The wavelet leaders spectrum: a new tool for Blind steganalysis?

Rémi CORNILLET[1,4], Marianne CLAUSEL[2] Farida ENIKEEVA[1], Laurent NAVARRO[3], Philippe CARRÉ[4]

[1]Laboratory of Mathematics and Applications, UMR CNRS 7348, University of Poitiers
BP 30179, 86962 Futuroscope Chasseneuil CEDEX, France

[2]Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France

[3]Mines Saint-Etienne, Univ Lyon, Univ Jean Monnet, INSERM, Centre CIS, Saint-Etienne, France

[4]XLIM Laboratory, UMR CNRS 7348, University of Poitiers, BP 30179, 86962 Futuroscope Chasseneuil CEDEX, France
philippe.carre@univ-poitiers.fr

Résumé – Les techniques de stéganalyses aveugles doivent être capables de détecter la présence d'un message secret enfouit dans un média quelconque comme un son, une image ou une vidéo, et la détection doit se faire sans connaissance a priori sur l'algorithme de stéganographie utilisé. Ce papier se propose d'étudier une nouvelle stratégie basée sur le spectre multifractal comme signature robuste. Nous montrons que le spectre multifractal d'un média corrompu change significativement par rapport au spectre d'origine. Afin de tester cette signature, nous simulons l'attaque associée à un algorithme de stéganographie par modification ponctuelle des bits LSB. Dans ce travail, nous proposons de nous placer dans une stratégie différente de l'état de l'art à savoir que nous privilégions la modélisation à l'apprentissage afin d'apporter des réponses dans des cadres à faible nombre de données.

Abstract – The techniques of blind steganalysis are able to detect the presence of a secret message embedded in digital media files, such as audio, images and video, without any prior knowledge of the steganography algorithm used for embedding. This paper presents a new steganalysis method based on the multifractal spectrum as a robust signature. We test the proposed approach on different data corrupted by an LSB steganography algorithm. In this work, we choose a strategy that differs from the state-of-the-art methods that are mainly based on machine learning. We emphasize the importance of data modelling in contrast to the learning based approaches in case of small training sets.

1 Introduction

Goals of steganalysis. The steganalysis aims at detecting the presence of a message hidden in a cover media using a steganography algorithm. Different approaches have been developed depending whether the embedding algorithm is known or not. In the first case, one usually tries to identify the statistical flaws of the steganography algorithm that has been used to insert the fraudulous message. Having a prior knowledge about the embedding algorithm is, most of the time, unrealistic, which makes important using of the so-called *universal blind steganalysis* methods, designed without any information about the message insertion procedure. In this paper, a new blind steganalysis will be proposed based on the measure of the local regularity of the signal.

State-of-the-art algorithms in the blind universal case. The recent and mostly used steganalysis techniques in the blind universal framework are based on the supervised learning procedures. One first identifies what characterizes a cover media with respect to a steganographed one by extracting the relevant features separating well these two classes (cover and stego). Thereafter, the decision about the class membership is made by using a suitable classifier as the SVM [1], deep learning [2] or the ensemble classifiers [3, 4]. A serious limitation of these methods is the need of storage of large learning databases. In our strategy, we try to replace the learning process by the modelization of local regularity.

A crucial part of an efficient steganalysis algorithm is then the choice of informative features of an image. Usually, these features are based on the wavelet or DCT coefficients (see for example [5, 6, 1] in the image framework). But several alternative approaches are possible. For example, [1] uses the features based on the mean, variance, skewness and kurtosis of the subband coefficients. In [5] the marginal distribution of the wavelet coefficients is modelled using the generalized Gaussian density. Another model proposed in [6] is based on the moments of the characteristic functions of the wavelet coefficients. Notice that a lot of number of method use Wavelet or others transforms for watermarking or steganography process but it is out of scope of this paper (the use of the coefficients is different).

Goals of the present contribution. We propose to use multifractal features in order to detect a steganography attack. Our conjecture is that the insertion of a message perturbs the interscale structure of the media of interest. Consequently, we suggest to use the multifractal attributes based on the so-called wavelet leaders [7, 8] as relevant features to discriminate the cover from the stego. These multifractal attributes describe the interscale structure of a signal or image and have been successfully used in [9] to classify physiological signals or in the Van Gogh challenge [10], to discriminate between different periods of the famous painter.

2 Multifractal attributes of a media

Here we only recall the definition of wavelet leaders and practical estimation procedures of multifractal features introduced in [7, 8].

The wavelet framework. Denote by $X : x \in \mathbb{R}^d \mapsto \mathbb{R}$ the media to be analyzed. We consider here only the two cases $d = 1, 2$ depending whether the analyzed input is a signal or an image. Note that the framework remains clearly valid in the case where $d > 2$. The wavelet coefficients of X are defined as $d_X(i, j, k) = 2^{-dj} \int_{\mathbb{R}^d} X(x)\psi^{(i)}(2^{-j}x - k)dx$. We assume that the mother wavelets are compactly supported and admit at least N vanishing moments, that is $\int_{\mathbb{R}^d} x_1^{\alpha_1} \cdots x_d^{\alpha_d}\psi^{(i)}(x)dx = 0$ for any integers $(\alpha_1, \cdots, \alpha_d)$ such that $\alpha_1 + \cdots + \alpha_d < N$. We also assume that $\{2^{-jd/2}\psi^{(i)}(2^{-j}x-k)\}_{i,j,k}$ forms a basis of $L^2(\mathbb{R}^d)$.

Interscale multiresolution quantities. In the image framework, one defines a dyadic cube as $\lambda_{j,k} = \prod_{\ell=1}^{d}[k_\ell 2^j, (k_\ell + 1)2^j)$ and $3\lambda_{j,k}$ as the union of $\lambda_{j,k+n}$ where $n \in \{-1, 0, 1\}^d$.

The wavelet leaders of X are then defined as $L_X(i, j, k) = \sup_{\lambda' \subset 3\lambda_{j,k}} |d_X(i, j, k)|$. Note that the wavelet coefficients can naturally be arranged over a nested multiscale structure and that the definition of the wavelet leaders at scale j involves all the wavelet coefficients belonging to the tree whose root is $d_X(i, j, k)$. If we insert a fraudulous message at some leaf of the tree, we change all the wavelet leaders located at a coarser level of the hierarchy. This fact is the key ingredient of our methodology. Note that the notion of coarse scale should be numerically limited because the modification associated with the fraudulous message is only limited to the first scales of the decomposition.

Multifractal features. Multifractal features are built from the wavelet leaders and reflect the multiscale properties of the data. The multifractal paradigm assumes that for any $q \geq 0$, the quantity $S_X(j, q) = n_j^{-1} \sum_{k=1}^{n_j} L_X^q(j, k)$ has an asymptotic power distribution of the form $F_q 2^{j\zeta(q)}$ as $n_j := 2^j \to 0$. The so-called multifractal spectrum is related to the structure function ζ by a Legendre transform and summarizes the roughness properties of X. Following [11, 7, 8], to estimate a parametric equation of the multifractal spectrum, one first performs a linear regression of $S_X(j, q)$ on the scale j at a range $\{j_1, \cdots, j_2\}$ which yields some weights $w_{j_1}, \cdots, w_{j_2}$. We then compute

$$h(q) = \sum_{j=j_1}^{j_2} w_j V(j, q) \text{ and } D(q) = d + \sum_{j=j_1}^{j_2} w_j U(j, q)$$

where $V(j, q) = \sum_k R^q(j, k) \log_2 L(j, k)$ and
$U(j, q) = \log_2(n_j) + \sum_k R^q(j, k) \log_2 R^q(j, k)$ with
$R^q(j, k) = L(j, k)^q / \sum_{k'} L(j, k')^q$.

The vector h contains all possible values of the roughness indices of the image at each point and the $D(q)$ the corresponding "size" (in the sense of Hausdorff measure) of the set of points whose roughness index equals $h(q)$.

3 Results and discussions

Now we illustrate by numerical experiments the potential efficiency of multifractal analysis for the detection of steganography attacks. We consider three types of cover data : one-dimensional synthetic data as a fractional Brownian motion, two-dimensional real image data, a real audio file. For each cover we have extracted its multifractal spectrum and look at its change after a steganography attack.

To simulate the insertion of a message in our cover media with a steganography algorithm we modify LSB (least significant bit) of the sample data. The simulation modifies one or two least significant bits of a chosen data entry (with a setting probability). We insert a message with the insertion rate p that is defined as the percentage of changed 8-bit units of information per data file. The computation of the multifractal spectrum is done with a Matlab library that the authors have developed in the context of CNRS project. The proposed library uses a numerical strategy similar to the one proposed by Wendt et al. in [8] with some variations. A detailed explanation and description of the numerical aspects of this work is not a subject of this article but one has to be aware of that the estimated spectrum is sensitive to several numerical factors. For the estimation of the spectrum, we propose to use Daubechies Wavelet with 8 vanishing moments. We have tested different decompositions : decimated or undecimated, orthogonal or biorthogonal, based on periodic convolution product or based on symmetric convolution product (for the border effect) in order to obtain the more robust estimation. For this paper, we propose to use the simple form of the decomposition : orthogonal decimated transform based on periodic convolution product. The number of levels of the wavelet decomposition depends on the length of the analyzed signal. It is necessary to use an important number of scales taking into account the length of the used filters in order to limit the effect of the boundary.

In order to analyze the influence of the LSB steganography algorithm, we first propose to consider a synthetic signal as an input : a fractional Brownian motion (fBm) $W_H(t)$, $t \in [0, 1]$ for different values of the Hurst parameter. The evolution of the spectrum is associated with the local regularity of the signal and the hypothesis proposed in this paper is that the steganography modification disturbs this local regularity and, consequently, disturbs the multifractal spectrum as well. We present the spectrum of the initial cover data and the spectrum of the simulated steganographed cover for different insertion rate $p \in \{0.1, 0.2 \ldots, 0.9\}$. Figure 1 shows the evolution of the spectrum when inserting a message.We have used a fractional Brownian motion with $H = 1/2$ with amplitude values spread on a full 8 bits scale (the amplitude of the signal is scaled to the interval $Z_{[0,255]}$). The insertion was performed by modification of two last bits of randomly chosen points of the data with the insertion rate p. The figure shows that there is a significant modification of that right-hand side of the spectrum when the insertion rate is significantly high (greater that 0.4). The spectrum of a modified data has a larger support shifted to the left. The modification of a significant number of samples increases the "irregularity" of the signal and thus reduces the part of the spectrum associated with uniform area.

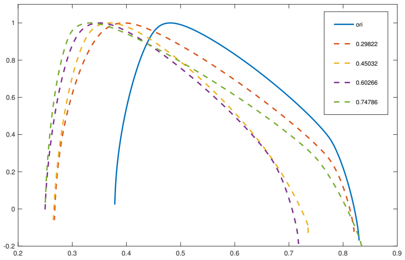

FIGURE 1 – The spectrum of fBm $W_H(t)$, $H = 1/2$, for different values of p, the original spectrum is in blue.

This first numerical result justifies our hypothesis : the evolution of the multifractal spectrum based on wavelet leaders allows to detect the presence of an inserted message. We have next tested our approach on the real data. In the first example we have inserted a message in a selected line of a real image. Figure 2 shows the image from which the line was extracted (associated with a highly textured content), as well as the luminance of one of the lines of the image that was modified.

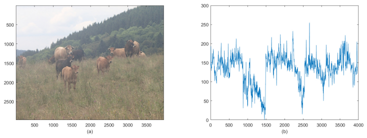

FIGURE 2 – The image (on the left) and the extracted cover media line (on the right).

Let us note that the insertion would be generally done only within a certain well localized area. It would be interesting to detect a significant irregularity in the spectrum with respect to the untouched parts. To simulate this setting, an area of 512 adjacent pixels on the line was modified by the LSB algorithm (with $p = 0.8$). Then the line was analyzed within a sequence of overlapping sliding windows with 50% overlap. Figure 3 below illustrates the obtained spectra. The values in the legend stand for the consecutive window numbers, the negative value (-4) corresponds to the window that contains mostly the steganographed area. We can clearly see the difference in the behavior of the corrupted interval spectrum. One can observe the change in the right part of the spectrum when the number of modified pixels increases. We can make the same conclusion as for the synthetic example : the application of a steganography algorithm perturbs the local regularity which results in the perturbation of the part of the spectrum associated with the "regular" area. The next step of our work will be an automatic detection of steganographic attacks of some local area. The detection algorithm is based on the following idea : we will track the evolution of the spectrum along adjacent windows and detect significant differences in the spectra using an appropriate statistical test.

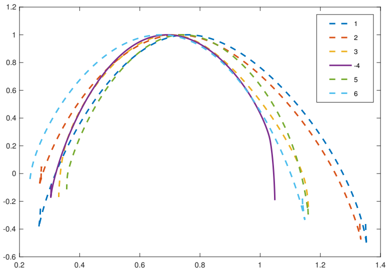

FIGURE 3 – The image line spectrum calculated within overlapping windows of size 512. The violet line shows the spectrum of the corrupted area. Each curve is associated with an interval, and the negative number for a window indicates that the interval is associated with modified samples.

Finally we illustrate our approach applied to the audio file. We have inserted a message into a 8192-length interval of a record of a single note on an 8-bit dynamic by modifying 2 bits with the insertion rate $p = 0.8$ for the interval. Figure 4 shows the steganographed sound.

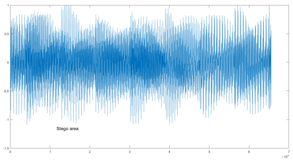

FIGURE 4 – The steganographed audio record.

Again, the sound was analyzed within half-size overlapping sliding windows. In order to obtain a more visual detection of the stego part of the signal, we propose a representation like the Time-Frequency representation, the evolution of the spectrum along the time with an image coding. In this case, the value of the spectrum for each interval is recorded on the gray value of the pixel, and the time coordinate of each block is the time coordinate of the middle of each window. Figure 5 illustrates the obtained representation, in this experimentation we use smallest windows (equal to 1024 samples). One can observe that this representation permits one to detect a significant change in the behavior of the spectrum associated with the stego area and thus to decide to do not guarantee the integrity of the data recorded. The "shift" of the spectrum around the time position 0.5 is only due to the message insertion. The spectrum of the original signal in this time area is close to the spectrum of intervals before and after.

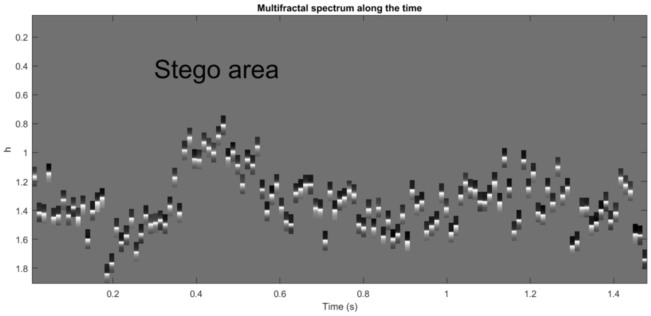

FIGURE 5 – The spectrum of audio record calculated within half-size overlapping windows of size 1024 : each block is associated with an interval, the x-axis is asscoiated with time (seconds), the y-axis is the h value, and the color value is associated with the value of the different multifractal spectrum.

4 Conclusion

In this paper, we presented a blind steganalysis scheme based on multifractal attributes and provided an evidence of a significant change in the interscale structure of an attacked image. We have used the multifractal spectrum estimation strategy based on the wavelet leaders as proposed by Wendt et al. In this first work, we tested our approach when applied to the detection of modifications made by the simplest steganography algorithm (LSB) that modifies the Least Significant Bit of the data entries at a given insertion rate. The numerical results obtained on synthetic signals, audio signals and image data justify our hypothesis that the insertion of a message affects the local regularity of the cover signal and thus modifies the right side of the multifractal spectrum, that corresponds to the regular part of the signal. Moreover, in this paper, we propose an original strategy of detecting the presence of an inserted message that is based on comparison of the spectra calculated withing sliding windows. The future work is to propose a relevant test statistic that will be able to detect the change in the multifractal spectrum and to study its theoretical properties. We will also focus on the application of the proposed approach to detecting a presence of a message in a sequence of image frames like a video file.

Références

[1] S. Lyu and H. Farid, "Steganalysis using higher-order image statistics," *Trans. Info. For. Sec.*, vol. 1, no. 1, pp. 111–119, Nov. 2006.

[2] L. Pibre, J. Pasquet, D. Ienco, and M. Chaumont, "Deep learning for steganalysis is better than a rich model with an ensemble classifier and is natively robust to the cover source-mismatch," in *Proceedings of Media Watermarking, Security, and Forensics*, 2016.

[3] B. Chen, G. Feng, and F. Li, "Steganalysis in high-dimensional feature space using selective ensemble classifiers," in *Communications in Computer and Information Science*, pp. 9–14. Springer Nature, 2012.

[4] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, apr 2012.

[5] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *Trans. Img. Proc.*, vol. 11, no. 2, pp. 146–158, Feb. 2002.

[6] G. Xuan, "Steganalysis based on multiple features formed by statistical moments of wavelet characteristic functions," pp. 262–277, 2005.

[7] S. Jaffard, B. Lashermes, and P. Abry, "Wavelet leaders in multifractal analysis.," in *Wavelet Analysis and Applications,*. Birkhkauser Verlag,, 2006.

[8] H. Wendt, S. Roux, S. Jaffard, and P. Abry, "Wavelet leaders and bootstrap for multifractal analysis of images," *Signal Process.*, vol. 89, no. 6, pp. 1100–1114, June 2009.

[9] J. Spilka, J. Frecon, R. Leonarduzzi, N. Pustelnik, P. Abry, and M. Doret, "Intrapartum fetal heart rate classification from trajectory in sparse svm feature space," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 2335–2338.

[10] P. Abry, H. Wendt, and S. Jaffard, "When van gogh meets mandelbrot : Multifractal classification of painting's texture," *Signal Processing*, vol. 93, no. 3, pp. 554 – 572, 2013, Image Processing for Digital Art Work.

[11] A.B. Chhabra, C. Meneveau, R.V. Jensen, and K.R. Sreenivasan, "Direct determination of the f(α) singularity spectrum and its application to fully developed turbulence," *Physical Review A*, vol. 40, no. 9, pp. 5284, 1989.