Steganalysis of Semi-fragile Watermarking Systems Resistant to JPEG Compression

Anna Egorova[1] and Victor Fedoseev[1,2]
[1]Samara National Research University, Samara, Russia
[2]Image Processing Systems Institute, Branch of the Federal Scientific Research Centre "Crystallography and Photonics" of Russian Academy of Sciences, Samara, Russia

Keywords: Image Protection, JPEG, Semi-fragile Watermarking, Targeted Steganalysis, LSB, QIM.

Abstract: Recently, dozens of semi-fragile digital watermarking systems have been designed to protect JPEG images from unauthorized changes. Their principle is to embed an invisible protective watermark into the image. Such a watermark is destroyed by any image editing operations, except for JPEG compression with the quality level in a given range of values. Watermarking systems of this type have been assessed in terms of watermark extraction accuracy and visual quality of the protected image. However, their steganographic security (i.e., robustness against detecting protective information traces by a third party) has not been sufficiently studied. Meanwhile, if an attacker detects the presence of a watermark in the image, he can get valuable information on the used image protection technique. It can let him develop a data modification method that alters the content of the protected image without destroying the embedded watermark. In this paper, we propose a specific attack to analyze steganographic security of known semi-fragile watermarking algorithms for JPEG images. We also investigate the efficiency of the proposed attack. In addition, we propose an approach to counter the attack that can be applied in the existing watermarking systems to enforce their steganographic security.

1 INTRODUCTION

At present, the role of visual information (in particular images) has increased in various areas of the digital economy: e-commerce, medicine, education, etc. Consequently, image authentication and malicious change detection in images have become the tasks of great importance. Note that images are mostly compressed in practice. For this reason, distortions caused by JPEG, JPEG 2000, and other lossy compression formats are often considered as legal modifications. For compressed images authenticating, semi-fragile watermarking systems can be used (Cox, 2008). In this paper, we consider the watermarking systems that are robust only to JPEG compression. They embed a watermark (security information) into the image immediately after image registration. Such the watermark has the property to be preserved after JPEG compression, but it is destroyed after any other modifications of the image. The performance of such watermarking systems has been proven (Egorova and Fedoseev, 2019), but their steganographic security (the ability to detect watermark traces by a third party) has not previously been examined. Meanwhile, if an intruder detects the presence of such an embedded watermark in the image, he can get information on the used image protection system. Thereby, by using this information, he can develop a data alteration method that does not change the protective watermark but distorts the image content.

In this study, we model a specific attack to analyze steganographic security of various semi-fragile watermarking systems designed for the JPEG compression standard (Lin and Chang, 2000; Mursi et al., 2009; Preda and Vizireanu, 2015; Fallahpour and Megias, 2016; Egorova and Fedoseev, 2019). The need for a new attack is caused by the fact that the existing targeted attacks for JPEG steganography methods (JSteg, F5, Model-based, etc.) (Fridrich, 2010) do not fit the semi-fragile embedding concept.

The essence of the proposed attack is that the number and the distribution of both odd and nonzero quantized DCT coefficients can be used as significant features to detect a watermark, i.e., to separate original images from watermarked ones.

Egorova, A. and Fedoseev, V.
Steganalysis of Semi-fragile Watermarking Systems Resistant to JPEG Compression.
DOI: 10.5220/0009129708210828
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 821-828
ISBN: 978-989-758-402-2; ISSN: 2184-4321

The study also answers the following questions:

- A) How efficient is the proposed attack?
- B) Which of the existing JPEG semi-fragile watermarking systems is more resistant to the attack?
- C) How to adjust the watermark embedding procedures, which are commonly used in existing JPEG semi-fragile watermarking systems to protect these systems against the specific attack?

The rest of the paper is organized as follows. Section 2 provides a brief description of the data embedding techniques that are commonly utilized in the JPEG-resistant watermarking systems. Section 3 introduces a method for targeted steganalysis of these systems. Section 4 presents the results of conducted numerical experiments and answers questions A and B, while Section 5 responds to question C and suggests a modification of watermark embedding procedure that protects the considered watermarking systems against the proposed method for targeted steganalysis.

2 SEMI-FRAGILE WATERMARKING SYSTEMS FOR JPEG

Let us consider the main steps of the lossy JPEG compression standard for grayscale images (see Figure 1). First, a source image I is transformed by 8×8 block discrete cosine transform (DCT), resulting in coefficients $B_i(j)$, where i is an index of 8×8 block and $j = 1..64$ is a DCT coefficient index in zig-zag scanning. Then each $B_i(j)$ is divided by a 8×8 quantization matrix $Q_{QF}(j)$ element-wise, where the index QF is the compression quality factor. Then the quotients are quantized, resulting in values $D_i(j)$. Finally, entropy coding of $D_i(j)$ is performed.

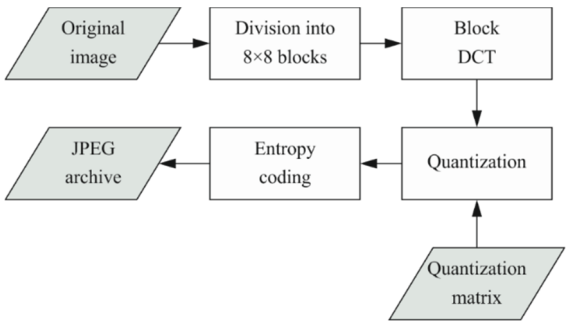

Figure 1: JPEG image compression scheme.

Existing JPEG semi-fragile watermarking systems embed the watermark by modifying either DCT $B_i(j)$ or quantized DCT $D_i(j)$ coefficients. Let N_W be a number of watermark bits to be embedded in each image block, and j_k be the position of the DCT coefficients in zig-zag scanning to be watermarked, where $k = 1..N_W$.

The simplest way to embed the watermark during JPEG compression process is to change the least significant bits of the quantized DCT coefficients (LSB method) (Barni and Bartolini, 2004):

$$D_i^W(j_k) = 2\lfloor D_i(j_k)/2 \rfloor + W_{i,k}, \quad (1)$$

where $W_{i,k}$ is the k-th bit of information embedded in the i-th block. This procedure is implemented in the system proposed in (Lin and Chang, 2000).

Another embedding approach is quantization index modulation (QIM) (Chen, 2001) that simultaneously quantizes DCT components and embeds the watermark. QIM may be implied in various forms. For example, in (Preda and Vizireanu, 2015), the authors use the following embedding rule:

$$B_i^W(j_k) = round\left(\frac{B_i(j_k)}{2Q_{QF}(j_k)} - W_{i,k}\right)2Q_{QF}(j_k) + W_{i,k}Q_{QF}(j_k). \quad (2)$$

Another QIM-based system is "Sign-QIM" proposed in (Egorova and Fedoseev, 2019). In this system, the watermark embedding is performed according to these equations:

$$B_i^W(j_k) = Br_i(j_k) + S_i(j_k)\cdot W_{i,k}\cdot Q_{QF}(j_k), \quad (3)$$

$$Br_i(j_k) = round\left(\frac{B_i(j_k)}{2Q_{QF}(j_k)}\right)2Q_{QF}(j_k), \quad (4)$$

$$S_i(j_k) = \begin{cases} 1, & B_i(j_k) \geq Br_i(j_k), \\ -1, & else. \end{cases} \quad (5)$$

There are many other JPEG semi-fragile watermarking systems based on LSB (Ho and Li, 2004; Huang, 2013) or QIM (Wang et al., 2011; Fan et al., 2011; Ye et al., 2003). They are described in detail in (Egorova and Fedoseev, 2019).

A system using another embedding algorithm is proposed in (Mursi et al., 2009). It is based on the mapping table approach. Such a table is randomly generated using a secret key. It defines a mapping between values of $D_i(j)$ and the set $\{0,1\}$. Suppose a bit "1" needs to be embedded in some coefficient. If its value corresponds to "1" in the mapping table, it does not change. Otherwise, the value is replaced by the nearest number corresponding to "1" in the table.

The system (Fallahpour and Megias, 2016) associates a watermark bit with the parity of the last nonzero (LNZ) $D_i(j)$ coefficient position. That is, this system cannot embed more than one bit per block. However, in our study, we use a generalized version of this system that embeds up to 4 bits per block.

Thus, in this study, we investigate the five watermarking systems (Lin and Chang, 2000; Mursi et al., 2009; Preda and Vizireanu, 2015; Fallahpour and Megias, 2016; Egorova and Fedoseev, 2019) implementing the four most common embedding approaches used in semi-fragile watermarking: LSB, QIM, LNZ, and mapping tables. In the experimental part, when we select the DCT coefficients in positions specified in "original" papers for watermarking, we call the selecting mode "original". To check different embedding approaches in the same conditions, we also test a "sequential" mode, where the first N_W AC coefficients in each block are modified. This mode reduces distortions in the watermarked image and increases the PSNR metric (Egorova and Fedoseev, 2019). In all the experiments, we use N_W values from the set $\{1, 2, 4\}$ and $QF = 50$.

3 PROPOSED TARGETED ATTACK AGAINST JPEG SEMI-FRAGILE WATERMARKING

Most of the quantized DCT coefficients D_i generated at the JPEG compression process are equal to zero. When the image is watermarked by using the system based on LSB, QIM, or the mapping table, the statistics of even and odd DCT coefficients of a block are leveled. Given this, we supposed that the number of odd and the number of nonzero coefficients per the quantized DCT block might be the significant features and can be used to detect the embedded information in JPEG images.

To test this hypothesis, we plotted two graphs (see Figure 2). The top plot shows the number of odd coefficients among the first *j* coefficients of a block in the zig-zag scanning. The lower plot illustrates the same statistics of the nonzero coefficients. The plots show an average result for 1000 halftone images obtained by each of the five selected watermarking systems using the "original" mode of coefficients selection. Black lines in Figure 2 correspond to 50 different host images for clarity. It can be seen that the scatter of the nonzero and odd statistics is quite significant. It means that the total numbers of nonzero and odd values in each block (the values on the graphs corresponding to *j=64*) are not sufficiently reliable signs of the watermark presence. However, the protected images generally have a higher growth rate at low *j* indices, since low-frequency coefficients are usually used for watermarking in order to reduce visual distortions.

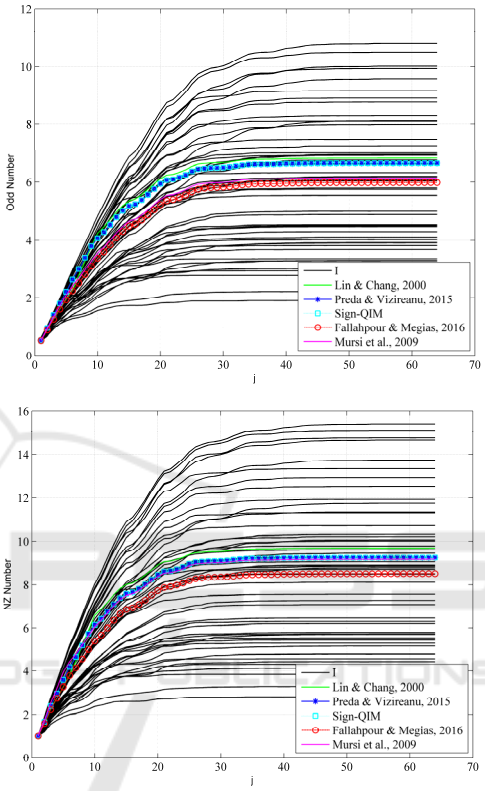

Figure 2: The average number of odd (up) and nonzero (down) DCT coefficients among the first j positions in 8×8 watermarked image block obtained by using "original" coefficient selection mode, compared with the same plots for 50 host images (black lines marked as I).

The same results are more clearly shown in Figure 3. Here instead of 50 lines for host images, only the one – averaged over 1000 images is shown (black line). To save space, in Figure 3, we show only graphs of odd statistics. Figure 3 also presents the result for the "sequential" mode of coefficients selection. The figure shows that the watermarked images have a higher growth rate at low *j*. Note that the "sequential" mode decreases this feature. Figure 3 illustrates that the line for the system (Fallahpour and Megias, 2016) is indistinguishable from the line for the empty containers. This is because the system does not change the values of the coefficients. Instead, it swaps some values. Therefore, in relation to our attack, this

system is secure. However, there is a much simpler and more effective attack for this system: testing the hypothesis that odd and even positions of LNZ are equally probable. It can be done by applying the statistical methods, for example, by calculating chi-square statistics (Cox, 2008).

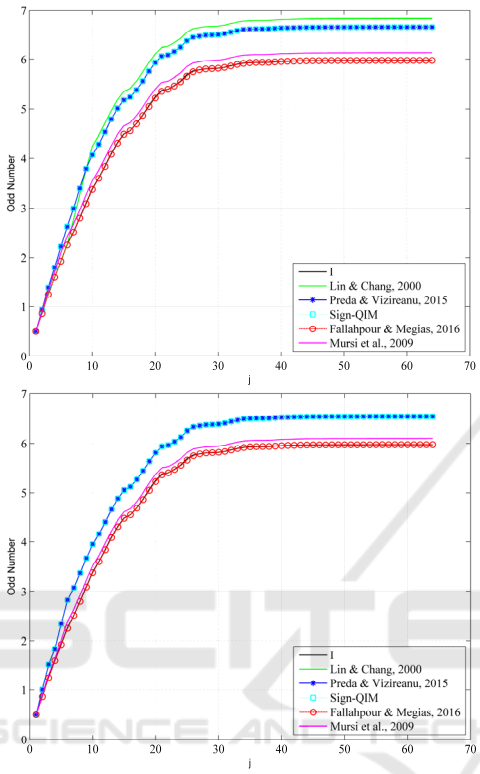

Figure 3: The average number of odd DCT coefficients among the first j positions in 8×8 source (I) and watermarked image block obtained by using "original" (up) and "sequential" (down) coefficients selection modes.

4 EXPERIMENTAL INVESTIGATION OF THE PROPOSED ATTACK

4.1 Model and Feature Selection

Based on the analysis of the results shown in Figure 3, we developed a targeted attack in the form of a linear SVM classifier, which determines whether a given image contains a watermark or not. This classifier operates with the following feature set:

$$(NZN(4)-NZN(1),\ NZN(9)-NZN(4),\ NZN(64),\ AUNZN,\ ON(4)-ON(1),\ ON(9)-ON(4),\ ON(64),\ AUON),$$

where $NZN(j)$ is the number of nonzero coefficients among the first j, $ON(j)$ is the corresponding number of odd coefficients, $AUNZN$ is the area under the curve of nonzero values, and $AUON$ is the area under the curve of odd values.

Differential features reflect the growth dynamics of $NZN(j)$ and $ON(j)$ among the most important DCT coefficients. $NZN(64)$ and $ON(64)$ show the total values of the measured statistics. The area features represent both dynamics and total numbers.

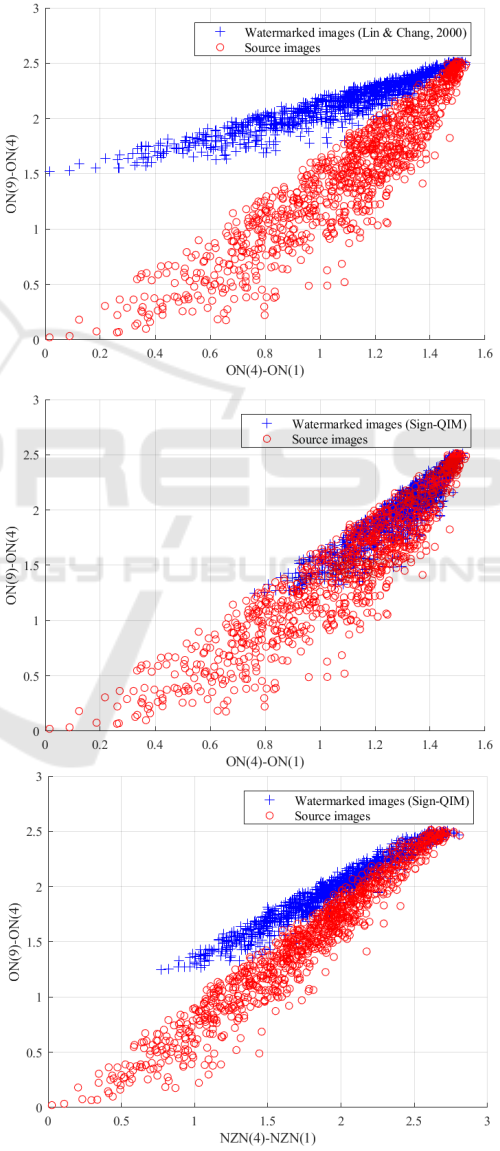

Figure 4: Importance of different features in detecting the watermarked images (obtained using the "original" coefficients selection mode).

In the preliminary tests, we analyzed the significance of single features from the selected feature set using the sequential forward selection procedure (Marcano-Cedeño et al., 2010). These tests showed that the differential features significantly outperform all others at all tested N_W values. Diagrams in Figure 4 confirm this property and show that the nonzero features and the odd features vary in importance for different watermarking algorithms.

However, in the experiments described below, we use the full 8-feature set, as well as two 4-feature subsets of nonzero and odd features for comparison. This was done in order to prevent the exclusion of potentially significant data.

4.2 Performance Evaluation of the Developed Attack

To investigate the effectiveness of the developed attack, we used 1000 images of size 512×512 from the BOWS-2 dataset (Westfeld, 2009). By applying each of the five selected systems, we generated 1000 images with watermarks. 70% of the resulting 2000 images were used for training the classifier, and 30% for testing.

The results on classification accuracy obtained using both "original" and "sequential" modes of coefficients selection at the different N_W are presented in Table 1. The higher the accuracy value, the higher the probability of a successful attack.

As predicted, the system (Fallahpour and Megias, 2016) is completely secure against the attack at any N_W. Other systems can be detected. If the "original" mode is used for coefficients selection, the least resistant system is (Lin and Chang, 2000) based on LSB. The detection accuracy for this system is 73 – 100%, depending on N_W. Three other systems show very similar results and can be detected with accuracy higher than 69% for $N_W > 1$. When using the "sequential" mode, all the accuracy values become lower, and the detection accuracy could be rated as acceptable only at $N_W = 4$. The highest values are gained by the system (Mursi et al., 2009).

Table 1: Accuracy of different systems detection by the developed attack (full feature set). The higher the accuracy value, the higher the probability of a successful attack.

N_W	1			2			4		
Positions selection method	Original	Sequential	Adaptive-2	Original	Sequential	Adaptive-2	Original	Sequential	Adaptive-2
Lin & Chang, 2000	0.73	0.56	0.54	0.97	0.61	0.53	1.00	0.69	0.63
Preda & Vizireanu, 2015	0.62	0.57	0.54	0.69	0.64	0.58	0.85	0.67	0.77
Sign-QIM	0.65	0.57	0.55	0.72	0.65	0.67	0.85	0.72	0.76
Fallahpour & Megias, 2016	0.50	0.51	0.51	0.50	0.51	0.50	0.50	0.51	0.51
Mursi et al., 2009	0.64	0.63	0.58	0.78	0.52	0.56	0.84	0.72	0.62

Table 2: Accuracy of different systems detection by the developed attack (odd features only). The higher the accuracy value, the higher the probability of a successful attack.

N_W	1			2			4		
Positions selection method	Original	Sequential	Adaptive-1	Original	Sequential	Adaptive-1	Original	Sequential	Adaptive-1
Lin & Chang, 2000	0.79	0.54	0.50	0.90	0.60	0.50	0.95	0.80	0.50
Preda & Vizireanu, 2015	0.66	0.54	0.51	0.78	0.59	0.51	0.83	0.79	0.50
Sign-QIM	0.69	0.53	0.52	0.71	0.60	0.52	0.86	0.74	0.55
Fallahpour & Megias, 2016	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
Mursi et al., 2009	0.57	0.51	0.51	0.59	0.54	0.51	0.60	0.57	0.50

Table 3: Accuracy of different systems detection by the developed attack (nonzero features only). The higher the accuracy value, the higher the probability of a successful attack.

N_W	1			2			4		
Positions selection method	Original	Sequential	Adaptive-1	Original	Sequential	Adaptive-1	Original	Sequential	Adaptive-1
Lin & Chang, 2000	0.72	0.57	0.51	0.93	0.59	0.52	0.99	0.67	0.58
Preda & Vizireanu, 2015	0.61	0.54	0.50	0.65	0.58	0.52	0.82	0.64	0.55
Sign-QIM	0.61	0.53	0.51	0.65	0.57	0.53	0.81	0.65	0.55
Fallahpour & Megias, 2016	0.51	0.50	0.51	0.51	0.51	0.50	0.51	0.51	0.50
Mursi et al., 2009	0.62	0.53	0.51	0.75	0.57	0.50	0.78	0.65	0.55

Table 4: Average PSNR of secured images after watermark embedding by the investigated systems.

N_W	1			2			4		
Positions selection method	Original	Sequential	Adaptive-2	Original	Sequential	Adaptive-2	Original	Sequential.	Adaptive-2
Lin & Chang, 2000	41.6	44.0	39.9	39.2	41.0	36.9	36.6	38.0	33.8
Preda & Vizireanu, 2015	45.7	46.0	40.5	42.3	43.0	37.4	39.1	40.0	34.3
Sign-QIM	49.3	49.5	42.1	45.9	46.7	39.1	42.6	43.7	35.9
Fallahpour & Megias, 2016	35.4	35.3	35.3	35.4	35.2	35.2	35.4	35.3	35.3
Mursi et al., 2009	47.6	47.5	36.1	44.3	42.0	38.6	40.8	41.6	30.3

Tables 2-3 show that the detection accuracy reduces if either nonzero only or odd only features are selected for classification in the attack. Overall, the obtained results claim the effectiveness of the proposed attack in some range of conditions.

5 MODIFYING THE WATERMARKING SYSTEMS TO MAKE THE ATTACK LESS EFFECTIVE

In this section, we suggest two methods for adjusting the watermark embedding procedures in the existing systems that allow enforcing the security of the systems against the proposed attack. To make an adjustment method universal, we modify the mode for selecting positions of DCT coefficients. The idea is to select a coefficient in a way that makes the nonzero and odd curves after the embedding (see Figure 3) close to graphs of host images. To get this effect, we generate these positions randomly. The probability of each position selection is proportional to the numerical derivative of the host image lines shown in Figure 3. This mode we name "Adaptive-1". It is investigated in the same conditions as other positions selection modes. Tables 2-3 present the classification accuracy when only nonzero or only odd features were used. It can be seen that the "Adaptive-1" mode is more secure than other modes because it provides lower accuracy. However, the full feature set does not provide lower classification accuracy. To save space, these results are not specified in Table 1.

To overcome the disadvantage of the "Adaptive-1", we modify this method slightly. In addition to watermark insertion in some positions, we replace with zero some existing nonzero coefficients. Their positions are also selected according to the numerical derivative of the host image nonzero line. The number of such coefficients is proportional to N_W . The experimental results obtained for this mode are shown in Table 2 in the column "Adaptive-2". These data show that the accuracy becomes lower for all cases except for the two QIM-based systems for $N_W = 4$. Thus, this mode has reasonable potential.

It is clear that zeroing some DCT coefficients reduces the watermarked image quality. To estimate the quality loss, we measured PSNR in all the cases tested earlier. Table 4 contains the obtained numerical results, while Figure 5 shows some examples of images protected using all the analyzed watermarking systems in combination with "Adaptive-2" at $N_W = 4$. The numbers and shown images prove that visual quality is reduced. However, the loss is not so dramatic, and the quality of the watermarked images can be recognized as acceptable in many applications.

6 CONCLUSIONS

In this paper, at first, we proposed a new targeted steganographic attack against semi-fragile watermarking systems designed for the JPEG compression standard. The goal was to analyze the steganographic security of such systems. The attack consists of the calculation of the nonzero and the odd DCT coefficients statistics, selection of significant features, and training an SVM classifier. We showed logically and experimentally the significance of the selected features for the given problem. To investigate the efficiency of the developed attack, we applied it for five different watermarking systems and measured the classification accuracy for different watermark length. The systems were tested in both "original" and "sequential" modes of coefficients selection. The results showed that the attack is effective for all the systems except one (Fallahpour and Megias, 2016) (but this system has more crucial security problems) when more than one bit per block is embedded. "Sequential" mode has been found more secure than "original", in addition to its higher quality investigated in (Egorova and Fedoseev, 2019) and also shown in Table 4. In addition, in this paper, we proposed and investigated two methods to counter the developed attack, i.e., ways to make the systems more secure.

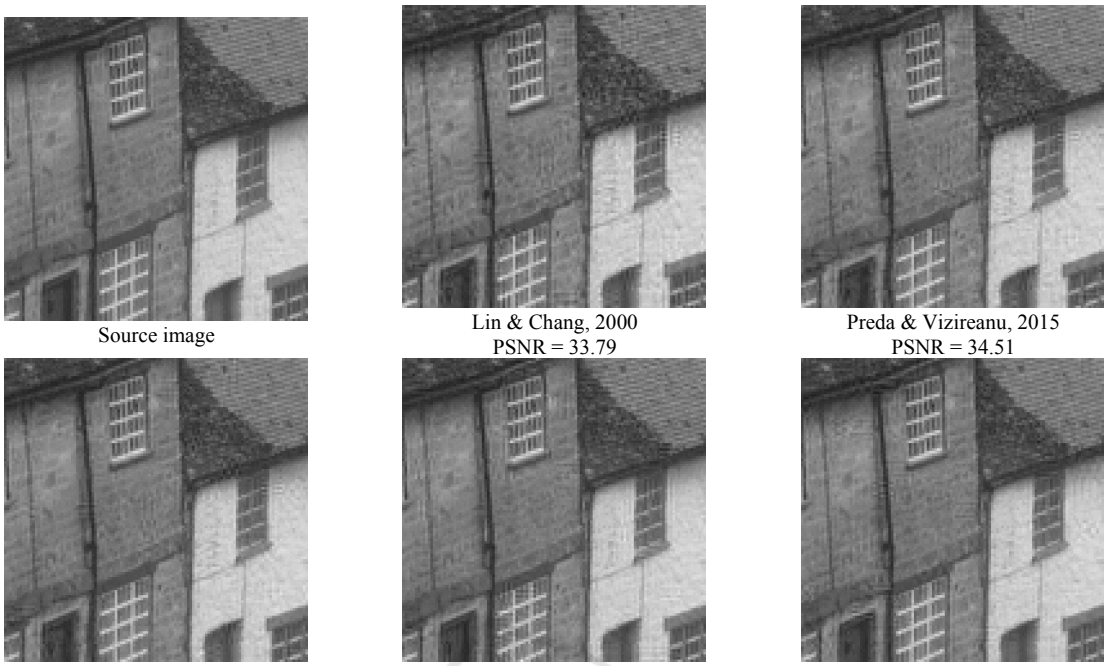

Figure 5: Examples of secured images protected using the proposed "Adaptive-2" positions selection method at $N_W = 4$.

These methods consist in specific modes of coefficients selection, "Adaptive-1" and "Adaptive-2", and allow obtaining more "natural" DCT coefficients statistics. The second method, "Adaptive-2", in addition to adaptive coefficients selection, zeroes some values. The experiments show that "Adaptive-1" improves security when the attack is applied using shortened feature sets (nonzero only features, or odd only features). However, for the full set, its effect is not so clear. The other method undoubtedly makes the systems more secure but reduces the visual quality of the protected images (by 5 dB on average in terms of PSNR). However, if we use the Sign-QIM system (Egorova and Fedoseev, 2019), which is the best in visual quality, we can still obtain 36 dB in the case of 4 bit per block watermarking and 42 dB in the case of 1 bit per block watermarking.

Overall, the paper draws the attention of researchers to the problem of steganographic security of semi-fragile watermarking systems and gives some practical methods to improve it.

ACKNOWLEDGMENTS

The work was funded by the RSF grant #18-71-00052.

REFERENCES

Barni, M. and Bartolini, F. (2004). *Watermarking systems engineering: Enabling digital assets security and other applications*, CRC Press.

BOWS-2. *Image Repository*. (n.d.). Retrieved from http://bows2.ec-lille.fr/.

Chen, B. and Wornell, G. (2001). Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. *IEEE transaction on Information Theory*, 47(4). 1423–1443. https://doi.org/10.1109/18.923725.

Cox, I., Miller, M., Bloom, J., Fridrich, J., and Kalker, T. (2008). *Digital watermarking and steganography,* Elsevier.

Egorova, A. A and Fedoseev, V. A. (2019). A classification of semi-fragile watermarking systems for JPEG images. *Computer Optics*, 43(3), 419–433. https://doi.org/10.18287/2412-6179-2019-43-3-419-433.

Fallahpour, M. and Megias, D. (2016). Flexible image watermarking in JPEG domain. In *2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 311–316. https://doi.org/10.1109/ISSPIT.2016.7886055.

Fan, C. H., Huang, H. Y., and Hsu, W. H. (2011). A robust watermarking technique resistant JPEG compression. In *Journal of Information Science and Engineering*, 27(1), 163–180.

Fridrich, J. (2009). *Steganography in digital media: Principles, algorithms, and applications*. Cambridge University Press.

Ho, C. K. and Li C. T. (2004). Semi-fragile watermarking scheme for authentication of JPEG images. In *Proceedings of the International Conference on Information Technology: Coding and Computing*, 1, 7–11. https://doi.org/10.1109/ITCC.2004.1286417.

Huang, L. Y. (2013). Authentication watermarking algorithm resisting JPEG compression based on preliminary quantization. In *Information Technology Journal*, 12(6), 3723–3728. https://doi.org/10.3923/itj.2013.3723.3728.

Lin, C. Y. and Chang, S. F. (2000). Semifragile watermarking for authenticating JPEG visual content. In *Electronic Imaging*. 140–151. https://doi.org/10.1117/12.384968.

Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, 2845–2850. https://doi.org/10.1109/IECON.2010.5675075

Mursi, M., Assassa, G. M. R., Aboalsamh, H., and Alghathbar, K. (2009). A DCT-based secure JPEG image authentication scheme. *World Academy of Science, Engineering and Technology*, 53, 681–687. https://doi.org/10.5281/zenodo.1070359.

Preda, R. O. and Vizireanu, D. N. (2015). Watermarking-based image authentication robust to JPEG compression. *Electronics Letters*, 51(23), 1873–1875. https://doi.org/10.1049/el.2015.2522

Wang, H., Ho, A., and Zhao, X. (2011). A novel fast self-restoration semifragile watermarking algorithm for image content authentication resistant to JPEG compression. *Digital Forensics and Watermarking*, 7128. 72–85. https://doi.org/10.1007/978-3-642-32205-1_8.

Westfeld, A. (2009). Fast Determination of Sensitivity in the Presence of Countermeasures in BOWS-2. *Information Hiding* (pp. 89–101). Springer. https://doi.org/10.1007/978-3-642-04431-1_7.

Ye, S., Zhou, Z., Sun, Q., Chang, E., and Tian, Q. (2003). A quantization based image authentication system. In *Fourth International Conference on Information, Communications and Signal Processing*, 2003 *and the Fourth Pacific Rim Conference on Multimedia,* 955–959. https://doi.org/10.1109/ICICS.2003.1292599.