UDC 004.[056.54:932.2]

Statistical stegdetectors performance by message re-embedding

Dmytro Progonov[1, a]

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

Abstract

State-of-the-art stegdetectors for digital images are based on pre-processing (calibration) of analyzed image for increasing stego-to-cover ratio. In most cases, the calibration is realized by image processing with enormous set of high-pass filters to obtain good estimation of cover image from the stego one. Nevertheless, the efficiency of this approach significantly depends on careful selection of filters for reliably extraction of cover image alterations that are specific for each embedding method. The selection is non-trivial and laborious operation that is realized today by training of convolutional neural networks, such as Ye-Net, SR-Net to name but a few.

The paper is devoted to performance analysis of alternative approach to image calibration, namely message re-embedding into analyzed image. The considered method is aimed to increasing stego-to-cover ratio by amplification of cover image alterations caused by message hiding. The analysis was performed on ALASKA and VISION datasets by usage of stegdetector based on SPAM model of covers. Messages were re-embedded according to state-of-the-art adaptive methods HUGO, S-UNIWARD, MG and MiPOD. Proposed approach allows significantly (up to 20%) decreasing detection error even in case of low payload of cover image (less than 10%) where modern stegdetectors are ineffective.

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

1. Introduction

Securing sensitive data of public and private organizations is topical task today. Special attention is paid to the data leakage prevention during files transmission in communication systems. The data leakage can be performed by an attackers with hidden (steganographic) communication systems [1]. These systems are aimed at sensitive data embedding into innocuous digital media, for instance digital images, and further transmission of obtained stego images to a recipient.

The wide range of digital image steganalysis methods is proposed for stego images detection. These methods are based on studying the differences between current image and used model of cover [1, 2]. Achieving high detection accuracy (more than 95%) requires improvement of stego-to-cover ratio that can be realized by image pre-processing (calibration). The state-of-the-art methods of image calibration are aimed on obtaining good estimation of cover image (CI) from the stego one. It is achieved by image processing using various high-pass filters [1]. Nevertheless, such approach requires using a priori information about embedding method for selection appropriate filters that may be inappropriate in real cases.

For overcoming mentioned issue, we proposed to improving stego-to-cover ratio by amplification of CI distortions caused by message hiding. The amplification can be achieved by message repetitive embedding (re-embedding) into analyzed images. The effectiveness of proposed approach was shown in paper [3] for the case of message re-embedding with same method and similar payload of cover image. Since steganalytic may have limited access to embedding module in real scenarios, the case of message re-embedding according to known steganographic methods should be considered.

This work focuses on performance analysis of statistical stegdetector in case of message re-embedding with state-of-the-art adaptive embedding methods (AEM) HUGO, S-UNIWARD, MG and MiPOD. In the next section describe the common approaches to digital image calibration. The notations and used functions are defined in Section 3. Section 4 describes modern approaches for adaptive message hiding into digital images, while state-of-the-art methods for steganalysis of obtained stego images are shown in Section 5. In Sections 6 and 7, we provide the results of all experiments aimed at comparing the detection accuracy by message re-embedding with state-of-the-art AEM and discussion of obtained results. The paper is concluded in Section 8.

2. Related works

Today, the common paradigm in digital image steganalysis is based on features learning from cover and stego images noise components [1]. Estimation of these components is non-trivial task due to theirs low energy and masking into image context. Therefore, it is used image calibration methods for suppression of image context by preserving statistical features of noises. The careful selection of calibration methods allows considerably improving stego-to-cover ratio and reliably revealing negligible alterations of CI pixels brightness caused by message hiding[2].

The state-of-the-art image calibration methods are based on applying high-pass filters [1]. As an example,

[a] progonov@gmail.com

we may mention the SRM model [4] and SR-Net convolutional neural network [5] based stegdetectors. The former method takes enormous set of two-dimensional high-pass filters for effective suppression context in real images. The latter method allows considerably reducing the number of used filters (convolution kernels) by training a network on heterogeneous dataset of photographic images. Effectiveness of these approaches was shown for both state-of-the-art AEM and advanced side-informed embedding methods that utilize knowledge of pre-covers (raw images) [6]. This became possible due to either utilization of a priori information about used embedding method for selection appropriate filters, or learning versatile filters by thorough training of a network on several datasets.

Therefore, this approach can be not suitale for real cases when steganalytics have limited awareness of applied embedding methods [2]. Thus, it is topical task to develop the universal calibration function that preserves high detection accuracy even in case of limited a priori information about used steganographic algorithm.

The proposed methods for image calibration can be divided into next groups [7]:

1) **Parallel reference** — the calibration can be seen as a constant shift of feature space.
2) **Eraser** — it relates to a transformation that is robust to embedding changes. Then, image calibration leads to erasing of embedding stego bits.
3) **Cover image estimate** — it corresponds to the original idea of image calibration, namely to obtain good estimation of CI.
4) **Stego image estimate** — the calibration provides an accurate estimation of features for stego images. This case is complementary to the previous one.
5) **Divergent reference** — the calibration leads to a shift of features for both cover and stego images to different directions.

It should be noted, the parallel and divergent references lead to failure of steganalysis due to making indistinguishable cover and stego images classes [7].

Practical application of eraser scheme needs a priori information about specific distortions of covers that may be unavailable in real cases. On the other hand, the cover image estimation approach requires careful selection of cover model [1]. In spite of models diversity, they cope with specific distortions of covers, such as compression, denoising to name but a few. It makes these models inapropriate for real cases where steganalytics may face several types of image distortions at the same time.

In contrast to considered types of image calibration, the stego image estimation scheme relies on amplification of CI alterations caused by message hiding. The example of such calibration is message re-embedding that leads to significantly changing of CI features while preserving negligible alterations of stego image features. Thus, steganalytic may use known embedding methods for increasing stego-to-cover ratio even by processing of stego images formed by unknown steganographic algorithm. Despite simplicity of this appoach, it did not get enough attention today. This work focuses on filling in these gaps — performance analysis of statistical stegdetectors in case of image calibration by message re-embedding according to state-of-the-art AEM.

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-calibrated), cover and stego images with size $N \times M$ pixels respectively. The feature extraction operator $F_e(\cdot)$ will be used for extraction a feature vector $\mathbf{F}$ from an image.

The embedded binary message will be represented as $\mathbf{M} = (m_{ij}) \in \{0, 1\}^{1 \times |\mathbf{M}|}$, where $|\mathbf{M}|$ is message size in bits.

The probability of event A will be denoted as $Pr(A)$. The Iverson bracket $[a]_I$ equals to one of the boolean expression a is true, and zero otherwise. The notation $\|\cdot\|$ will correspond to either Euclidean norm for a scalar, or Frobenius norm for a matrice.

4. Adaptive embedding methods for digital images

Steganographic methods are aimed for message hiding into cover file, such as digital image, while preserving cover's perceptual quality and minimal changes of its statistical parameters. The known embedding methods can be divided into two groups depending on the way of stego bits hiding [1]:

- **Cover domain** — stego bits are embedded by alteration of cover elements values, e.g. brightness of cover image pixels;
- **Transformation domain** — a cover is pre-processed with some transformation that allows easy estimation cover's distortion. Then, message is embedded by manipulation of obtained coefficients.

The well-known example of transformation based embedding method is JPEG-steganography [1]. In this case, stego bits are hiding by alteration coefficients of Discrete Cosine Transform of cover image. Practical applications of such methods are limited due to introducing specific distortions into CI that can be easily detected by modern stegdetectors [2, 8].

Today, message hiding into cover (spatial) domain takes leading positions in digital image steganography. The methods for data embedding into spatial domain can be divided into next groups [9]:

1) **Distortion-minimizing methods** — are aimed on usage of empirical functions for estimation CI distortion. A message is embedded by selection CI pixels with minimal expected cost.
2) **Side-informed (SI) methods** — are based on usage of pre-cover during data hiding. Generally, the pre-cover is subjected to some sort of cover processing or format conversion before embedding the

secret message. Nevertheless, pre-covers are rarely available in real cases.

3) **Methods with synchronized embedding changes** — take asymmetric embedding probabilities for each stego bits. It encourages synchronization (clustering) of polarities of neighboring modification that decrease performance of state-of-the-art stegdetectors.

The first and second groups of methods are widely used today. They are based on minimization of total cost by message hiding into CI [10]:

$$D(\mathbf{X},\mathbf{Y}) = \sum_{i,j} \rho_{ij}(\mathbf{X},\mathbf{Y}) \xrightarrow[|\mathbf{M}|=const]{} \min, \quad (1)$$

where $D(\mathbf{X},\mathbf{Y})$ — the empirical distortion estimation function. The majority of proposed functions $D(\mathbf{X},\mathbf{Y})$ in eq. (1) is based on additivity assumption — the representation of CI distoriton as weighted sum of individual distortions caused by embedding of stego bits.

During message hiding into CI according to eq. (1), each pixel is assigned two costs $\rho_{ij}(1)$ and $\rho_{ij}(-1)$ that measures the impact on detectability when the $(i,j)^{\text{th}}$ element is modified by $(+1)$ or (-1) respectively. Selection of pixels for stego bit embedding usually is performed by some heuristic rules that assesse noise level in a local neighborhood of pixel (i,j). This allows achieving high empirical security while preserving computational effective optimization methods for cost estimation function in eq. (1).

The methods with synchronized embedding changes are promising and currently undeveloped domain of digital images steganography [9]. The effectiveness of these methods are based on curbing the range of local pixel differences or noise residuals, for instance counteracting sign-alternating kernels used in cover rich models [11]. Therefore, the paper is focused on advanced distortion-minimizing methods HUGO [12], S-UNIWARD [13], MG [14] and MiPOD [15]. Let us consider these methods in details.

The HUGO embedding method is based on minimization of CI overall distortion [12]:

$$\min_{\pi} \mathrm{E}_{\pi}[D] = \sum_{y\in\mathcal{Y}} \pi(y)\cdot D(y), \mathrm{H}(\pi) = |\mathbf{M}|, \quad (2)$$

where $y \in \mathcal{Y}$ — stego image y from set of all possible stego images $\mathcal{Y}$; π — probability distribution function of selection the certain y from $\mathcal{Y}$; $\mathrm{E}_{\pi}[D]$ — averaging operator for function $D(\mathbf{X},\mathbf{Y})$ over distribution π; $H(\pi) = -\sum_{y\in\mathcal{Y}} \pi(y)\cdot\log(\pi(y))$ — entropy function.

The optimization problem (2) can be solved by sampling stego images from Gibbs probability distribution [12]:

$$\pi_{\lambda_G}(y) = \frac{\exp(-\lambda_G D(y))}{\sum_{\tilde{y}\in\mathcal{Y}}\exp(-\lambda_G D(\tilde{y}))}. \quad (3)$$

The scalar $\lambda_G > 0$ is determined by solving of eq. (3). Filler *et al.* [12] suggested to use adjacency matrix $\mathbf{C}_{kl}(\mathbf{X})$ in eq. (1) for estimation distortion of CI during message hiding according to HUGO method:

$$D(\mathbf{Y}) = \sum_{c\in\mathcal{C}} \sum_{(k,l)\in\mathcal{I}} \omega_{k,l}\mathbf{H}^{c}_{(k,l)}(\mathbf{Y}),$$

where $\mathcal{C} = \mathcal{C}^{\rightarrow}\cup\mathcal{C}^{\leftarrow}\cup\mathcal{C}^{\uparrow}\cup\mathcal{C}^{\downarrow}$ — set of three-elements cliques for four-pixels adjacency directions; $\omega_{k,l} > 0$ — weighting factor; $\mathbf{H}$ — normalized adjacency matrix that is calculated for each type of cliques $\mathcal{C}$. For example, the matrix $\mathbf{H}$ can be calculated according to the next formulae in the case of CI processing in row-wise order:

$$\mathbf{H}^{\rightarrow}_{(k,l)}(\mathbf{Y}) = \frac{1}{N(M-2)}\cdot \cdot\sum_{i,j} \left|\left[\left(\mathbf{D}^{\rightarrow}_{i,j},\mathbf{D}^{\rightarrow}_{i,j+1}\right)(\mathbf{Y}) = (k,l)\right]_I - \left[\left(\mathbf{D}^{\rightarrow}_{i,j},\mathbf{D}^{\rightarrow}_{i,j+1}\right)(\mathbf{X}) = (k,l)\right]_I\right|,$$

$$\left(\mathbf{D}^{\rightarrow}_{i,j},\mathbf{D}^{\rightarrow}_{i,j+1}\right)(\mathbf{X}) = (k,l) \Leftrightarrow \Leftrightarrow \left(\mathbf{D}^{\rightarrow}_{i,j}(\mathbf{X}) = k\right)\wedge\left(\mathbf{D}^{\rightarrow}_{i,j+1}(\mathbf{X}) = l\right),$$

where $\mathbf{D}^{\rightarrow}_{i,j}(\mathbf{X}) = (\mathbf{X}_{i,j+1} - \mathbf{X}_{i,j})$ — adjacency matrix for the case of left-to-right pixels scanning. Normalized adjacency matrices $\mathbf{H}$ for other types of cliques can be calculated in the same way [12].

In contrast to HUGO method, the empirical distortion estimation funciton for S-UNIWARD embedding method is based on spectral transformation, namely two-dimensional discrete wavelet transform (2D-DWT) [13]:

$$D(\mathbf{X},\mathbf{Y}) = \sum_{k}\sum_{u,v} \frac{\left|\mathbf{W}^{k}_{uv}(\mathbf{X}) - \mathbf{W}^{k}_{uv}(\mathbf{Y})\right|}{\sigma + |\mathbf{W}^{k}_{uv}(\mathbf{X})|}, \quad (4)$$

where $\mathbf{W}^{k}_{uv}(\mathbf{X}), \mathbf{W}^{k}_{uv}(\mathbf{Y})$ — correspondingly, 2D-DWT coefficients of cover $\mathbf{X}$ and stego $\mathbf{Y}$ images with coordinates (u,v) in the k^{th} frequency subband; $\sigma > 0$ — stabilizing constant. Variation of basis functions for 2D-DWT in eq. (4) allows revealing specific distortions of CI caused by message hiding. The function $D(\mathbf{X},\mathbf{Y})$ in eq. (4) can be easily adapted for the cases of message hiding in spatial or transformation domains.

The MG and MiPOD embedding methods are aimed on minimization both CI distortion and statistical detectability of formed stego image [14, 15]. Feature of these methods is usage of locally-estimated multivariate Gaussian model of cover image. The model gives opportunity to derive a closed-form expression for a stegdetector performance and to capture the non-stationary character of natural images [15].

The pipeline of message $\mathbf{M}$ hiding into CI is similar for MG [14] as well as MiPOD [15] methods. Firstly, the CI context is suppressed using denoising filter F:

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

Secondly, variance σ_l^2 of obtained residuals is measured with linear model:

$$\mathbf{r}_l = \mathbf{G}\mathbf{a}_l + \xi, l \in \{1,\ldots,M\cdot N\}, \quad (5)$$

where $\mathbf{r}_l$ — the residuals $\mathbf{r}$ inside $p \times p$ block surrounding the l^{th} cover image pixel; $\mathbf{G}_{p^2 \times p}$ — the matrix that defines the parameteric model of remaining expectation; $\mathbf{a}_{p \times 1}$ — the vector of linear model parameters; $\xi_{p^2 \times 1}$ — the signal whose variance is need to be estimated. For practical cases, Maximum Likelihood Estimation can be used for calculation model parameters in eq. (5) [15]:

$$\sigma_l^2 = \frac{\|\mathbf{P}_{\mathbf{G}}^{\perp}\mathbf{r}_l\|^2}{p^2 - q}, \tag{6}$$

$$\mathbf{P}_{\mathbf{G}}^{\perp} = \mathbf{I}_l - \mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T,$$

where $\mathbf{P}_{\mathbf{G}}^{\perp}$ — the orthogonal projection of residual $\mathbf{r}_l$ in eq. (5) on $(p^2 - q)$, $q \in \mathbb{N}$, dimensional sub-space spanned by the left eigenvectors of $\mathbf{G}$. For the MG embedding method, the simplified estimation of variance in eq. (6) is used [14]:

$$\sigma_l^2 = \frac{\|\mathbf{r}_l - \tilde{\mathbf{r}}_l\|}{p^2 - q},$$

$$\tilde{\mathbf{r}}_l = \mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\mathbf{r}_l.$$

Thirdly, the embedding change β_l, $l \in \{1, \ldots, M \cdot N\}$ that minimizes deflection coefficient ζ^2 between cover and stego images distributions is determined:

$$\zeta^2(\beta_l) = 2\sum_{l=1}^{M \cdot N} \beta_l^2 \sigma_l^{-4} \xrightarrow[\sum_{l=1}^{M \cdot N} H(\beta_l) = const]{} \min, \tag{7}$$

where $H_4(z) = -2z\log(z) - (1-2z)\log(1-2z)$ — ternary entropy function. Solving of eq. (7) can be performed by applying the Lagrange multipliers method [15]. The change rate β_l and Lagrange multiplier λ can be determined by numerical solving of next 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 cover image:

$$\rho_l = \ln\left(\frac{1}{\beta_l - 2}\right). \tag{8}$$

Finally, the message $\mathbf{M}$ is embedded into CI using syndrome-trellis codes with pixels costs determined according to eq. (8).

The locally-estimated multivariate Gaussian model allows precisely measuring local distortions of CI caused message hiding [15]. It makes possible achieving state-of-the-art empirical security of stego images without usage of compute-intensive statistical models.

5. Statistical steganalysis of digital images

The state-of-the-art paradigm in digital image steganalysis is based on investigation of differences between image and its statistical model [1]. In most cases, these differences are informative enough to be used as features for classifiers, such as Support Vector Machines, Random Forest to name but a few. Nevertheless, development of «universal» model of CI that is suitable for both known and unknown embedding methods remains open problem in steganalysis [16].

The most of digital image statistical models is based on analysis the adjacency pixels brightness dependencies [4, 17]. Generally, this analysis is performed in several stages [4, 17]. Firstly, the image is pre-processed with high-pass filters for extraction noise components used for message hiding. Then, obtained residuals are truncated and quantized for limiting theirs dynamic range. Finally, prepared residuals are used for estimation the co-occurence matrices, which are applied as features for stegdetector training.

The well-known examples of statistical models that are based on mentioned approach are SRM [4] and SPAM [17] models. The difference between these models is image pre-processing stage. The SRM model is based on utilization huge range of high-pass filters for reliably detection of CI distortions caused by message hiding [4]. In contrast to SRM, the SPAM model takes an image without additional pre-processing [17].

As it was mentioned in Section 2, image high-pass filtering for SRM model is the particular case of cover image estimation approach. Effectiveness of this approach highly depends on usage the variety of filters for comprehensive analysis of image noise components [4]. On the other hand, usage of huge range of filters for SRM model leads to considerable increasing the requirements to used image dataset and duration of stegdetector training.

For overcoming mentioned drawbacks, we proposed to use stego image estimation based on message re-embedding. The effectiveness of this approach was shown in paper [3] for the case of message re-embedding with similar payload. In real cases, steganalytics may have limited opportunity to determine used embedding methods and estimate CI payload. Therefore, the paper is devoted to performance analysis of statistical stegdetector in case of image calibration by message re-embedding with known steganographic method. Therefore, we focuse on SPAM model that does not include any image pre-processing methods.

The calculation of SPAM-features starts by compution the difference array $\mathbf{D}$. For example, the array $\mathbf{D}$ for the case of row-wise image processing and left-to-right pixels scanning can be calculated as [17]:

$$\mathbf{D}_{i,j}^{\rightarrow} = \mathbf{U}_{i,j} - \mathbf{U}_{i,j+1}, i \in \{1, \ldots, M\}, j \in \{1, \ldots, N-1\}.$$

The first-order SPAM features $\mathbf{F}_1$ are used for modeling $\mathbf{D}$ by first-order Markov process [17]. For the horizontal direction, it leads to:

$$\mathbf{M}_{u,v}^{\rightarrow} = Pr(\mathbf{D}_{i,j+1}^{\rightarrow} = u | \mathbf{D}_{i,j}^{\rightarrow} = v), \quad u, v \in \{-T, \ldots, T\}, \tag{9}$$

where $T \in \mathbb{N}$ — threshold parameter. If probability $Pr(\mathbf{D}_{i,j}^{\rightarrow} = v)$ is equal to zero, then $\mathbf{M}_{u,v}^{\rightarrow} = 0$ as well.

The second-order SPAM features $\mathbf{F}_2$ are taken for modeling difference arrays $\mathbf{D}$ by second-order Markov process. Similarly to eq. (9), we obtain:

$$\mathbf{M}^{\rightarrow}_{u,v,w} = Pr(\mathbf{D}^{\rightarrow}_{i,j+2} = u|\mathbf{D}^{\rightarrow}_{i,j+1} = v, \mathbf{D}^{\rightarrow}_{i,j} = w), \quad u, v, w \in \{-T, \ldots, T\},$$

where $\mathbf{M}^{\rightarrow}_{u,v,w}$ is equal to zero if $Pr(\mathbf{D}^{\rightarrow}_{i,j+1} = v, \mathbf{D}^{\rightarrow}_{i,j} = w) = 0$. The features $\mathbf{F}_1$ and $\mathbf{F}_2$ for others scanning directions, denoted by a syperscript $c \in \{\leftarrow, \rightarrow, \uparrow, \downarrow, \nearrow, \swarrow, \searrow, \nwarrow\}$, can be estimated analogically.

For decreasing features dimensionality, the assumption that statistics in natural images are symmetric with respect to mirroring and flipping is used [17]. Thus, we can separately average matrices for horizontal, vertical an diagonal directions to form the final features sets:

$$\mathbf{F}^{\cdot}_{1,\ldots,k} = \frac{1}{4}\left[\mathbf{M}^{\rightarrow}_{\cdot} + \mathbf{M}^{\leftarrow}_{\cdot} + \mathbf{M}^{\uparrow}_{\cdot} + \mathbf{M}^{\downarrow}_{\cdot}\right],$$

$$\mathbf{F}^{\cdot}_{k+1,\ldots,2k} = \frac{1}{4}\left[\mathbf{M}^{\nearrow}_{\cdot} + \mathbf{M}^{\swarrow}_{\cdot} + \mathbf{M}^{\searrow}_{\cdot} + \mathbf{M}^{\nwarrow}_{\cdot}\right],$$

where $k = (2T + 1^2)$ for the first-order and $k = (2T + 1)^3$ for the second-order features.

6. Experiments

Performance analysis of statistical stegdetectors by message re-embedding was performed on ALASKA dataset [18]. The sub-set of $10,000$ grayscale images pseudo-randomly chosen from initial dataset was used. The case of stego image formation according to adaptive embedding methods HUGO, S-UNIWARD, MG and MiPOD was considered. The CI payload p_{init} was changed in range — 3%, 5%, 10%, 20%, 30%, 40%, 50%.

Since steganalytics do not known the CI payload in advance, the case of message hiding with random payload p_{re-emb} is considered. The payload p_{re-emb} is uniformly sampled from the range $p_{re-emb} \in [1; 50]$. Then, stegdetector can be tuned with next features:

1) **Non-calibrated features**— corresponds to the case of initial (non-calibrated) image $\mathbf{U}$ usage:

$$\mathbf{F}_{nc} = F_e(\mathbf{U}), \tag{10}$$

2) **Features of calibrated image** — corresponds to features obtained after message re-embedding into image $\mathbf{U}$:

$$\mathbf{F}_{re-emb} = F_e(C(\mathbf{U})), \tag{11}$$

3) **Linearly transformed features of calibrated image** — corresponds to the difference between features of calibrated and initial images:

$$\mathbf{F}_{DF} = \mathbf{F}_{re-emb} - \mathbf{F}_{nc}; \tag{12}$$

4) **Cartesian calibrated features** — correpsonds to the case of features merging for initial and calibrated images:

$$\mathbf{F}_{CC} = [\mathbf{F}_{nc}; \mathbf{F}_{re-emb}]. \tag{13}$$

The stegdetector includes ensemble classifier [19] trained with usage of SPAM model [17]. According to recommendation of Pevny *et al* [17], we used second-order SPAM features with $T = 3$, leading to 686 features. Stegdetector was tested according to cross-validation procedure with minimization of total error P_E [19]:

$$P_E = \min_{P_{FA}} \frac{1}{2}\left(P_{FA} + P_{MD}\left(P_{FA}\right)\right),$$

where P_{FA}, P_{MD} — are false alarm (misclassification cover images as stego ones) and missed detection (assignment of stego images as covers) probabilities respectively. The dataset was divided into training (90%) and testing (10%) sub-sets during cross-validation. The division was performed 10 times for estimation averaged values of total error P_E.

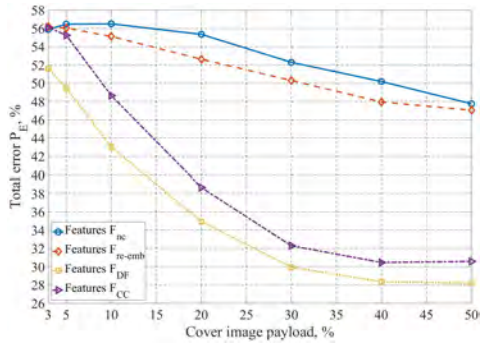
Stegdetector performance significantly depends on amount of cover-stego images pairs in training set [19, 20]. The majority of research in digital image steganalysis considers the case of forming training dataset only from cover-stego images pairs that allows

The majority of research in digital image steganalysis uses assumption that steganalycs have access to stego images generator for creation stego images from covers. This case may be unrealistic in some scenarios, such as revealing of stego images formed according to a priori unknown embedding method. In this situation steganalytics have access only to stego images, formed from inaccessible (unseen) covers. Therefore, we investigated this case during experiments for estimation of stedetectors performance in scenarios as close to reality as possible.

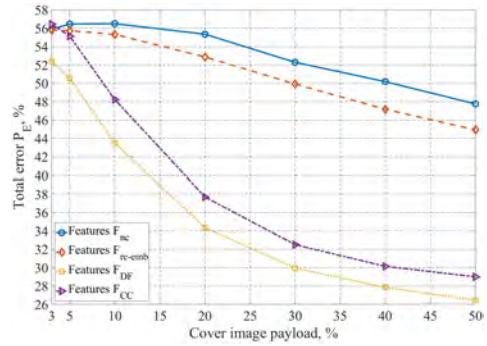
The dependencies of total error P_E on CI payload for stego images formed according to HUGO embedding method by usage of features (eqs. (10) to (13)) are represented at Fig. 1.

Message re-embedding by HUGO (Fig. 1a) and S-UNIWARD (Fig. 1b) methods allows considerably decreasing of total error P_E in comparison with MG (Fig. 1c) and MiPOD (Fig. 1d) methods. Images calibration with these methods does not considerably decrease total error P_E in comparison with usage of non-calibrated features $\mathbf{F}_{nc}$. Applying of $\mathbf{F}_{re-emb}$ and $\mathbf{F}_{CC}$ features allows negligibly decreasing P_E values (about $0.5\% - 0.75\%$) only for low payloads of CI (less than 10%). Similarly to HUGO (Fig. 1a) and S-UNIWARD (Fig. 1b) methods, usage of $\mathbf{F}_{DF}$ features leads to decreasing of P_E error up to 5% even for low payload of CI.

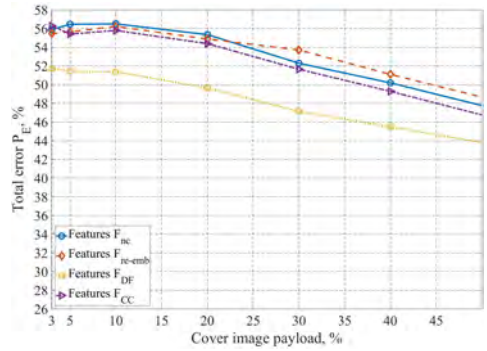
Usage of calibrated images features $\mathbf{F}_{re-emb}$ allows decreasing of error values only for medium and high payloads ($p_{re-emb} \geq 10$) — up to 4% for HUGO (Fig. 1a) and up to 5% for S-UNIWARD (Fig. 1b) methods. On the other hand, applying of cartesian calibrated $\mathbf{F}_{CC}$



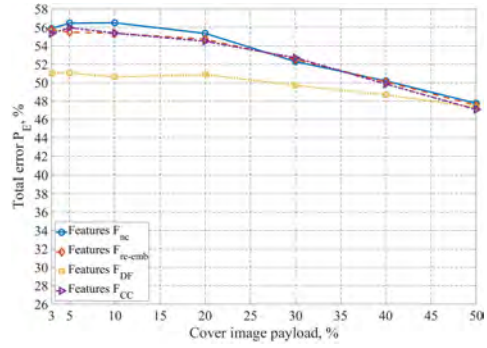
(a) Message re-embedding with HUGO method



(b) Message re-embedding with S-UNIWARD method



(c) Message re-embedding with MG method



(d) Message re-embedding with MiPOD method

Fig. 1. Dependencies of total error P_E on cover image payload for stego images formed according to HUGO embedding method. The stegdetector was tuned with usage of non-calibrated features $\mathbf{F}_{nc}$ (solid lines), features of calibrated image $\mathbf{F}_{re-emb}$ (dashed lines), linearly transformed features of calibrated image $\mathbf{F}_{DF}$ (dotted lines) and cartesian calibrated features $\mathbf{F}_{CC}$ (dash-dot lines).

and linearly transformed $\mathbf{F}_{DF}$ features leads to considerably decreasing of P_E error — up to 20% for high payload of cover image. The obtained results are non-trivial, since cartesian calibrated $\mathbf{F}_{CC}$ features have doubled dimensionality and contain information (features) for both initial and calibrated images. Contrariwise, linearly transformed $\mathbf{F}_{DF}$ features allows aditionally decreasing of detection error despite of operation over differences between SPAM-features for initial and calibrated images. The differences between values of P_E for $\mathbf{F}_{CC}$ and $\mathbf{F}_{DF}$ features (Fig. 1a-1b) are about 4% even for the case of low payloads of CI (less than 10%). It proved our assumption that usage of special type of image calibration and features post-processing allows considerably improving stegdetector performance.

The HUGO method is widely used as typical adaptive embedding methods today. Therefore, it is represented the interest to analyze the influence of message re-embedding into stego images formed according to modern S-UNIWARD method that preserves changes of cover's spectral features. The dependencies of total error P_E on cover image payload for stego images formed according to S-UNIWARD embedding method by usage of features (eqs. (10) to (13)) are represented at Fig. 2.

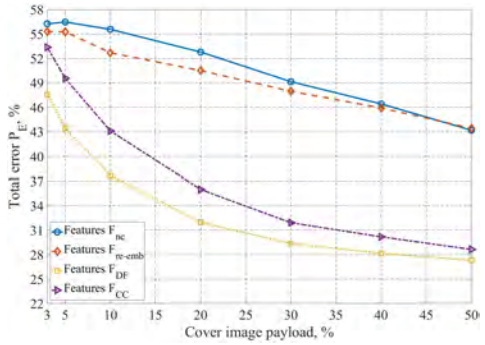
Similarly to HUGO methods (Fig. 1), message re-embedding by HUGO (Fig. 2a) and S-UNIWARD (Fig. 2b) methods allows considerably improving stegdetectors performance — the total error is decreasing about 20% for high payload and near 4% for low payload of cover image. Also, the biggest reducing of P_E is achieved for $\mathbf{F}_{CC}$ and $\mathbf{F}_{DF}$ features, whereas usage of $\mathbf{F}_{re-emb}$ leads to total error decreasing only up to 6%.

Also, it is revealed that the biggest improvement of stegdetector performance is achieved by message re-embedding according to same embedding method, which was used for stego image forming (Fig.1-2). It can be explained by amplification of method-specific distortions by message re-embedding. On the other hand, applying of MG (Fig. 2c) and MiPOD (Fig. 2d) methods allows incdeasing of detection error even for low payload of CI (about 5% for $\mathbf{F}_{DF}$ features).

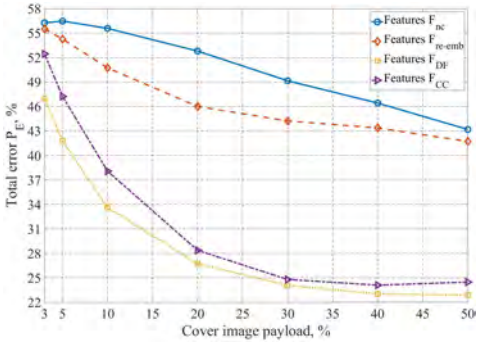
Obtained results for S-UNIWARD embedding methods proved our conclusions for HUGO method (Fig. 1). So, it is represent interest to further analysis of stegdetector performance by message re-embedding by advanced MG and MiPOD methods. The dependencies of total error P_E on cover image payload for stego images formed according to these methods by usage of features (eqs. (10) to (13)) are represented at Fig. 3 and Fig. 4 respectively.

Likewise HUGO (Fig. 1) and S-UNIWARD (Fig. 2) methods, message re-embedding according to steganographic algorithm that was used for stego image formation allows considerably decreasing of total error P_E in all range of CI payload (Fig. 3 – 4). It is achieved about 6% decreasing for MG (Fig. 3c) and 5% for MiPOD (Fig. 4d) methods even for low cover image payloads.

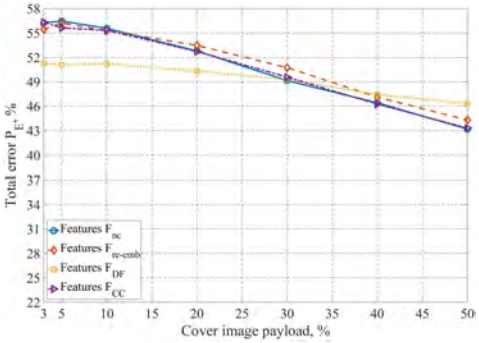
Also, it is revealed that applying of HUGO method for data re-embedding lead to decreasing of P_E values (about 3%) for both MG (Fig. 3a) and MiPOD (Fig. 4a)



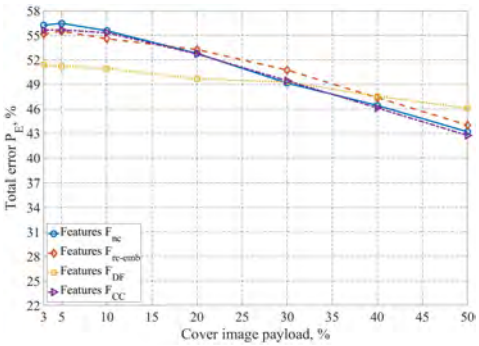
(a) Message re-embedding with HUGO method



(b) Message re-embedding with S-UNIWARD method

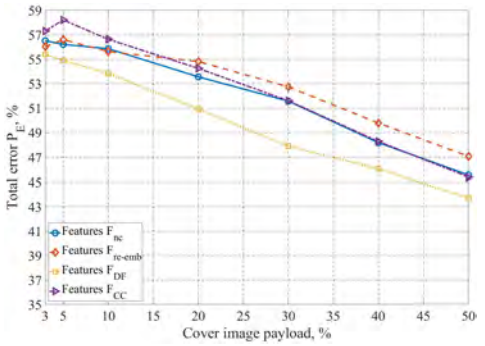


(c) Message re-embedding with MG method

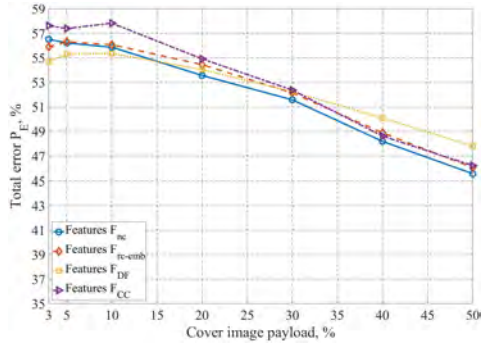


(d) Message re-embedding with MiPOD method

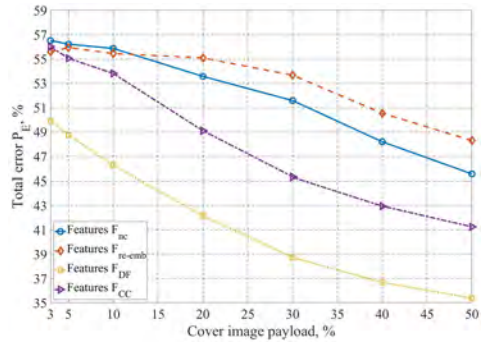
Fig. 2. Dependencies of total error P_E on cover image payload for stego images formed according to S-UNIWARD embedding method. The stegdetector was tuned with usage of non-calibrated features $\mathbf{F}_{nc}$ (solid lines), features of calibrated image $\mathbf{F}_{re-emb}$ (dashed lines), linearly transformed features of calibrated image $\mathbf{F}_{DF}$ (dotted lines) and cartesian calibrated features $\mathbf{F}_{CC}$ (dash-dot lines).



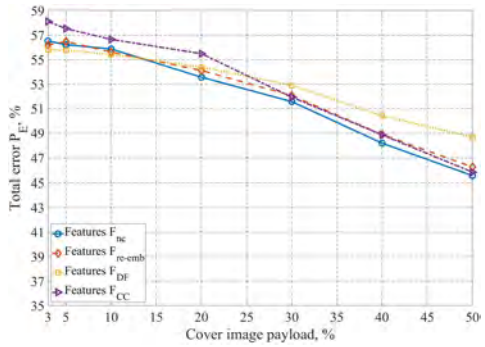
(a) Message re-embedding with HUGO method



(b) Message re-embedding with S-UNIWARD method

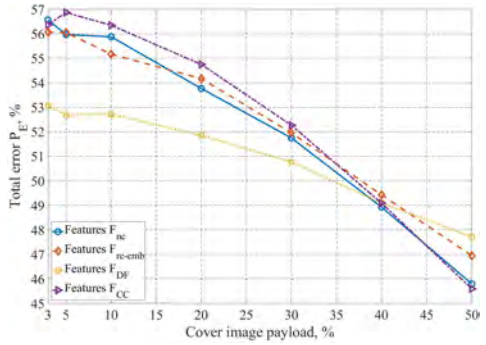


(c) Message re-embedding with MG method

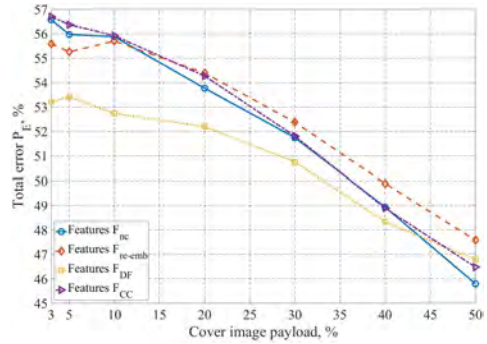


(d) Message re-embedding with MiPOD method

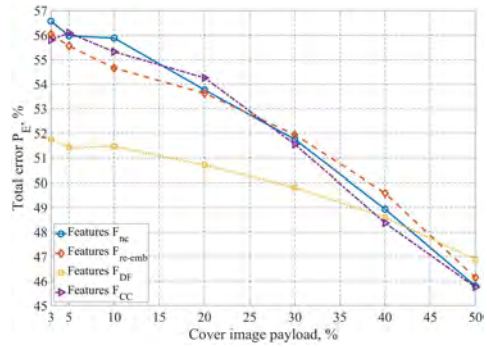
Fig. 3. Dependencies of total error P_E on cover image payload for stego images formed according to MG embedding method. The stegdetector was tuned with usage of non-calibrated features $\mathbf{F}_{nc}$ (solid lines), features of calibrated image $\mathbf{F}_{re-emb}$ (dashed lines), linearly transformed features of calibrated image $\mathbf{F}_{DF}$ (dotted lines) and cartesian calibrated features $\mathbf{F}_{CC}$ (dash-dot lines).



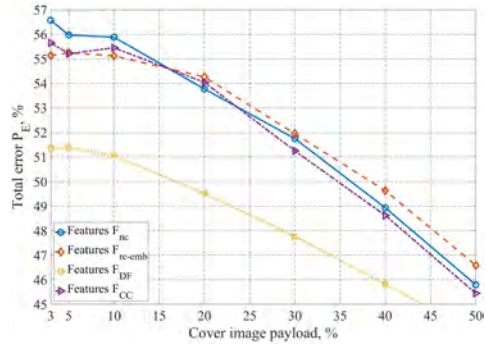
(a) Message re-embedding with HUGO method



(b) Message re-embedding with S-UNIWARD method



(c) Message re-embedding with MG method



(d) Message re-embedding with MiPOD method

Fig. 4. Dependencies of total error P_E on cover image payload for stego images formed according to MiPOD embedding method. The stegdetector was tuned with usage of non-calibrated features $\mathbf{F}_{nc}$ (solid lines), features of calibrated image $\mathbf{F}_{re-emb}$ (dashed lines), linearly transformed features of calibrated image $\mathbf{F}_{DF}$ (dotted lines) and cartesian calibrated features $\mathbf{F}_{CC}$ (dash-dot lines).

methods. On the other hand, usage of S-UNIWARD method gives opportunity to decrease detection error only for MiPOD method (Fig. 4b).

Decreasing of P_E was achieved by usage of $\mathbf{F}_{DF}$ features (Fig. 3 – 4). Applying of $\mathbf{F}_{CC}$ has low effect on detection accuracy in comparison with the case of usage the non-calibrated images. Taking $\mathbf{F}_{re-emb}$ features allows improving stegdetector preformance only by MG method re-embedding (Fig. 3c).

7. Discussion

The presented results for AEM (Fig. 1-4) proved our hypothesis that stego image estimation approach allows considerably improving stegdetector performance. This approach shown the biggest impact for HUGO and S-UNIWARD methods when simple message re-embeding with random payload can significantly (up to 20%) decrease detection error. For the advanced MG and MiPOD methods, proposed approach allows decreasing total error up to 7% even for low payload of CI (less than 10%) where known steganalysis methods are ineffective.

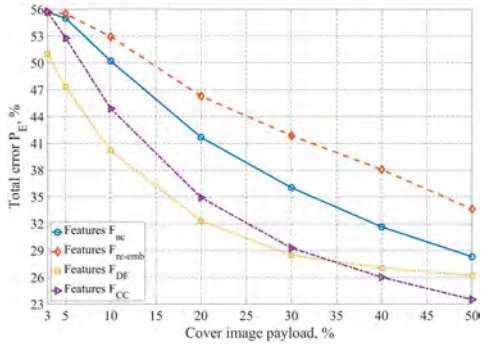
Obtained results may be explained by features of ALASKA dataset used in experiments, namely by low level of image noise. Therefore, we carried out additional verification on VISION dataset [21]. The dataset consists of images captured in-the-wild by 35 different portable devices of 11 major brands. The dependencies of total error P_E on cover image payload for stego images formed according to HUGO and SUNI methods by usage of features (eqs. (10) to (13)) are represented at Fig. 5 and Fig. 6 respectively.

Obtained results (Fig. 5-6) for VISION dataset are similar to results obtained for ALASKA dataset (Fig. 1-2). The negligible dissimilarity between error range for both datasets can be explained differences in images noises level. Therefore, we may conclude that revealed improving of stegdetector performance does not connect with dataset-specific features.

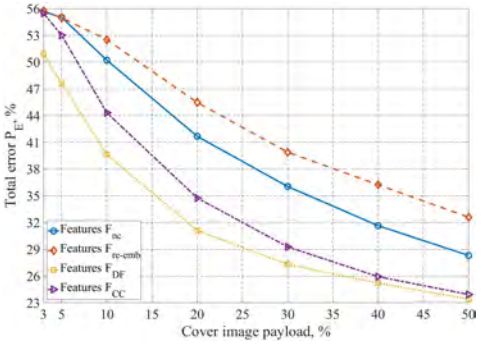
Revealed decreasing of detection error shown that message re-embedding even by another steganographic method leads to significantly amplification of CI distrotions caused stego formation. It is notable that the amplification is achived even by applying known HUGO and S-UNIWARD methods to stego images formed according to advanced MG (Fig. 3) and MiPOD (Fig. 4) methods. Therefore, the message re-embedding is analogue to well-known side-channel technique in digital image steganography — usage pre-cover by steganographer for additional masking the distorions caused by message hiding. For the steganalysis, message re-embedding into analyzed image can be represented as utilization by steganalytic the side-channel — to use similar sets of CI pixels as was used by stego bits embedding. This effect can be used in blind steganalysis when steganalytics do not have access to embedding module.

8. Conclusion

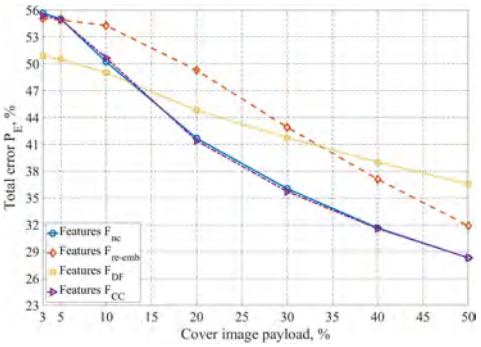
The paper devoted to performance analysis of statistical stegdetectors in case of image calibration by



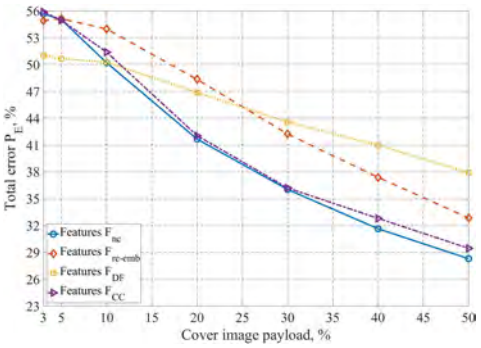
(a) Message re-embedding with HUGO method



(b) Message re-embedding with S-UNIWARD method

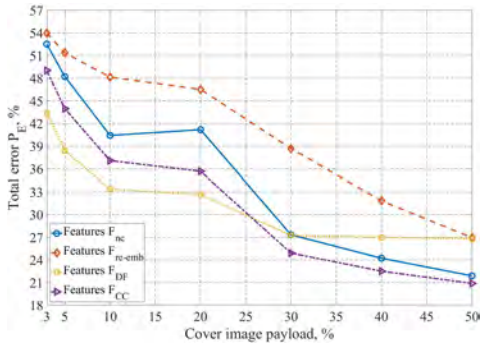


(c) Message re-embedding with MG method

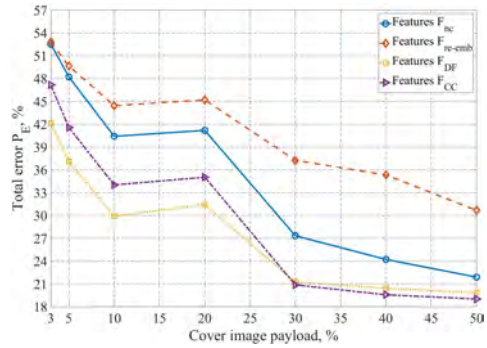


(d) Message re-embedding with MiPOD method

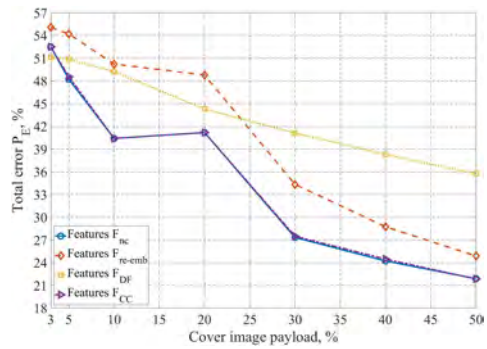
Fig. 5. Dependencies of total error P_E on cover image payload for stego images formed according to HUGO embedding method on VISION dataset. The stegdetector was tuned with usage of non-calibrated features $\mathbf{F}_{nc}$ (solid lines), features of calibrated image $\mathbf{F}_{re-emb}$ (dashed lines), linearly transformed features of calibrated image $\mathbf{F}_{DF}$ (dotted lines) and cartesian calibrated features $\mathbf{F}_{CC}$ (dash-dot lines).



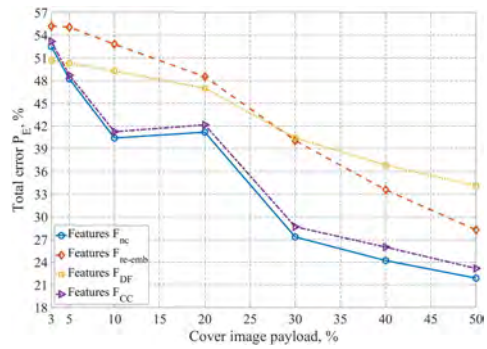
(a) Message re-embedding with HUGO method



(b) Message re-embedding with S-UNIWARD method



(c) Message re-embedding with MG method



(d) Message re-embedding with MiPOD method

Fig. 6. Dependencies of total error P_E on cover image payload for stego images formed according to S-UNIWARD embedding method on VISION dataset. The stegdetector was tuned with usage of non-calibrated features $\mathbf{F}_{nc}$ (solid lines), features of calibrated image $\mathbf{F}_{re-emb}$ (dashed lines), linearly transformed features of calibrated image $\mathbf{F}_{DF}$ (dotted lines) and cartesian calibrated features $\mathbf{F}_{CC}$ (dash-dot lines).

message re-embedding with randomly chosen payload. We obtained further results during verification of proposed approach on stegdetector based on SPAM model of cover image:

1) Message re-embedding into cover and stego images allows considerably improving stegdetectors performance (up to 20%) even in case of low payload of cover image (less than 10%). The biggest decreasing of detection error was achieved in case of stego data re-embedding with usage of same steganographic method that was applied for forming stego images.
2) The value of detection error decreasing significantly varies for considered adaptive embedding methods. The biggest improvement of stegdetector accuracy was achieved for HUGO and S-UNIWARD (up to 20%), while for advanced MG and MiPOD methods the gain was about 7%. This became possible due to usage of proposed linearly transformed features instead of widely used Cartesian calibrated ones.

In the future, we would like to investigate the accuracy of stegdetector in case of message re-embedding with multiple steganographic methods. We also plan to investigate the influence of source-target domain mismatch problem on effectiveness of this approach, where the major challenge would be features adaptation to new image sources.

References

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

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

[3] D. Progonov and V. Lucenko, "Steganalysis of adaptive embedding methods by message re-embedding into stego images," *Information Technologies & Knowledge*, 2020. under review.

[4] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, pp. 868–882, 4 2012.

[5] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 14, pp. 1181–1193, 5 2018.

[6] T. Denemark, V. Sedighi, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *Proceedings of the IEEE Workshop on Information Forensic and Security*, IEEE, IEEE, 12 2014.

[7] J. Kodovsky and J. Fridrich, "Calibration revisited," in *Proceedings of the 11th ACM workshop on Multimedia and security*, pp. 63–74, ACM, 2009.

[8] V. Holub and J. Fridrich, "Random projections of residuals for digital image steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 8, pp. 1996–2006, 12 2013.

[9] 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.

[10] 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.

[11] T. Denemark and J. Fridrich, "Improving steganographic security by synchronizing the selection channel," in 3^{rd} *ACM IH&MMSec. Workshop* (J. Fridrich, P. Comesana, and A. Alattar, eds.), ACM Press, 6 2015.

[12] T. Filler and J. Fridrich, "Gibbs construction in steganography," *IEEE Trans. Inf. Forensics Security*, vol. 5, pp. 705–720, 12 2010.

[13] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 1, 2014.

[14] V. Sedighi, J. Fridrich, and R. Cogranne, "Content-adaptive pentary steganography using the multivariate generalized gaussian cover model," in *Proceedings of Electronic Imaging, Media Watermarking, Security, and Forensics*, SPIE, 2015.

[15] 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.

[16] R. Cogranne, Q. Gilboulot, and P. Bas, "Moving steganography and steganalysis from the laboratory into the real world," in *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, pp. 45–58, ACM, 2013.

[17] 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.

[18] 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.

[19] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Trans. Inf. Forensics Security*, vol. 7, pp. 432–444, 4 2012.

[20] D. Progonov, "Performance of statistical stegdetectors in case of small number of stego images in training set," in *Proceedings of International Conference "Problems of Infocommunications Science and Technology"*, IEEE, 2020.

[21] D. Shullani, M. Fontani, M. Iuliani, O. A. Shaya, and A. Piva, "Vision: a video and image dataset for source identification," *EURASIP Journal on Information Security*, vol. 2017, p. 15, 10 2017.