IMPROVING SPATIAL IMAGE ADAPTIVE STEGANALYSIS INCORPORATING THE EMBEDDING IMPACT ON THE FEATURE

Chao Xia[1,2], *Qingxiao Guan*[1,2*], *Xianfeng Zhao*[1,2], *Jing Dong*[1,3], *Zhoujun Xu*[4]

[1]State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China
[3]Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[4]Beijing Information Technology Institute, Beijing 100094, China

ABSTRACT

Recently, in order to attack the adaptive steganograhpy more accurately, steganalysis features are associated with the content adaptivity. The adaptive σ version of the steganalysis features incorporates the impact of embedding on the residual to improve the detection. However, this method does not consider whether the embedding impact brings the change on the feature (histogram in the PSRM) which will be utilized by the detectors. Thus, we calculate the expectation of the residual L_1 distortion under the condition when the corresponding stego and cover residual values are within different quantization intervals, which will be accumulated in the histograms. This adaptive steganalytic scheme, with the relative position of the residual value in the quantization interval, only utilizes the residual distortion that leads to the change on the final feature. The experimental results demonstrate the potential of the proposed idea, especially for small payloads. This idea can also be applied to JPEG phase-aware features.

Index Terms— Adaptive steganalysis, adaptive steganography, detection, distinguishable L_1 distortion

1. INTRODUCTION

The purpose of image steganography is to embed the secret information into cover images without arousing a warder's suspicion. In early image steganography, the embedding changes are randomly spread throughout the cover image, such as LSB-MR [1]. Recently, with the advent of STCs (Syndrome-Trellis Codes) coding technique [2], several modern adaptive embedding methods are proposed, such as HUGO (Highly Undetectable steGO) [3], WOW (Wavelet Obtained Weights) [4], UNIWARD (UNIversal WAvelet Relative Distortion) [5], HILL (HIgh-pass, Low-pass, and Low-pass) [6]. Such advanced adaptive schemes encourage modifications to occur in complex textures and noisy regions which are hard to model. Compared to the non-adaptive steganography, hence, the above modern adaptive methods can achieve better security performance against the popular universal steganalysis feature sets, such as SRM (Spatial Rich Model) [7], PSRM (Projected Spatial Rich Model) [8].

With the development of increasingly more sophisticated adaptive steganography, much attention has been paid to designing adaptive steganalytic schemes. Unlike universal steganalysis, adaptive steganalysis, so-called selection-channel-aware steganalysis, utilizes the knowledge of embedding probabilities (the selection channel) to improve the detection of adaptive steganography. In 2014, Tang et al. [9] proposed an adaptive steganalytic scheme, tSRM (thresholded SRM), to detect the WOW algorithm. In tSRM, SRM features are extracted only from those suspicious regions where pixels are assigned high embedding probabilities. The tSRM removes lots of unchanged pixels and substantially improves the detection performance. In [10], Denemark et al. modified SRM into maxSRM by accumulating the maximum of the four pixel embedding probabilities into the co-occurrence of four residual elements. It has been shown that the maxSRM is more effective against most adaptive steganographic schemes. In [11], the authors improved the adaptive features by replacing the embedding probability with the expectation of the residual distortion (σspamPSRM and σmaxSRM) because each element in residual depends on more than one pixel. This idea is extended to JPEG phase-aware features, thus forming $\delta^{1/2}_{uSA}$DCTR, $\delta^{1/2}_{uSA}$GFR and $\delta^{1/2}_{uSA}$PHARM in [12].

For modern image steganalysis, cover and stego images are represented with features formed by histograms (or co-occurrences) of quantized residuals. The difference between the histograms (co-occurrences) of the cover and stego images is used to distinguish them from each other. In some cases, although some residual elements are changed after data embed-

*Corresponding author (E-mail: guanqingxiao@iie.ac.cn). This work was supported by the NSFC under U1536105 and U1636102, and National Key Technology R&D Program under 2014BAH41B01, 2016YFB0801003 and 2016QY15Z2500.

ding, they still fall into their original quantization intervals. Those changes of the residual values do not lead to the impact on the histogram (co-occurrence), thus making no contribution to the distinction between cover and stego images. Thus, it is more reasonable to design adaptive steganalytic schemes considering the impact of embedding on the feature. But the adaptive σ version of steganalysis features in [11] evaluates the expectation of the residual distortion, which does not consider whether the residual distortion leads to the impact on the final feature. To further improve the detection, with the quantization step and the given residual value, we propose the expectation of the so-called distinguishable L_1 distortion with which the corresponding cover and stego residual values fall into different quantization intervals, thus only taking into account the residual distortion that brings the change on the feature.

This modification is straightforward for linear residuals. But, as mentioned in [11], there is a major complication for non-linear residuals due to the demand to compute marginals of a high-dimensional probability mass function. Due to limited space, we in this paper only focus on the linear residuals in PSRM (spamPSRM) which have been used in [11]. The rest of this paper is organized as follows. In the next section, we introduce a brief review of the spamPSRM, and its σ-version. In Section 3, we describe the proposed quantity that will be accumulated in the histograms. In Section 4, experiments are conducted to test our idea. Conclusions are drawn in the fifth section.

2. RELATED WORK

In this section, we briefly review the concepts of the spamPSRM [8] and the σspamPSRM [11] to make this paper self-contained and better understand our proposed method. An $n_1 \times n_2$ grayscale cover and stego image will be denoted as $\mathbf{X}$, $\mathbf{Y} \in \{0, ..., 255\}^{n_1 \times n_2}$, respectively.

2.1. spamPSRM

The PSRM is one of the most effective steganalysis feature sets in the spatial domain. Note that the 1980-dimensional spamPSRM is a subset of the PSRM, corresponding to linear ('spam' type) residuals.

For a given image $\mathbf{X}$ ($\mathbf{Y}$), a linear residual projection $\mathbf{Z} = (z_{ij})$ in the spamPSRM is formed as follows.

$$\begin{aligned} \mathbf{Z}^{(\mathrm{pred})} &= \mathbf{K}^{(\mathrm{pred})} \star \mathbf{X} - \mathbf{X} \triangleq \mathbf{K}' \star \mathbf{X}, \\ \mathbf{Z} &= \mathbf{\Pi} \star (\mathbf{K}' \star \mathbf{X}) = (\mathbf{\Pi} \star \mathbf{K}') \star \mathbf{X} \triangleq \mathbf{K} \star \mathbf{X}, \end{aligned} \tag{1}$$

where $'\star'$ denotes the convolution. First, the noise residual $\mathbf{Z}^{(\mathrm{pred})}$ is obtained using local linear pixel predictors $\mathbf{K}^{(\mathrm{pred})}$ to suppress the image content largely. Then, the linear residual projection $\mathbf{Z}$ is generated by convolving the noise residual with the projection kernel $\mathbf{\Pi}$. To diversify the features, the spamPSRM selects ν random projection kernels $\mathbf{\Pi} \in \mathbb{R}^{k \times l}$, where k and l are uniformly randomly selected from $\{1, \ldots, s\}$. The elements of $\mathbf{\Pi}$ are $k \cdot l$ independent realizations of a standard normal random variable $\mathcal{N}(0, 1)$ normalized to a unit Frobenius norm $\|\mathbf{\Pi}\|_2 = 1$.

For a linear residual projection $\mathbf{Z}$, the histogram is obtained using the following formula,

$$\mathbf{h}(l) = \sum_{i,j} \left[Q_T \left(|z_{ij}| / q \right) = l + 1/2 \right], \; l \in \{0, \ldots, T-1\}, \tag{2}$$

where Q_T is a quantizer with $T+1$ centroids $\{1/2, \ldots, T + 1/2\}$, T is the truncation threshold, q is the quantization step, and $[P]$ is the Iverson bracket equal to 0 when the statement P is false and 1 when P is true.

The parameter setup for the spamPSRM is described in [8]. The number of projections per residual $\nu = 55$, the maximum projection matrix size $s = 8$, the quantization step $q = 1$, and the histogram threshold $T = 3$. This setup gives the spamPSRM the dimension of 1980.

2.2. σspamPSRM

From (1), it can be seen that the residual projection $\mathbf{Z}$ in the spamPSRM depends linearly on the pixels of the image because both $\mathbf{\Pi}$ and $\mathbf{K}'$ are linear kernels.

$$z_{ij}(\mathbf{X}) = \sum_{k,l} K_{kl} x_{i-k, j-l}, \tag{3}$$

where the indices k and l depend on the kernel support.

The message embedding is equivalent to adding noise to covers. For adaptive steganographic schemes, the distribution of the noise depends on the pixel location,

$$y_{ij} = x_{ij} + \xi_{ij}, \tag{4}$$

where ξ_{ij} are independent random variables taking their values in {-1, 0, 1} with probabilities β_{ij},$1 - 2\beta_{ij}$,β_{ij}. Thus, each element of the difference $\mathbf{Z}(\mathbf{Y}) - \mathbf{Z}(\mathbf{X})$, $z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X}) = \sum_{k,l} K_{kl} \xi_{i-k, j-l}$, is a random variable with

$$E\left[z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X}) \right] = 0, \tag{5}$$

$$Var\left[z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X}) \right] = 2 \sum_{k,l} K_{kl}^2 \beta_{i-k, j-l}. \tag{6}$$

It is assumed that $z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X})$ is a zero-mean Gaussian random variable with variance (6). In this case, the expectation of the residual L_1 norm is evaluated as

$$\begin{aligned} &E\left[|z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X})| \right] \\ &= \sqrt{\frac{4}{\pi} \sum_{k,l} K_{kl}^2 \beta_{i-k, j-l}} \triangleq \sqrt{\frac{2}{\pi}} \sigma_{ij} \propto \sigma_{ij}. \end{aligned} \tag{7}$$

The parameter σ_{ij} is proportional to the expectation of the L_1 norm which is used to measure the impact of embedding

on the residual. Thus, σ_{ij} is accumulated into the histograms in the spamPSRM. For the σspamPSRM, the histogram of a residual projection is expressed as

$$\mathbf{h}^{\sigma}(l) = \sum_{i,j} \sigma_{ij} \left[Q_T \left(|z_{ij}| / q \right) = l + 1/2 \right], l \in \{0, \ldots, T-1\}. \tag{8}$$

3. PROPOSED METHOD

In the σspamPSRM, the contribution of the elements in residual projection to the histograms does depend on the value of σ. The parameter σ is proportional to $E\left[|z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X})|\right]$ and used to evaluate the impact of embedding on the residual but not considering the change on the final feature (histogram in the spamPSRM). For example, two elements (Z_a and Z_b) in the linear residual projection of a cover image have the same σ, but their relative positions within a quantization interval are different, the value of element Z_a in the middle of a quantization interval and Z_b close to the boundary of an interval. After embedding, it seems obvious that the element Z_b is more likely to fall into another quantization interval (altering the original histogram) than Z_a. That is, although Z_a and Z_b have the same σ, the embedding in Z_b has a greater impact on the histogram. As the steganalysis detectors are built based on the difference between cover and stego features, it is need to take into account whether the change on the feature is brought. In this section, when calculating the expectation of the L_1 norm, we only consider the residual distortion that results in the change on the feature. The details will be described as follows.

For the sake of simplicity, we assume that $z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X})$ is a uniform random variable from a to b, $\mathcal{U}(a, b)$. From its mean (5) and variance (6), it is easy to compute the parameters a and b,

$$\begin{cases} \dfrac{a+b}{2} &= E\left[z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X})\right] = 0 \\ \dfrac{(b-a)^2}{12} &= Var\left[z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X})\right] \triangleq \sigma_{ij}^2 \end{cases} \tag{9}$$
$$\Longrightarrow \begin{cases} a &= -\sqrt{3}\sigma_{ij} \\ b &= \sqrt{3}\sigma_{ij} \end{cases}.$$

If the given image is a cover, $z_{ij}(\mathbf{X})$ is a known real number and can be obtained using (3). Thus, the corresponding one, $z_{ij}(\mathbf{Y})$, conforms to a uniform distribution $\mathcal{U}(m, n)$, where $m = z_{ij}(\mathbf{X}) - \sqrt{3}\sigma_{ij}$, $n = z_{ij}(\mathbf{X}) + \sqrt{3}\sigma_{ij}$. The probability density function of $z_{ij}(\mathbf{Y})$ can be expressed as

$$f(y) = \begin{cases} \left(2\sqrt{3}\sigma_{ij}\right)^{-1}, & m < y < n \\ 0, & \text{otherwise} \end{cases}. \tag{10}$$

If $z_{ij}(\mathbf{X})$ and $z_{ij}(\mathbf{Y})$ are in the same quantization interval, there is no difference between the histogram features of the cover and stego images. Only when $z_{ij}(\mathbf{Y})$ falls into another interval, can the embedding changes have effect on the final histogram which is utilized by the detector. Thus, we only consider the condition when $z_{ij}(\mathbf{X})$ and $z_{ij}(\mathbf{Y})$ are not in the same quantization interval. That is, the expectation of the distinguishable L_1 distortion, E_{DL_1}, can be calculated as

$$\begin{aligned} E_{ij}^{DL_1} &= E\left[|z_{ij}(\mathbf{X}) - z_{ij}(\mathbf{Y})| \left| I_{z_{ij}(\mathbf{X})} \neq I_{z_{ij}(\mathbf{Y})} \right.\right] \\ &= \int_{y \in \left[(m,n) \cap \overline{I_{\pm z_{ij}(\mathbf{X})}}\right]} \frac{|y - z_{ij}(\mathbf{X})|}{2\sqrt{3}\sigma_{ij}} dy, \end{aligned} \tag{11}$$

where $I_{z_{ij}(\mathbf{X})}$ and $I_{z_{ij}(\mathbf{Y})}$ denote the quantization intervals of $z_{ij}(\mathbf{X})$ and $z_{ij}(\mathbf{Y})$, respectively. If $z_{ij}(\mathbf{X})$ is known, the quantization interval of $z_{ij}(\mathbf{X})$ is also known, $I_{z_{ij}(\mathbf{X})} = [M, N)$. $\overline{I_{\pm z_{ij}(\mathbf{X})}}$ represents all the other quantization intervals except $I_{z_{ij}(\mathbf{X})}$ and $I_{-z_{ij}(\mathbf{X})}$ due to the sign-symmetry, namely $(-\infty, -N] \cup (-M, M) \cup [N, +\infty)$.

For example, as shown in Fig. 1, $(m, n) \cap \overline{I_{\pm z_{ij}(\mathbf{X})}} = [N, n)$, $z_{ij}(\mathbf{X}) = \mu_{ij}$, and the E_{DL_1} can be calculated as

$$E_{ij}^{DL_1} = \int_N^n \frac{y - \mu_{ij}}{2\sqrt{3}\sigma_{ij}} dy = \frac{n - N}{2\sqrt{3}\sigma_{ij}} \left(\frac{1}{2}(n + N) - \mu_{ij} \right), \tag{12}$$

where $\frac{1}{2}(n + N) - \mu_{ij} = \frac{\sqrt{3}\sigma_{ij} - \mu_{ij} + N}{2} \propto \sigma_{ij}$ reflects the expectation of L_1 distortion, and the part of $\frac{n-N}{2\sqrt{3}\sigma_{ij}}$ represents the probability that $z_{ij}(\mathbf{Y})$ and $z_{ij}(\mathbf{X})$ are not within the same interval, which is the difference from the σspamPSRM. Thus, E_{DL_1} can reflect the case where the histogram is altered.

Since the distribution of $z_{ij}(\mathbf{Y}) - z_{ij}(\mathbf{X})$ is zero-mean and symmetrical about y axis, if the given image is a stego, $z_{ij}(\mathbf{Y})$ is a known real value and $z_{ij}(\mathbf{X})$ is uniform on $\left(z_{ij}(\mathbf{Y}) - \sqrt{3}\sigma_{ij},\ z_{ij}(\mathbf{Y}) + \sqrt{3}\sigma_{ij}\right)$, analogically. Accordingly, the E_{DL_1} can be calculated as

$$E_{ij}^{DL_1} = \int_{x \in \left[(m,n) \cap \overline{I_{\pm z_{ij}(\mathbf{Y})}}\right]} \frac{|x - z_{ij}(\mathbf{Y})|}{2\sqrt{3}\sigma_{ij}} dx. \tag{13}$$

The E_{DL_1} version of histogram can be calculated by accumulating the quantity E_{DL_1} into the histogram,

$$\mathbf{h}^{E_{DL_1}}(l) = \sum_{i,j} E_{ij}^{DL_1} \left[Q_T \left(|z_{ij}| / q \right) = l + 1/2 \right], \tag{14}$$
$$l \in \{0, \ldots, T-1\}.$$

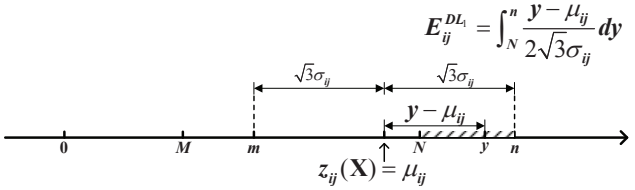

Fig. 1. Example of the expectation of the distinguishable L_1 distortion.

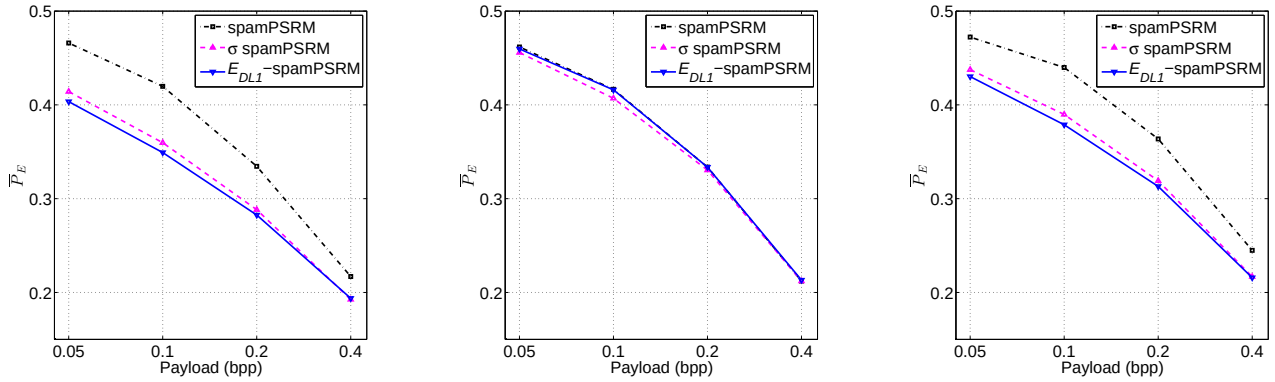

Fig. 2. $\overline{P}_E$ for WOW (left), S-UNIWARD (middle) and HILL (right) with spamPSRM, σspamPSRM and E_{DL_1}-spamPSRM.

The parameters ($\nu = 55$, $s = 8$, $q = 1$, $T = 3$) are set the same as in the spamPSRM and the σspamPSRM to keep the same feature dimensionality.

Fig. 3. Embedding probability for payload 0.4 bpp using WOW (top right), S-UNIWARD (bottom left) and HILL (bottom right) for a 128 $\times$ 128 crop of '1013.pgm' from BOSSbase (top left).

4. EXPERIMENTAL RESULT

Numerous experiments are conducted to test the effectiveness of the proposed E_{DL_1} version of the spamPSRM. In the experiments, 10000 512×512 grayscale images from BOSSbase are used as cover images. Three advanced adaptive steganographic schemes WOW, S-UNIWARD and HILL are used to generate stego images with different embedding rates. The detection accuracy is quantified using $P_{\rm E} = \min_{P_{\rm FA}} \frac{1}{2}(P_{\rm FA} + P_{\rm MD})$, where $P_{\rm FA}$ and $P_{\rm MD}$ are the false-alarm and missed-detection probabilities. The $\overline{P}_{\rm E}$ is averaged over ten random 5000/5000 database splits. The FLD ensemble classifier [13] is used as the detector in the training and testing stages.

From Fig. 2, it can be seen that the proposed adaptive feature set, E_{DL_1}-spamPSRM, achieves much better performance than the non-adaptive spamPSRM for embedding schemes WOW and HILL. Depending on the payload size, the improvement ranges from 2.32% to 6.25% for WOW and from 2.91% to 4.21% for HILL. With the increase of payload, the embedding probability is more spread out and the embedding schemes are less adaptive, thus reducing the improvement. When compared with the adaptive σspamPSRM, our feature set can detect WOW and HILL more accurately. The proposed feature set can improve on the σspamPSRM by 1.10% for HILL with 0.1 bpp and by 1.07% for WOW with 0.05 bpp. However, in the case of S-UNIWARD, the E_{DL_1}-spamPSRM does not achieve improvements. This is likely because S-UNIWARD's adaptivity is weaker than WOW and HILL (see Fig. 3). Although the topic of this paper is steganalysis, it is interesting to find that S-UNIWARD is more resistant to adaptive steganalysis features.

5. CONCLUSIONS

Steganalysis of adaptive steganography needs to incorporate the content adaptivity within the features. Considering the detectors built as classifiers trained on cover and stego features (histograms in the spamPSRM), it is necessary for adaptive steganalysis features to take into account the quantization effect used in the calculation of the histograms. In this paper, thus, we propose the expectation of L_1 distortion under the condition when $z_{ij}(\mathbf{X})$ and $z_{ij}(\mathbf{Y})$ are not in the same quantization interval. Experimental results show the advantage of our proposed idea for highly adaptive algorithms WOW and HILL. It is meaningful to extend this idea to SRM (with nonlinear residuals) and JPEG phase-aware features (DCTR [14], PHARM [15] and GFR [16]). The assumption of modeling the residual distortion as a uniform random variable is strict and particular, and thus we will make use of more realistic assumptions (e.g., gaussian assumption) in the future.

6. REFERENCES

[1] J. Mielikainen, "LSB matching revisited," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 285–287, May 2006.

[2] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 920–935, September 2011.

[3] T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Proceedings of the 12th International Conference on Information Hiding*, 2010, pp. 161–177.

[4] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2012, pp. 234–239.

[5] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 2014, no. 1, pp. 1–13, 2014.

[6] B. Li, M. Wang, J. Huang, and X. Li, "A new cost function for spatial image steganography," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4206–4210.

[7] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, June 2012.

[8] V. Holub and J. Fridrich, "Random projections of residuals for digital image steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 1996–2006, December 2013.

[9] W. Tang, H. Li, W. Luo, and J. Huang, "Adaptive steganalysis against WOW embedding algorithm," in *Proceedings of the 2nd ACM Workshop on Information Hiding and Multimedia Security*, 2014, pp. 91–96.

[10] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *2014 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2014, pp. 48–53.

[11] T. Denemark, J. Fridrich, and P. Comesaña-Alfaro, "Improving selection-channel-aware steganalysis features," *Electronic Imaging*, vol. 2016, no. 8, 2016.

[12] T. Denemark, M. Boroumand, and J. Fridrich, "Steganalysis features for content-adaptive JPEG steganography," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pp. 1736–1746, August 2016.

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

[14] V. Holub and J. Fridrich, "Low-complexity features for JPEG steganalysis using undecimated DCT," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 219–228, February 2015.

[15] V. Holub and J. Fridrich, "Phase-aware projection model for steganalysis of JPEG images," in *Proc. SPIE*, 2015, vol. 9409, pp. 94090T–94090T–11.

[16] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang, "Steganalysis of adaptive JPEG steganography using 2D Gabor filters," in *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security*, 2015, pp. 15–23.