding pixel $p^k_{i,j}$ is considered as watermarkable if it satisfies the following constraints:

$$\hat{B}^k_{\min} > T_{min} \text{ and } \hat{B}^k_{\max} < T_{max} \qquad (2)$$

otherwise, it is considered as non-watermarkable and will be left unchanged in the image. Again, classification is conducted on invariant measures; the reader will re-identify easily these non-watermarkable pixels.

E. Proposed scheme

To sum up, our algorithm goes through the image between in one to four times depending on the needs in terms of capacity and distortion. At each pass a quarter of the image pixels are watermarked and a classification process conducted so as to identify PHS and DEHS image regions. For DEHS regions, a second classification process is conducted in order to avoid under/overflows.

In order to minimize the distortion, we also propose two other refinements while preserving the capacity is to not watermark pixels for which the estimator bias is too important and for which the reference class cannot be identified accurately. We thus introduce two more constraints to be satisfied by a DEHS watermarkable pixel. Pixels with high bias belong to blocks which are highly textured. They can be identified through the standard deviation of their block of reference (see section II.D). Thus

$p^k_{i,j}$ is watermarked if $\hat{B}^k_{std} < T_{std}$, where $\hat{B}^k_{std}$ is the standard deviation of $\hat{B}^k$ and T_{std} is a threshold we define in this study as the standard deviation mean of all reference blocks. Again, the reader will retrieve T_{std} and achieve the same classification. In the same vein, pixels with a non accurate reference class C_r, are pixels for which the predict-error neighborhood standard deviation $e^k_{i,jstd}$ is too high. Thus $p^k_{i,j}$ is modified if $e^k_{i,jstd} < T_e$ where T_e corresponds to the mean of { $e^k_{i,jstd}$ } in the whole image.

III. Experiments

A. Image database and measures of performance

The proposed watermarking method has been tested over several test set of medical images issued from three distinct image modalities: three volumes of 12 bits encoded magnetic resonance images (MRI) with 79, 80 and 99 axial slices of 256x256 pixels respectively; three 16 bits encoded PET (positron emission tomography) image volumes of 234, 213 and 212 axial slices of 144x144 pixels respectively; three sequences of 8 bits encoded US (ultrasound image) images (14 of 480x592 pixels, 9 and 30 of 480x472 pixels respectively). We have also considered some well known and common natural test images: Lena and Baboon.

To objectively quantify algorithms' performances, different indicators have been considered: the capacity rate *C* expressed in *bpp* (bit of message per pixel of image) and, in order to quantify the distortion between an image *I* and its watermarked version I_w, the Peak Signal to Noise Ratio:

$$PSNR = 10\log_{10}\left(\frac{NM\left(2^d - 1\right)^2}{\sum_{i,j=1,1}^{N,M}\left(I(i,j) - I_w(i,j)\right)^2}\right) \qquad (5)$$

where *d* corresponds to the image depth, *N* and *M* correspond to the image dimensions.

Δ = 1	use ¼ of the *I*		use ½ of the *I*		use the whole *I*	
	C	*PSNR*	*C*	*PSNR*	*C*	*PSNR*
MRI	0.064	84.99	0.13	81.94	0.25	79.06
([6])	(0.011*bpp* / 74.07*dB*)					
PET	0.088	108.16	0.17	105.35	0.32	102.58
([6])	(0.057*bpp* / 97*dB*)					
US	0.043	62.55	0.084	59.77	0.16	57.067
([6])	(0.2*bpp* / 51.1*dB*)					
Lena	0.04	61.375	0.078	58.545	0.15	55.72
([7])	(0.02)	(61.42)	(0.04)	(58.51)	(0.09)	(55.29)
Baboon	0.0127	63.026	0.025	60.077	0.049	57.167
([7])	(0.005)	(63.66)	(0.01)	(60.46)	(0.02)	(57.11)

Tab. 1. Capacity and distortion measurements for our approach in application to MRI, PET, US, and for Lena and Baboon grayscale images. Results in parenthesis correspond to methods [6] or [7].

B. Experimental results

Results are given in Table 1 in terms of capacity and distortion depending on Δ, the pixel shifting magnitude, and the number of pixel considered for embedding. For medical images, results in terms of capacity and distortion correspond to the mean per image modality. Compare to our previous work in [6], our new scheme allows a watermark capacity close to 0.2 *bpp* with *PSNR* about 79.06 *dB* for MRI, 105.35 *dB* for PET and 57.067 *dB* for US images. Our new approach considers the signal specificities; it is not limited by the black background which occupies a large space in this kind of image.

The last two rows of the table compare our technique with the recent method of Sachnev *et al.* [7] which actually outperforms all other approaches of the literature. Results are given for the grayscale Lena and baboon images. As it can be seen, our method gives a compromise from 0.04 *bpp* / 61.375 *dB* to 0.15 *bpp* / 55.72 *dB* for Lena and 0.0127 *bpp* / 63.026 *dB* to 0.049 *bpp* / 57.167 *dB* for Baboon simply by watermarking successively each quarter of the image. For high PSNR values, our method allows twice the capacities of the Sachnev *et al.* approach.

IV. Conclusion

In this article, we have proposed a new reversible watermarking scheme which originality stands in identifying parts of the image that can be watermarked with two distinct HS modulations: Pixel Histogram Shifting and Dynamical Error Histogram Shifting (DEHS). The later modulation is another original contribution of this work. By considering the specificity of the signal content, our scheme offers a very good compromise in terms of capacity and low distortion for both medical images and natural images. However, this method is fragile and any modifications will impact the watermark. It can serves within applications for verifying the image integrity. However, questions about watermark robustness are largely open. Up to now, a few methods have been proposed. This is one of the upcoming challenges.

Acknowledgment

The work presented in this paper is partly supported by the ANR through the funding of the ANR SELKIS project.

References

[1] G. Coatrieux, L. Lecornu, B. Sankur, and Ch. Roux, "A Review of Image Watermarking Applications in Healthcare," in *Proc. of the IEEE EMBC Conf.*, New York, USA, 2006, pp. 4691–4694.

[2] Mintzer, F., J. Lotspiech, and N. Morimoto, Safeguarding digital library contents and users: Digital watermarking, D-Lib Mag., 1997.

[3] C. W. Honsinger, P. Jones, M. Rabbani, and J. C. Stoffel, "Lossless recovery of an original image containing embedded data," US Patent application, Docket No.:77102/E-D, 1999.

[4] Z. Ni, Y. Shi, N. Ansari, and S.Wei, "Reversible data hiding," in *Proc. IEEE Int. Symp. Circuits and Systems*, May 2003, vol. 2, pp. 912–915.

[5] D. M. Thodi and J. J. Rodriquez, "Expansion Embedding Techniques for Reversible Watermarking,," in *IEEE Trans. Image Processing*, vol.16, no.3, pp. 721-730, March 2007.

[6] W. Pan, G. Coatrieux, N. Cuppens, F. Cuppens and Ch. Roux, "An Additive and Lossless Watermarking Method Based on Invariant Image Approximation and Haar Wavelet Transform," in Proc. of the IEEE EMBC Conf., Buenos Aires, Argentina, 2010, pp. 4740 -4743.

[7] V. Sachnev, H. J. Kim, J. Nam, S. Suresh, and Y.-Q. Shi, "Reversible watermarking algorithm using sorting and prediction," *IEEE Trans. on Circuit Syst. and Video Technol.*, vol. 19, no. 7, pp. 989-999, 2009.

[8] G. Coatrieux, C. Le Guillou, J.-M. Cauvin, C. Roux, "Reversible watermarking for knowledge digest embedding and reliability control in medical images," *IEEE Trans. Inf. Technol. Biomed.*, 2009 Mar., 13(2):158-165.