UDC 004.[056.54:932.2]

Destruction of stego images formed by adaptive embedding methods with dictionary learning methods

Dmytro Progonov[1, a]

[1] *National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Institute of Physics and Technology*

Abstract

Counteraction to sensitive information leakage that processed by state and private organizations is topical task today. Of special interest are methods for prevention data leakage by usage of hidden (steganographic) communication channels by attackers. Despite wide range of proposed steganalysis methods for detection of embedded messages, theirs performance highly depends on prior information about used embedding methods. As an example, we may mention modern stegdetectors for digital images, which are based on cover rich models and deep convolutional neural networks. Therefore, the stego image destruction methods are widely applied as preventive action.

Modern methods for stego image destruction are based on widespread image denoising methods, like median filter and lossy compression. The limitation of such methods is significant changes of image's statistical features that may disclosure the steganalysis process to attacker. Therefore, development of stego images processing methods that provide reliable destruction of embedded data, and preserving cover image statistical features is needed. The paper is aimed at performance evaluation of applying the novel methods of spectral analysis, namely dictionary learning, for solving this tasks. The obtained results showed limitation of state-of-the-art methods for destruction of stego image formed by adaptive embedding methods, namely considerable changes of image's statistical parameters. The proposed method allows preserving both minimal changes of a Cover Image (CI) parameters, and ratio of survived bits of embedded message (less than 7%). This makes proposed solution an attractive candidate for reliable destruction of stego images formed by novel embedding methods. However, practical usage of proposed solution requires further improvement of dictionary learning methods, namely decreasing of computation complexity of dictionary forming procedure.

Keywords: digital image steganalysis, adaptive embedding methods, message re-embedding

1. Acronyms

AEM Adaptive Embedding Methods
AEN Autoencoder Network
BF Bilateral Filtering
CI Cover Image
DAE Denoising Autoencoder
DI Digital Image
MVG Multivariate Gaussian model
TVM Total Variation Minimization

2. Introduction

Reliable protection of critical information infrastructure that belongs to state or private organizations is topical task today. Of special interest are methods for early detection and counteraction to sensitive information leakage caused by steganographic data transmission from local to global networks by attackers [1]. Feature of such attack is data embedding into a innocuous files, like Digital Image (DI), that are processed and transmitted in communication systems.

The majority of research in the domain of digital images steganalysis is aimed at development of stegdetectors with extra low error rate [2]. These detectors allows reliably detecting wide range of known embedding methods even for the most difficult cases of low CI payload (less than 10%) [3]. However, performance of modern stegdetectors considerably depends on prior information about used embedding methods. Therefore, effectiveness of stegdetectors may drastically reduce inc case of processing of stego images formed by unknown embedding method (zero-day problem). One of solutions to overcome this limitation of modern stegdetectors is applying of stego image destruction methods as a preventive action.

Modern methods of stego image destruction are aimed at suppression of noise components that are widely used for message embedding. Despite removing of huge part of embedded message, statistical and spectral features of processed images can significantly differs from initial stego images. This may disclosure the steganalysis process to an attackers, which may select another digital media as cover files. Therefore, development of method for reliable stego image destruction while preserving low alteration of CI parameters is needed.

Solving of mentioned task is complicated by appearance of novel Adaptive Embedding Methods (AEM) that preserves minimal impact on CI statistical parameters. Therefore, destruction of embedded message requires rigorous analysis of local perturbances of pixels brightness for removing embedded bits. This makes ineffective widepsread image denoising methods that do not take into account local statistical parameters. Of

[a] progonov@gmail.com

special interest are methods of anisotropic filtering that can adjust parameters for each region of image. In spite of mentioned promising features of anisotropic filtering, the information about its performance for stego image destruction is limited in open literature. The paper is aimed at filling this gap and analyse performance of such methods, namely novel dictionary learning methods, for the case of destruction stego images formed by advanced AEM.

The rest of this paper is organized as follows. Notations are presented in section 3. The results of review of modern methods for stego image destruction are presented in section 4 that is concluded with purpose and tasks of the paper in section 5. Then, features of modern embedding methods are presented in section 6, while proposed method is described in section 7. Results of performance evaluation are presented in section 8. Section 9 summarizes the paper.

3. Preliminaries

High-dimensional arrays, matrices, and vectors will be typeset in boldface. Their individual elements will be represented with the corresponding lower-case letters in italic. For example, the identity matrix with size $L \times L$ elements will be denoted as $\mathbf{I}_L$.

The symbols $\mathbf{U} = (u_{ij}) \in \mathcal{I}^{N\times M}$, $\mathbf{X} = (x_{ij}) \in \mathcal{I}^{N\times M}$ and $\mathbf{Y} = (y_{ij}) \in \mathcal{I}^{N\times M}$, $\mathcal{I} = \{0, 1, \ldots, 255\}$, will always represent pixel values of 8-bit grayscale initial (non-processed), cover and stego images with size $N \times M$ pixels respectively. The embedded binary message will be represented as $\mathbf{M} \in \{0,1\}^{1\times|\mathbf{M}|}$, where $|\mathbf{M}|$ is message size in bits.

The notation $\|\cdot\|$ will correspond to either Euclidean norm for a scalar, or Frobenius norm for a matrix.

4. Related works

In most cases, message embedding into a CI by the novel AEM is performed with usage of noise-like areas, such as textures [4, 5]. Therefore, the widespread approach to stego image destruction is applying of image denoising methods, such as median and Wiener filters. However, non-local character of image processing with these methods may negatively impact on effectiveness of stego image destruction [6]. Therefore, of special interest are advanced methods for image denoising, such as anisotropic filtering [7]. The feature of these methods is adjusting of methods parameters for each area of image by taking into account of local statistics, for example the variance of pixels brightness values.

The example of modern methods for anisotropic filtering is Bilateral Filtering (BF) that is based on reducing the impact of additive interference while maintaining the contours of objects in the image [7]:

$$F_{BF}(\mathbf{U}_{x,y}) = \frac{1}{N_{BF}(i,j)} \times \\ \times \sum_{k=-(h_k-1)/2}^{(h_k-1)/2} \sum_{n=-(h_n-1)/2}^{(h_n-1)/2} \mathbf{U}_{x+k,y+n} \cdot \\ \cdot h(k,n) \cdot g(\mathbf{U}_{x+k,y+n} - \mathbf{U}_{x,y}), \quad (1)$$

$$N_{BF}(i,j) = \sum_{k=-(h_k-1)/2}^{(h_k-1)/2} \sum_{n=-(h_n-1)/2}^{(h_n-1)/2} h(k,n) \cdot \\ \cdot g(\mathbf{U}_{x+k,y+n} - \mathbf{U}_{x,y}),$$

where $h(k,n)$ is smoothing filter with size of $h_k \times h_k$ (pixels); $g(\cdot)$ is the function that reduces impact of smoothing filter near the contours of objects; $N_{BF}(i,j)$ is normalizing factor for the current position of the sliding window. The value of function $g(\cdot)$ is close to one for the areas with relatively low variation of pixels brightness that does not significantly affect the smoothing filter $h(k,n)$ in eq. 1. In other cases, value of function $g(\cdot)$ tends to zero near the contours that suppress influence of smoothing filter. The Gaussian smoothing is widely used as a function $h(k,n)$ for BF that allows to minimize influence of additive noise [7]. Still, BF tends to produce cartoon-like images by processing of high-textured areas, like grass, foliage, fur etc. This negatively impacts on stego image destruction, since it disclosure the fact of additional processing.

The alternative approach to anisotropic image denoising is based on usage of artificial neural networks, such as convolutional neural networks, recurrent-convolutional networks and Autoencoder Network (AEN). Of special interest is Denoising Autoencoder (DAE) due to ability of initial (pristine) image restoration from the noisy one [8]. The DAE belongs to the class of AEN that consists of encoder and decoder modules (Fig. 1).

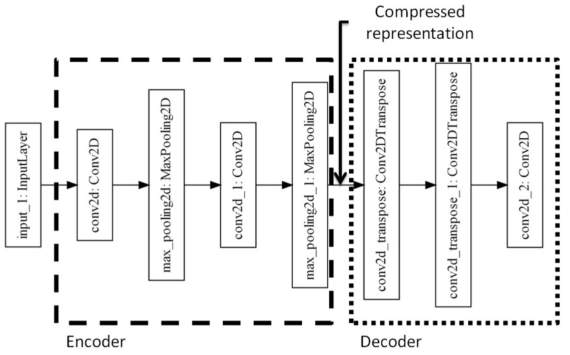
Compressed representation
input_1: InputLayer
conv2d: Conv2D
max_pooling2d: MaxPooling2D
conv2d_1: Conv2D
max_pooling2d_1: MaxPooling2D
conv2d_transpose: Conv2DTranspose
conv2d_transpose_1: Conv2DTranspose
conv2d_2: Conv2D
Encoder
Decoder

Fig. 1. The typical architecture of autoencoder neural network.

The first (encoder) part of AEN is aimed at projection of a given multidimensional signal, for example digital image, into a lower dimensional space, while maintaining its statistical features. The restoration of the pristine image is performed by the decoder network according to the obtained representation $\mathbf{h}$ (Fig. 1).

The set up of additional requirements for the parameters of encoder and decoder (Fig. 1) makes possible obtaining special features of AEN, for example denoising property [8]:

$$-\mathbb{E}_{\mathbf{U}\sim\hat{p}_{data}}\mathbb{E}_{\tilde{\mathbf{U}}\sim\mathtt{C}(\tilde{\mathbf{U}}|\mathbf{U})}\log(p_{decoder}(\mathbf{U}|\mathbf{h} = \\ = f(\tilde{\mathbf{U}}))) \to \min,$$

where $\mathbf{U}, \tilde{\mathbf{U}}$ are pristine and noisy images respectively; $\mathtt{C}(\tilde{\mathbf{U}}|\mathbf{U})$ is distortion introducing operator; $\hat{p}_{data}(\mathbf{U})$ is distribution of pristine images to be learnt; $p_{decoder}(\cdot)$ is distribution of output (processed) images after DAE.

Ability to learning of an appropriate transformation for restoring of pristine image from noisy ones makes DAE an attractive candidate for stego images destruction related task. However, practical usage of such network requires its time-consuming training with usage of examples of “expected” alterations. This may be impractical in cases when steganalytics havae limited ability to obtain cover and stego images forme by unknown embedding methods.

The idea of learning the appropriate transformation of noise image is closely related with Total Variation Minimization (TVM) techniques [9]. The TVM is aimed at decreasing the total variation $\sigma_{\mathbf{I}^2}$ of image $\mathbf{I}$ pixels brightness by preserving minimal impact on textures.

The value of $\sigma_{\mathbf{I}^2}$ for grayscale image $\mathbf{U}$ of size $N \times M$ (pixels) can be estimated as [9, 10]:

$$V(\mathbf{U}) = \sqrt{|\mathbf{U}_{x+1,y} - \mathbf{U}_{x,y}|^2 + |\mathbf{U}_{x,y} - \mathbf{U}_{x,y+1}|^2}. \quad (2)$$

Then, image denoising task can be presented as equivalent optimization problem of minimization the overall level of variation of image’s pixels brightness [9]:

$$\min_{\mathbf{U}}\left(\|\mathbf{U}\|_2^2 + \lambda\cdot V(\mathbf{U})\right), \quad (3)$$

where $\|\mathbf{U}\|_2^2$ is estimation of image’s energy; $\lambda > 0$ is regularization weight. Note that estimation $V(\mathbf{U})$ in eq. (2) is non-differentiable function that makes impossible applying of widespread optimization methods for solving of eq. (3). Therefore, the following approximation of value $\sigma_{\mathbf{I}^2}$ is used in most cases [11]:

$$V_a(\mathbf{U}) = \sum_{x,y}|\mathbf{U}_{x+1,y} - \mathbf{U}_{x,y}| + |\mathbf{U}_{x,y+1} - \mathbf{U}_{x,y}|. \quad (4)$$

Plug-in of approximation $V_a(\mathbf{U})$ in eq. (3) makes possible usage of widespread optimization methods for solving image denoising task. This can be represented as following optimization task [11]:

$$\min_{\mathbf{U}\in\mathtt{BV}(\Omega)}\|\mathbf{U}\|_{\mathtt{TV}(\Omega)} + \frac{\lambda}{2}\iint_{x,y\in\Omega}(\hat{\mathbf{U}} - \mathbf{U})^2dxdy, \quad (5)$$

where $\mathtt{BV}(\Omega)$ is a set of functions with limited variation of values over the domain Ω; $\mathtt{TV}(\Omega)$ is the operator for estimation the total variation of signals values in domain Ω; $\lambda > 0$ is regularization weight; $\hat{\mathbf{U}}$ is the estimated pristine image after applying of TVM-method. Note that operator $\mathtt{TV}(\Omega)$ is equal to gradient of a signal in case of processing signals with high degree of smoothness, like images:

$$\|\mathbf{U}\|_{\mathtt{TV}(\Omega)} = \int_{x,y\in\Omega}\|\nabla\mathbf{U}\|_2dxdy.$$

Then, the optimization problem in eq. (5) can be solved using numerical methods, such as the Euler-Lagrange method [12]:

$$\begin{cases}\nabla\left(\frac{\nabla_{\mathbf{U}}}{\|\mathbf{U}\|_2}\right) + \lambda(\hat{\mathbf{U}} - \mathbf{U}) = 0, & \mathbf{U}\in\Omega,\\ \frac{\partial\mathbf{U}}{\partial x\partial y} = 0, & \mathbf{U}\in\partial\Omega.\end{cases}$$

The TVM-based image denoising methods provide an effective way to image denoising by preserving high perceptual quality [9]. However, minimization of total variation of pixels brightness is aimed at suppression of noises with high amplitude. Therefore, TVM-method may be ineffective in case of stego image destruction due to negligible changes of pixels brightness during message hiding. Therefore, of special interest are methods for noise suppression regardless of noises magnitude and limited prior information about noise statistical features. The novel and promising approaches for solving this task is sparse representation of signals [13]. The idea of this approach is create a redundant set of basis functions (dictionary) that provide sparse representation of pristine image, while preserving minimal changes of decomposition coefficients for noisy ones. Also, formation of a dictionary can be performed on predefined examples of pristine images only [13]. This makes this approach a promising candidate for stego image destruction task. However, there are no information about performance of dictionary learning methods for steganalysis related task in open sources. Therefore, the paper is aimed at filling this gap and analyse performance of such methods for destruction of stego images formed by novel embedding methods.

5. The scope of research

The paper is aimed at performance analysis of destruction of stego images formed by AEM with usage of dictionary learning methods. To achieve this aim it is proposed to solve the following tasks:

1) to review features of novel adaptive embedding methods for digital images;
2) to review modern image denoising methods based on spectral analysis, namely sparse and redundant representation of signals;
3) to compare performance of state-of-the-art and proposed methods for stego images destruction.

The object of study is methods for steganalysis of stego images formed according to AEM. The subject of study is methods for destruction of embedded messages by preserving low distortions of a cover image statistical, spectral and structural parameters.

6. Adaptive embedding methods for digital images

Today, the methods for message hiding into cover (spatial) domain of DI are of special interest [2]. This is caused by well-founded theoretical background to achieve near-the-optimal empirical security of formed stego images [4, 5]. The proposed methods for message embedding into the spatial domain of CI can be divided into the following groups [14]:

1) **Distortion-minimizing methods** — are aimed at minimization of empirical function for estimation CI distortion. This is achieved by thorough preselection of pixels whose changes have as minimal as possible impact on CI statistical parameters.
2) **Side-informed methods** — are based on usage of additional information about pre-cover during message embedding. Generally, the pre-cover is subjected to some sort of CI processing or format conversion before message embedding. Still, precovers are rarely available in real cases that makes impractical wide usage of such methods.
3) **Methods with synchronized embedding changes** — take asymmetric embedding probabilities for each stego bits. It encourages synchronization (clustering) of polarities of neighboring modification that effectively counteracts to state-of-the-art stegdetectors.

Among considered groups, of special interest are methods for distortion minimization. This is caused by wide range of proposed functions $D(\cdot,\cdot)$ [15] that provide accurate estimations of CI distortions by preserving of low complexity of embedding procedure. Methods with synchronized embedding changes relate to class of clustering Modification Direction steganography [16]. Feature of such methods is providing similar (synchronized) changes of neighboring pixels while single bit is embedded. This decreases alterations of CI statistical parameters caused by stego image formation that that negatively impact on stegdetector performance.

The paper is focused on the case of destruction of stego images formed by mentioned state-of-the-art steganographic approaches. We considered the case of usage the novel MiPOD method [17] based on distortion-minimization technique as well as advanced Synch methods [18] with synchronized embedding changes. Let us consider these methods in details.

The MiPOD embedding method is aimed at minimization both CI distortion, and statistical detectability of formed stego image [17]. This is achieved by applying of locally-estimated Multivariate Gaussian model (MVG) of cover image. The model allows deriving a closed-form expression for a stegdetector performance as well as modeling the non-stationary character of natural images [17].

The pipeline of message $\mathbf{M}$ hiding into a cover image $\mathbf{X}$ by MiPOD method can be divided into several steps [17]. Firstly, the CI context is suppressed using denoising filter F:

$$\mathbf{r} = \mathbf{X} - F\left(\mathbf{X}\right).$$

At the second stage, the variance σ_l^2 of obtained residuals $\mathbf{r}$ is measured with the following linear model:

$$\mathbf{r}_l = \mathbf{G}\mathbf{a}_l + \xi, l \in \{1, \ldots, M \cdot N\}, \tag{6}$$

where $\mathbf{r}_l$ is residuals inside block of size $p \times p$ (pixels) surrounding the l^{th} cover image pixel; $\mathbf{G}_{p^2 \times p}$ is the mixing matrix for model's parametrs; $\mathbf{a}_{p\times 1}$ is the vector of model parameters; $\xi_{p^2\times 1}$ is the signal whose variance is need to be estimated. For practical cases, Maximum Likelihood Estimation can be used for evaluation model parameters in eq. (6) [17]:

$$\sigma_l^2 = \frac{\|\mathbf{P}_{\mathbf{G}}^{\perp}\mathbf{r}_l\|^2}{p^2 - q}, \tag{7}$$

where $\mathbf{P}_{\mathbf{G}}^{\perp} = \mathbf{I}_l - \mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T$ is the orthogonal projection of residual $\mathbf{r}_l$ onto $(p^2 - q)$, $q \in \mathbb{N}$, dimensional sub-space spanned by the left eigenvectors of matrix $\mathbf{G}$.

On the next stage, the embedding change β_l, $l \in [1; M \cdot N]$, that minimizes differences between cover and stego images distributions is estimated:

$$\zeta^2\left(\beta_l\right) = 2\sum_{l=1}^{M\cdot N} \beta_l^2 \sigma_l^{-4} \xrightarrow[\sum_{l=1}^{M\cdot N} H(\beta_l)=const]{} \min, \tag{8}$$

where ζ^2 is deflection coefficient used for estimation differences between cover and stego images distributions; $H_4(z) = -2z\log(z) - (1-2z)\log(1-2z)$ is ternary entropy function. Solving of eq. (8) can be done with usage of Lagrange multipliers method [17]. Therefore, the change rate β_l and Lagrange multiplier λ can be determined by numerical solving of following equations:

$$\beta_l\sigma_l^{-4} = \frac{1}{2\lambda}\ln\left(\frac{1-2\beta_l}{\beta_l}\right), l \in \{1, \ldots, M \cdot N\}.$$

Then, estimated change rate β_l is converted to the corresponding cost ρ_l of stego bit hiding in l^{th} pixel of a cover image:

$$\rho_l = -\ln(\beta_l - 2). \tag{9}$$

Finally, the pixels set with minimal value of total cost after estimation the cost ρ_l for each pixel. The set is used for embedding of message $\mathbf{M}$ that is pre-processed with usage of syndrome-trellis codes with pixels costs determined according to eq. (9). The results of performance evaluation for MiPOD method [17] proved effectiveness of usage of MVG to achieve state-of-the-art empirical security of stego images without usage of compute-intensive statistical models.

The alternative approach for message embedding into a CI is based on synchronization of changes of pixels brightness during stego image formation. However, practical usage of such approach needs time-consuming selection of an appropriate pixels set [16]. Therefore, the empirical methods for selection these set with usage of adjacent pixels were proposed [18]. These methods use similar processing pipeline that consists of the following steps. At first, the groups of adjacent pixels are split into sets [14]:

$$\begin{aligned}\mathcal{L}_1 &= \{(i,j)|\ \bmod (i,2)=1 \wedge \ \bmod (j,2)=1\},\\ \mathcal{L}_2 &= \{(i,j)|\ \bmod (i,2)=1 \wedge \ \bmod (j,2)=0\},\\ \mathcal{L}_3 &= \{(i,j)|\ \bmod (i,2)=0 \wedge \ \bmod (j,2)=1\},\\ \mathcal{L}_4 &= \{(i,j)|\ \bmod (i,2)=0 \wedge \ \bmod (j,2)=0\},\end{aligned}$$

where (i,j) is coordinate of current pixel; $\bmod (\cdot,\cdot)$ is modulo operation. Then, a message $\mathbf{M}$ is split into four parts $\mathbf{M}=\mathbf{M}_m, m\in[1;4]$. Each part is embedded with usage of appropriate set $\mathcal{L}_m$ and pre-selected AEM, while magnitude of pixels brightness change is equal to $v>0$ [14]. This leads to the effect of "alteration synchronization" — additional reduction of CI alterations $\rho_{i,j}$ in $q>0$ times. The value of the parameter q is determined empirically by comparison values of empirical function for estimation CI distortions $D(\mathbf{X},\mathbf{Y})$ by using either of "synchronized" changes, or applying of only AEM. According to evaluation results [16], the value of parameter q equals to nine for widespread adaptive embedding methods.

7. Advanced methods for stego images destruction

A message embedding into a cover image by AEM is performed with usage of noise components of a cover [4, 5]. Therefore, widepsread approach to destruct of message is based on applying of image denoising techniques, like median filter and lossy compression. Despite low ratio of survived bits after these transformations, this leads to considerable changes of statistical and spectral parameters of DI. Therefore, a sender and recipient of stego images may easily detect such intrusion into a steganographic channel and use another type of cover files instead [5]. We proposed to apply novel methods of spectral analysis, namely dictionary learning, for mitigation with mentioned limitation of state-of-the-art solutions.

The dictionary learning methods are aimed at performing sparse and redundant representation of analyzed signal in term of its approximation using the biggest M-elements [13]:

$$\min_{\mathbf{A},\{\mathbf{x}_i\}_{i=1}^{M}} \sum_{i=1}^{M} \|\mathbf{x}_i\|_0, \|\mathbf{y}_i-\mathbf{A}\mathbf{x}_i\|_2 \le \varepsilon, \varepsilon \ge 0, \qquad (10)$$

where $\mathbf{y}_i$ is current signal; $\mathbf{x}_i$ is i^{th} vector of decomposition coefficients for signal $\mathbf{y}_i$; $\mathbf{A}$ is matrix (dictionary) formed by concatenation elements of the basis functions.

Image denoising with usage of dictionary learning can be performed by DI restoration by solving of optimization task eq. (10) [13]. In most cases, the task is solved with usage of block-coordinate relaxation methods, like method of optimal directions [13]. These methods are aimed at solving the initial problem eq. (10) in an iterative way, namely to use dictionary $\mathbf{A}_{(k-1)}$, obtained at previous step, for solving to minimize the reconstruction error of a sample $\mathbf{y}_i$ at the k^{th} step. Then, the obtained decomposition matrix $\mathbf{X}_{(k)}$ is used to adjust the elements of the dictionary $\mathbf{A}_{(k)}$ using the least squares [13]:

$$\begin{aligned}\mathbf{A}_{(k)} = \arg\min_{\mathbf{A}} \left\|\mathbf{Y}-\mathbf{A}\mathbf{X}_{(x)}\right\|_F^2 =\\ = \mathbf{Y}\mathbf{X}_{(k)}^T \left(\mathbf{X}_{(k)}\mathbf{X}_{(k)}^T\right)^{-1} = \mathbf{Y}\mathbf{X}_{(k)}^{+}, \qquad (11)\end{aligned}$$

where $\|\cdot\|_F$ is Frobenius norm. These steps are repeated until providing M-elemental approximation of analyzed signals. However, practical usage of such procedure may face with low convergence of the optimization problem in eq. (10) due to necessity of optimization the whole dictionary $\mathbf{A}$ at each step of the algorithm [13]. For overcoming the limitation, the K-SVD method for sequential estimate each element (atom) from a dictionary $\mathbf{A}$ was proposed [19].

The estimation of j_0-th atom by K-SVD method is done by using only $\mathbf{a}_{j_0}$ column of matrix $\mathbf{A}$, while other atoms are fixed. This allows rewriting of eq. (11) as follow [19]:

$$\begin{aligned}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_F^2 = \left\|\mathbf{Y}-\sum_{j=1}^{M}\mathbf{a}_j\mathbf{x}_j^T\right\|_F^2 =\\ = \left\|\left(\mathbf{Y}-\sum_{j\neq j_0}\mathbf{a}_j\mathbf{x}_j^T\right)-\sum_{j=1}^{M}\mathbf{a}_{j_0}\mathbf{x}_{j_0}^T\right\|_F^2,\end{aligned}$$

where $\mathbf{x}^T$ corresponds to j^{th} row of matrix X. The solution of this optimization task can be reformulated as minimizing the error matrix $\mathbf{E}_{j_0}$ by updating the values $\mathbf{a}_j$ and $\mathbf{x}_j^T$ only [19]:

$$\mathbf{E}_{j_0} = \left(\mathbf{Y}-\sum_{j\neq j_0}\mathbf{a}_j\mathbf{x}_j^T\right) \to \min. \qquad (12)$$

The projection operator in the subspace $\mathbf{P}_{j_0}$ is proposed to obtain j_0-th column of matrix $\mathbf{E}_{j_0}$. The operator can be represented as a matrix multiplied on the right by the matrix $\mathbf{E}_{j_0}$ to zeroing all other columns. Therefore, the matrix $\mathbf{P}_{j_0}$ has M rows and M_{j_0} columns (the number of elements of the sample that requires the use of j_0-th atom of dictionary $\mathbf{A}$).

Let us denote $\left(\mathbf{x}_{j_0}^R\right)^T = \mathbf{x}_{j_0}^T\mathbf{P}_{j_0}$ as result of applying the operator $\mathbf{P}_{j_0}$ to the eq. (12). Then, the approximation of product $\mathbf{E}_{j_0}\mathbf{P}_{j_0}$ can be obtained by applying a singular value decomposition. The estimated approximation is used to update of both atom $\mathbf{a}_{j_0}$, and decomposition coefficients of the current vector $\mathbf{x}_{j_0}^T$.

To speed up the computation, the block-coordinate optimization method can be applied to solving of eq. (12) [19]. Then, $\mathbf{x}_{j_0}^T$ and $\mathbf{a}_{j_0}$ can be updated as follow [19]:

$$\min_{\mathbf{x}_{j_0}^R} \left\|\mathbf{E}_{j_0}\mathbf{P}_{j_0}-\mathbf{a}_{j_0}\left(\mathbf{x}_{j_0}^R\right)^T\right\|_F^2 \Rightarrow \mathbf{x}_{j_0}^R = \frac{\mathbf{P}_{j_0}^T\mathbf{E}_{j_0}^T\mathbf{a}_{j_0}}{\|\mathbf{a}_{j_0}\|_2^2},$$

$$\min_{\mathbf{a}_{j_0}} \left\| \mathbf{E}_{j_0} \mathbf{P}_{j_0} - \mathbf{a}_{j_0} \left(\mathbf{x}_{j_0}^R \right)^T \right\|_F^2 \Rightarrow \mathbf{a}_{j_0} = \frac{\mathbf{P}_{j_0}^T \mathbf{E}_{j_0}^T \mathbf{x}_{j_0}^R}{\|\mathbf{x}_{j_0}^R\|_2^2}.$$

The K-SVD method is characterized by high accuracy of dictionary $\mathbf{A}$ estimation while providing a given degree of sparseness of decomposition vectors $\mathbf{X}$ [13, 19]. This makes this method an attractive candidate for stego image denoising by preservation of low alteration of statistical and spectral features for a CI.

8. Experiments

Performance evaluation of state-of-the-art and proposed methods for stego images destruction was performed with usage of standard ALASKA dataset [20]. The subset of 250 images was sampled from the dataset, then images were converted to grayscale color mode. Finally, images were resized to fixed size of 512 × 512 (pixels) for preserving tractable computation complexity of dictionary learning with considered methods.

The stego images were formed according to considered MiPOD [17] and Synch [18] embedding methods. The case of low ($\Delta_\alpha^S = 3\%$) and middle ($\Delta_\alpha^S = 10\%$) cover image payloads was considered.

The effectiveness of stego image destruction was estimated by ratio Δ_p of pixels used for stego bit hiding whose brightness are remained after destruction. The distortions of a CI cause by applying of stego images destruction methods were estimated with using of following parameters:

- Statistical parameters — the standard statistical SPAM model [21] was used for estimation alteration of correlation of adjacent pixels brightness;
- Spectral parameters — were estimated with usage of decomposition coefficients for two-dimensional discrete wavelet transform. The Haar wavelet and corresponding scaling function were used as basis for transformation.
- Structural parameters — were used to evaluate changes of statistical parameters for analyzed image components. The Renyi (D_R) and multifractal ($f(\alpha)$) spectra were used as standard structural features of an image [22].

The proposed method for stego image destruction was compared with common (median filter and lossy JPEG compression) as well as state-of-the-art TVM methods. The estimated values of ratio Δ_p, statistical (Δ_F^{SPAM}), spectral (Δ_F^{DWT}) and structural (D_R and $f(\alpha)$) parameters of processed stego images after applying of state-of-the-art and proposed methods are presented at table 1.

Usage of TVM-method allows considerably (up to four times) reducing alteration of statistical and spectral parameters of analyzed images in comparison with median filtering and lossy JPEG compression (table 1) for low cover image payload ($\Delta_\alpha^S = 3\%$). On the other hand, efficiency of TVM-methods decreases for middle cover image payload ($\Delta_\alpha^S = 10\%$), when changes of mentioned parameters is similar to applying of widepsread destruction methods. This can be explained by spreading of altered image over the whole CI instead of localization into individual regions that complicates detection of local perturbance of pixels brightness.

Processing of stego images formed by Synch embedding method leads to negligible decreasing of statistical and spectral parameters changes in comparison with MiPOD method (table 1). Despite synchronization of embedded changes, effectiveness of state-of-the-art methods for stego image destruction remains high. Obtained effect can be explained by introducing of similar patterns of pixels brightness changes into a CI by Synch method that can be effectively suppressed with widespread denoising methods.

Let us note that applying both widespread (median filtering and lossy compression), and state-of-the-art TVM-method leads to similar changes of Renyi D_R and multifractal $f(\alpha)$ spectra for MiPOD and Synch embedding methods (table 1). This can be explained by non-local influences of these methods that leads to changes of the majority of image's components. Also, this effect leads to high ratio Δ_p (more than 15%) for MiPOD embedding methods (table 1).

The important result is relatively high ratio (more than 15%) of survived bits after applying of state-of-the-art destruction methods. Thus, even applying of "aggressive" image denoising does not ensure reliable destruction of the whole embedded message. This disclosures limitations of used destruction methods in modern intrusion prevention systems.

Applying of proposed method leads to drastically decreasing of CI related parameters in comparison with considered methods (table 1). The decreasing of parameters changes achieves up to seven times for both cases of low and middle cover image payloads. Therefore, proposed dictionary learning method allows effectively suppressing only noise components of DI by preserving low changes other components. Also, the ratio Δ_p for proposed method remains low (close to zero) that counteract restoration of embedded message at the receiver's side. This can be explained by high "selectivity" of proposed methods to change of pixels brightness.

9. Conclusion

The paper is aimed at performance analysis of dictionary learning methods for stego image destruction task. The case of stego image formation according to novel MiPOD and Synch embedding method was considered. Based on results of evaluation, we may conclude that:

1) Stego images destruction with using of widespread image denoising methods, like median filtering and lossy JPEG compression, does not provide reliable destruction of embedded message (ratio of survived bits is about 20%). Also, these methods considerably change statistical, spectral and structural parameters of DI that disclosures the steganalysis process. This limits practical usage of such image denoising methods for reliable destruction of stego images in modern intrusion prevention systems.
2) Applying of state-of-the-art TVM-method for stego image destruction allows up to four times decreasing changes of CI parameters, while preserving mid-

Table 1. The estimated values of ratio Δ_p, statistical (Δ_F^{SPAM}), spectral (Δ_F^{DWT}) and structural (D_R and $f(\alpha)$) parameters of processed stego images after applying of state-of-the-art and proposed methods for MiPOD embedding method.

	Ideal case	Stego image destruction methods			
		Median filter (5×5 pixels)	Lossy JPEG compression (quality index = 75%)	TVM method	Proposed method
MiPOD embedding method ($\Delta_\alpha^S = 3\%$)					
Δ_F^{SPAM}	0.00	96.81	87.47	72.90	12.58
Δ_F^{DWT}	0.00	78.12	48.70	21.91	7.34
D_R	0.00	6.63	9.24	2.09	0.75
$f(\alpha)$	0.00	3.27	7.85	6.54	2.69
Δ_p	0.00	17.95	42.75	28.96	3.49
MiPOD embedding method ($\Delta_\alpha^S = 10\%$)					
Δ_F^{SPAM}	0.00	92.15	80.76	66.97	8.06
Δ_F^{DWT}	0.00	82.95	53.72	22.48	4.54
D_R	0.00	10.80	13.03	5.14	1.19
$f(\alpha)$	0.00	5.42	8.18	8.91	4.57
Δ_p	0.00	23.79	18.22	14.95	1.95
Synch embedding method ($\Delta_\alpha^S = 3\%$)					
Δ_F^{SPAM}	0.00	98.65	89.57	82.03	15.57
Δ_F^{DWT}	0.00	81.84	55.91	17.93	13.40
D_R	0.00	7.67	8.87	3.75	1.77
$f(\alpha)$	0.00	2.74	5.31	5.39	1.22
Δ_p	0.00	89.65	23.55	12.17	7.12
Synch embedding method ($\Delta_\alpha^S = 10\%$)					
Δ_F^{SPAM}	0.00	90.70	81.60	76.03	11.18
Δ_F^{DWT}	0.00	85.27	59.69	22.04	10.62
D_R	0.00	9.09	12.71	6.82	2.35
$f(\alpha)$	0.00	4.69	5.48	5.31	3.71
Δ_p	0.00	56.95	18.02	10.03	4.44

dle (about 15%) ratio of survived bits. This makes this method an attractive candidate for reliable destruction of stego images, albeit minimization of CI parameters changes requires further improvements.

3) Dictionary learning methods allows preserving both minimal changes of a CI parameters, and ratio of survived bits of embedded message (less than 7%). This makes proposed solution an attractive candidate for reliable destruction of stego images in next-generation intrusion prevention systems. However, practical usage of proposed solution requires further improvement of dictionary learning methods, namely decreasing of computation complexity of dictionary forming procedure.

References

[1] D. Legezo, “MontysThree: Industrial espionage with steganography and a Russian accent on both sides.” https://securelist.com/montysthree-industrial-espionage/98972/. Accessed: 2022-Apr-12.

[2] M. Hassaballah, *Digital Media Steganography: Principles, Algorithms, and Advances.* Academic Press, 1 ed., 2020.

[3] D. Progonov and M. Yarysh, “Analyzing the accuracy of detecting steganograms formed by adaptive steganographic methods when using artificial neural networks,” *Eastern-European Journal of Enterprise Technologies*, vol. 1, pp. 45–55, 2022.

[4] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications.* Cambridge University Press, 1 ed., 2009.

[5] G. Konachovych, D. Progonov, and O. Puzyrenko, *Digital steganography processing and analysis of multimedia files.* Tsentr uchbovoi literatury, 2018. In Ukrainian.

[6] D. Progonov, “Analysis of changes the renyi divergence for pixel brightness distributions by stego images wiener filtering,” *Information Technologies and Knowledge*, vol. 12, pp. 3–25, 2018.

[7] R. C. Gonzalez and R. E. Woods, *Digital image processing.* Pearson, 4 ed., 2018.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* The MIT Press, 1 ed., 2016.

[9] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way.* Academic Press, 3 ed., 2008.

[10] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D*, vol. 60, pp. 259–268, 1992.

[11] A. Chambolle, “An algorithm for total variation minimization and applications,” *Journal of Mathematical Imaging and Vision*, vol. 20, pp. 89–97, 2004.

[12] P. Getreuer, "Rudin–osher–fatemi total variation denoising using split bregman," *Image Processing On Line*, vol. 2, pp. 74–95, 2012.

[13] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing.* Springer, 1 ed., 2010.

[14] M. Boroumand and J. Fridrich, "Synchronizing embedding changes in side-informed steganography," in *Electronic Imaging, Media Watermarking, Security, and Forensics Symposium*, Society for Imaging Science and Technology, 2020.

[15] T. Filler and J. Fridrich, "Design of adaptive steganographic schemes for digital images," in *Proceedings of SPIE – The International Society for Optical Engineering*, SPIE, 2 2011.

[16] B. Li, M. Wang, X. Li, S. Tan, and J. Huang, "A strategy of clustering modification directions in spatial image steganography," *IEEE Transactions on Information Forensics and Security*, vol. 10, pp. 1905–1917, 2015.

[17] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Trans. Inf. Forensics Security*, vol. 11, pp. 221–234, 2 2015.

[18] T. Denemark and J. Fridrich, "Improving steganographic security by synchronizing the selection channel," in *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security*, ACM, 2015.

[19] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcompletes dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311–4322, 2006.

[20] R. Cogranne, Q. Gilboulot, and P. Bas, "The alaska steganalysis challenge: A first step towards steganalysis," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, pp. 125–137, ACM, 2019.

[21] T. Pevny, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Trans. Inf. Forensics Security*, vol. 5, pp. 215–224, 6 2010.

[22] D. Progonov and S. Kushch, "Spectral analysis of steganograms," *Radio Electronics, Computer Science, Control*, vol. 2, pp. 71–80, 2015.