Boosting Steganalysis with Explicit Feature Maps

Mehdi Boroumand
Binghamton University
Department of ECE
Binghamton, NY 13902-6000
mboroum1@binghamton.edu

Jessica Fridrich
Binghamton University
Department of ECE
Binghamton, NY 13902-6000
fridrich@binghamton.edu

ABSTRACT

Explicit non-linear transformations of existing steganalysis features are shown to boost their ability to detect steganography in combination with existing simple classifiers, such as the FLD-ensemble. The non-linear transformations are learned from a small number of cover features using Nyström approximation on pilot vectors obtained with kernelized PCA. The best performance is achieved with the exponential form of the Hellinger kernel, which improves the detection accuracy by up to 2–3% for spatial-domain content-adaptive steganography. Since the non-linear map depends only on the cover source and its learning has a low computational complexity, the proposed approach is a practical and low cost method for boosting the accuracy of existing detectors built as binary classifiers. The map can also be used to significantly reduce the feature dimensionality (by up to factor of ten) without performance loss with respect to the non-transformed features.

Keywords

Steganography, steganalysis, machine learning, explicit feature maps, support vector machine, kernel, Nyström approximation, Hellinger

1. INTRODUCTION

Steganalysis of modern content adaptive steganography [17, 21, 26] requires detectors built as classifiers trained on cover and stego objects represented with rich media models [12, 18, 4, 3, 33, 10, 9]. The prohibitive complexity of training a non-linear classifier in high dimensional feature spaces and on large training sets gave rise to alternative machine learning approaches with lower complexity, such as the FLD-ensemble classifier [25], its linear version [5], regularized linear discriminants [6], and the Online Average Ensemble Perceptron [27]. This works well when the classes of cover and stego features are approximately linearly separable, which seems to be the case in the JPEG domain with features built from co-occurrences of quantized DCT coefficients [23] because of the linear relationship between the features and the embedding domain. In contrast, steganalysis of spatial-domain steganography with co-occurrences of quantized noise residuals benefits from using non-linear classifiers, such as kernelized support vector machines (SVMs). The creators of the popular spatial rich model (SRM) [12] report that a Gaussian SVM trained on carefully selected submodels of the SRM of total dimension 3,300 outperformed the entire 12,753-dimensional SRMQ1 model with the FLD-ensemble classifier (Table II in [12]). Low dimensional variable quantization co-occurrences coupled with a Gaussian SVM were also recently shown to match the performance of the entire SRM with the ensemble classifier [3].[1] Thus, there appears to be an untapped potential to improve steganalysis detectors with non-linear classifiers applied to rich feature sets. What hampers their use in practice is the unfeasibly high computational complexity associated with their training – the complexity of training a kernelized SVM in the primal or dual formulation is $\mathcal{O}(\max\{M, D\} \times \min\{M, D\}^2)$, where M is the number of training examples and D the feature dimension [2].

A kernelized SVM is essentially a linear classifier on features embedded in an infinite dimensional Hilbert space [30]. The classifier can be built thanks to the so-called kernel trick because the training and detector evaluation only require dot products in such space, which can be evaluated using the kernel. The transformation that maps the original features is only *implicit* in the sense that one does not explicitly work with the mapped features. In an alternative approach explored in this paper, the features are transformed using an *explicit* non-linear mapping to improve the classes separability with a hyperplane. Recently, efficient methods have been developed [34, 28] for learning such a non-linear transformation from a portion of the training set. The advantage of this approach is that the classification itself is achieved using a low complexity classifier while the non-linear mapping becomes a mere feature preprocessing. This methodology has found applications, e.g., in object retrieval [1] and digital forensics. In [7], the authors report that applying the square root non-linearity to features in the form of a three-dimensional co-occurrence of the third-

[1]Note that, according to [5] and [6], the non-linearity in the FLD-ensemble is not essential and almost identical performance can be achieved with linear classifiers.

IH&MMSec 2016, June 20–23, 2016, Vigo, Spain.

DOI: http://dx.doi.org/10.1145/2909827.2930803

order noise residual lead to a substantial improvement of their digital image forgery detector.

In this paper, we follow the methodology proposed in [28] and learn the feature map using kernelized PCA coupled with Nyström approximation. For building the map, we explore the Hellinger, linear, chi-square, and Jensen–Shannon kernels and their exponential forms. Based on experiments with individual submodels of the SRM, we identify kernels that provide the biggest detection boost and investigate how the boost depends on the dimensionality of mapped features and the size of the training set. The approach is then scaled to high dimensional feature vectors by learning the mapping separately for each submodel. Experiments with four modern spatial-domain steganographic algorithms on standard databases of grayscale and color images indicate that the detection accuracy can be improved by 2–4% depending on the payload and embedding scheme.

In the next section, we introduce the main idea behind the explicit feature map. In Section 3, we list the kernels used in this study and explain the kernel PCA for learning the transform. A set of initial experiments on two submodels of the SRM is used to gain insight into which kernels are the best performers and assess the boost obtained from them as a function of the dimensionality of the transformed features and the training set size. The procedure for determining the feature map is extended to high-dimensional rich models in Section 5, where we also discuss the results of all experiments with four modern steganographic algorithms and the maxSRMd2 and SCRMQ1 (Spatio-Color Rich Model with $q = 1$) features. A summary of the paper appears in Section 6

2. INTRODUCING THE MAIN IDEA

The problem of steganalysis is binary classification – the Warden monitoring the traffic between Alice and Bob needs to decide whether they exchange information in an overt or covert manner. A useful tool (but not the only one [22]) for the Warden is a detector that can be applied to individual images and provides a binary answer of whether or not a given image contains a secret message. The current paper deals with the problem of building a detector of this type that is as accurate as possible using existing features and classifiers by non-linearly transforming the features as a preprocessing step.

Let us start with a given feature representation, such as the SRM [12]. Assuming the Warden has access to N_{trn} cover images, she embeds them with a specific steganographic method to create a set of N_{trn} corresponding stego images. Then, the cover and stego images are represented with features built as concatenations of histograms or co-occurrences (which are high dimensional histograms): $\mathbf{x}^{(k)} \in \mathbb{R}_+^D$, $k = 1, \ldots, N_{trn}$, for covers and $\mathbf{y}^{(k)} \in \mathbb{R}_+^D$ for stego images, where D is the feature dimensionality and $\mathbb{R}_+$ is the set of non-negative real numbers. As the next step, a binary classifier is trained to distinguish between cover and stego features. One of the best choices for the classifier is kernelized SVM with a positive semi-definite kernel $k : \mathbb{R}_+^D \times \mathbb{R}_+^D \to \mathbb{R}_+$. One can think of such an SVM as a *linear* classifier in a space of features transformed into a Hilbert space $\mathcal{H}$, $\varphi : \mathbb{R}_+^D \to \mathcal{H}$, endowed with dot product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}_+$ for which

$$\langle \varphi(\mathbf{x}), \varphi(\mathbf{y}) \rangle = k(\mathbf{x}, \mathbf{y}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D. \quad (1)$$

The main principle behind SVMs stems from the fact that the transformation φ is only implicit in the sense that when building the SVM classifier or evaluating the detector, one does not need to work directly with $\varphi(\mathbf{x}) \in \mathcal{H}$ and only needs to evaluate dots product via the kernel (1).

As pointed out in the introduction, kernelized SVMs outperform linear classifiers (and the FLD-ensemble) in many typical setups in steganalysis of spatial-domain steganography. Their drawback is a high training complexity, which is why the community resorted to simpler machine learning paradigms. In this paper, we explore an alternative approach, which employs an *explicit* transform $\varphi : \mathbb{R}_+^D \to \mathbb{R}^E$ that *approximates* a given kernel in combination with a simple classifier in $\mathbb{R}^E$ trained on features $\varphi(\mathbf{x}^{(k)})$ and $\varphi(\mathbf{y}^{(k)})$, $k = 1, \ldots, N_{trn}$. To classify a feature $\mathbf{z} \in \mathbb{R}_+^D$, the classifier is presented with $\varphi(\mathbf{z})$. The mapping φ will be learned from a portion of the training set as shown in the next section.

3. LEARNING THE TRANSFORM

In this section, we first introduce several kernels that will be investigated in this paper. Then, we show that the problem of finding a mapping that approximates the kernel with dot products of transformed features coincides with kernelized principal component analysis (kPCA). The general mapping of the feature space is realized using Nyström approximation.

3.1 Kernels

A kernel is a symmetric positive semi-definite[2] mapping $k : \mathbb{R}_+^D \times \mathbb{R}_+^D \to \mathbb{R}_+$ that, loosely speaking, measures the *similarity* between two features. Let us assume that vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ are L_2-normalized, meaning that $\|\mathbf{x}\|_2^2 = \|\mathbf{y}\|_2^2 = \sum_{i=1}^D x_i^2 = 1$. Their square Euclidean distance can be written as:

$$\|\mathbf{x} - \mathbf{y}\|_2^2 = 2\left(1 - k(\mathbf{x}, \mathbf{y})\right), \quad (2)$$

where we introduced $k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D x_i y_i$. The reader recognizes $k(\mathbf{x}, \mathbf{y})$ as the classical dot product, which, due to the normalization, coincides with the cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$.

Generalizing this idea, the following are popular choices for kernels in machine vision [28, 34]:

1. Linear kernel $k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D x_i y_i$ with $\mathbf{x}$ and $\mathbf{y}$ L_2-normalized;
2. Hellinger kernel (also called Bhattacharyya kernel) $k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D \sqrt{x_i y_i}$ with $\mathbf{x}$ and $\mathbf{y}$ L_1-normalized;
3. Chi-square kernel $k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D \frac{2x_i y_i}{x_i + y_i}$ with $\mathbf{x}$ and $\mathbf{y}$ L_1-normalized;
4. Jensen–Shannon kernel $k(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \sum_{i=1}^D x_i \log \frac{x_i + y_i}{x_i} + y_i \log \frac{x_i + y_i}{y_i}$ with $\mathbf{x}$ and $\mathbf{y}$ L_1-normalized.

Whenever a term in the chi-square and the Jensen–Shannon kernel is not defined (due to division by zero or log of zero), the term is set to zero, which coincides with the limit from

[2]Kernel k is positive semi-definite if for any n and any $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(n)} \in \mathbb{R}^D$, the $n \times n$ matrix $K_{ij} = k(\mathbf{u}^{(i)}, \mathbf{u}^{(j)})$ is positive semi-definite.

the right. Also note that with the specified normalization, $0 \le k(\mathbf{x},\mathbf{y}) \le 1$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}_+^D$ and for all kernels. The Hellinger kernel corresponds to the linear kernel on square-rooted features.

It can be easily proved that for a symmetric positive semi-definite kernel k and $\gamma > 0$,

$$e^{\gamma(k(\mathbf{x},\mathbf{y})-1)} \tag{3}$$

is also symmetric positive semi-definite and bounded $0 \le e^{\gamma(k(\mathbf{x},\mathbf{y})-1)} \le 1$. Thus, the above four kernels have their exponential counterparts, which we name with the preposition 'exp', such as exp-Hellinger, etc.

3.2 Finding the transformation

The task of finding a transform such that the dot products of two transformed vectors coincide with the kernel evaluated on them can be formulated as follows. Given $M \ge D$ vectors $\mathbf{x}^{(1)},\dots,\mathbf{x}^{(M)} \in \mathbb{R}_+^D$ for training the map φ, find vectors $\phi(\mathbf{x}^{(i)}) \in \mathbb{R}^M$ so that for all $i,j \in \{1,\dots,M\}$:

$$k(\mathbf{x}^{(i)},\mathbf{x}^{(j)}) \approx \phi(\mathbf{x}^{(i)})\cdot\phi(\mathbf{x}^{(j)}). \tag{4}$$

This can be solved by the following optimization problem. Denoting the ath coordinate of $\phi(\mathbf{x}) \in \mathbb{R}^M$ with $\phi_a(\mathbf{x})$, $1 \le a \le M$, minimize

$$\sum_{i,j=1}^{M}\left(k(\mathbf{x}^{(i)},\mathbf{x}^{(j)}) - \phi(\mathbf{x}^{(i)})\cdot\phi(\mathbf{x}^{(j)})\right)^2 \tag{5}$$

subject to

$$\sum_{i=1}^{M}\phi_a(\mathbf{x}^{(i)})\phi_b(\mathbf{x}^{(i)}) = 0 \text{ for all } 0 \le a \ne b \le M. \tag{6}$$

The constraint (6) expresses our desire that the description in the M dimensional space be non-redundant – we essentially require each pair of coordinates a,b of the transformed feature vectors be uncorrelated.

Using the method of Lagrange multipliers, it is easily established that $\phi_a \triangleq (\phi_a(\mathbf{x}^{(1)}),\dots,\phi_a(\mathbf{x}^{(M)}))' \in \mathbb{R}^M$ are eigen-vectors of the kernel matrix $\mathbf{K} = (K_{ij}) \in \mathbb{R}_+^{M\times M}$, $K_{ij} = k(\mathbf{x}^{(i)},\mathbf{x}^{(j)})$:

$$\mathbf{K}\phi_a = \lambda_a^2\phi_a,\ 1 \le a \le M, \tag{7}$$

where λ_a^2 are the corresponding eigenvalues sorted from the largest to the smallest. We note that $\lambda_a = \|\phi_a\|_2$.

The mapping $\varphi:\ \mathbb{R}_+^D \to \mathbb{R}^E$ is defined using the so-called Nyström approximation. For any $\mathbf{z} \in \mathbb{R}_+^D$,

$$\varphi_a(\mathbf{z}) = \frac{1}{\lambda_a^2}\mathbf{K}(\mathbf{z},\cdot)\phi_a,\ 1 \le a \le E, \tag{8}$$

where

$$\mathbf{K}(\mathbf{z},\cdot) = \left(k(\mathbf{z},\mathbf{x}^{(1)}),\dots,k(\mathbf{z},\mathbf{x}^{(M)})\right). \tag{9}$$

Note that in building φ, we retain the first E coordinates a corresponding to the largest eigenvalues λ_a^2. When $E = D$, the feature transform preserves the feature dimensionality.

The Hellinger kernel corresponds to the linear kernel on L_1-normalized features that have been square-rooted elementwise. This is the only non-linear kernel for which the explicit map φ adopts a simple closed-form expression – the square root executed elementwise. Because of this simplification, the feature preprocessing is very cheap and the complexity is essentially negligible in comparison to the classifier training. This is why in our experiments, we include results obtained with square-rooted features. They should always be very close to the results obtained with the transform learned for the Hellinger kernel. Square-rooting features has been reported in computer vision [1] and digital forensics [7] as a way to improve detection accuracy with a heuristic justification that the non-linear transformation evens out the differences between the individual features (histogram bins).

We furthermore note that it is possible to use only cover features for the map training rather than cover-stego feature pairs because the kernel is continuous and the features of cover and stego images are close and would not provide good constraints for learning the map. We verified this experimentally but do not report the details of these findings in this paper.

3.3 Complexity considerations

The complexity of learning the map includes the time needed to form the matrix $\mathbf{K}$, which is $\mathcal{O}(DM^2)$ and the cost of solving the eigenvector problem (7), which is $\mathcal{O}(M^3)$ if implemented, e.g., using the Cholesky decomposition. Thus, the total complexity is $\mathcal{O}(DM^2 + M^3)$. Fortunately, these computations only need to be executed once for a given cover source because the map is trained on cover images only. Of course, this makes the map independent of the embedding payload and the steganographic scheme. The cost of transforming a new feature is part of the training as well as testing and is $\mathcal{O}(MD)$ to evaluate (9) and then $\mathcal{O}(ME)$ to compute all E coordinates of $\varphi(\mathbf{z})$ (8).

4. INITIAL EXPERIMENTS

To get a feeling for the ability of explicit maps to boost steganalysis and to assess the influence of various parameters, such as E, the number of retained coordinates in the map, and M, the number of images for training the map, in this section we experiment with S-UNIWARD [21] and HILL [26] and their detection with two submodels of the SRM: the four-dimensional co-occurrence matrix of the 'SQUARE 3x3' submodel (also sometimes called "KB residual") and the 'minmax22h' submodel for the first-order residual. Both feature sets were computed with the quantization step $q = 1$ and symmetrized as in SRM, which means that the KB residual co-occurrence had dimensionality 169 while the 'minmax22h' submodel had dimensionality of 101 after removing from it elements that are always equal to zero (see Section 4.1).

The experiments in this section were conducted on BOSSbase 1.01 [11] containing 10,000 512×512 8-bit grayscale images. After randomly splitting the database into two disjoint parts of equal size (5,000 images), the feature transformation $\varphi:\ \mathbb{R}_+^D \to \mathbb{R}^E$ was learned on M randomly selected images from the training set. The FLD-ensemble was then trained and tested on the transformed features. This was repeated ten times while evaluating the empirical security using the minimal total error under equal priors achieved on the testing set:

$$P_{\mathrm{E}} = \min_{P_{\mathrm{FA}}}\frac{1}{2}(P_{\mathrm{FA}} + P_{\mathrm{MD}}), \tag{10}$$

where P_{FA} and P_{MD} are the probabilities of false alarm and missed detection. The symbol $\overline{P}_{\mathrm{E}}$ is the average P_{E} over the ten splits. The statistical spread is reported using the mean absolute deviation. The constant γ in exponential kernels (3) was chosen as the reciprocal of the mean of the non-exponential kernel over all training pairs:

$$\gamma = \frac{1}{\frac{1}{M^2}\sum_{i,j=1}^{M} k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}. \tag{11}$$

4.1 Removing zero features

Before explaining the results of all experiments in this section, we make one note. Exactly 224 elements of the 325-dimensional 'minmax22h' submodel of SRM are zero in both cover and stego images. This is due to the nature of the residuals used in this submodel and the scan direction for forming the co-occurrence.[3] For example, the 'minmax22v' submodel does not have any zeros. Besides 'minmax22h', there are two other submodel types in the SRM with elements that are identically equal to zero no matter what the input image is. They are 'minmax34h' and 'minmax41'. The zeros occur for first-order differences and third-order differences. Because first-order differences are quantized only with $q = 1$ and $q = 2$, there are total 2×3 first-order submodels and 3×3 third-order submodels (because third-order residuals are quantized with three quantization steps), each effectively containing only $325 - 224 = 101$ non-zero elements instead of 325. These zeros need to be removed before learning the transformation because Matlab eigenvector solver may otherwise return negative eigenvalues and complex-valued eigen-vectors due to finite machine precision. We note that after removing the zero elements from the SRM feature vector, its dimensionality becomes $34{,}671 - 15 \times 224 = 31{,}311$. The dimensionality of the SRMQ1 decreases from 12,753 to 11,409 (6×224 fewer).

Because the selection-channel-aware version of the SRM called maxSRMd2 [10] uses a different scan for forming co-occurrences, the so-called 'd2' scan,[4] the number of non-zero elements in the above-mentioned submodels is different. For quantization $q = 2$ the 'minmax22h', 'minmax34h', and 'minmax41' submodels for the first and third order residuals have dimensionality 190. For $q = 1$ and $q = 1.5$, their dimensionality is 120. This gives the maxSRMd2 feature set a dimensionality of 32,016.

Finally, we note that this peculiarity largely escaped the attention of the community because the ensemble classifier first prunes the cover and stego features and automatically removes zero elements from both cover and stego features before building the classifier. In our case, however, because we learn the transformation φ before applying the ensemble, we remove such zero features prior to learning φ.

[3] Anecdotal evidence exists among researchers that the SRM feature contains many zeros but, according to the best knowledge of the authors, this issue has never been investigated in detail. In this paper, we merely state the true dimensionality of the affected submodels without providing any further analysis in order not to digress from the main topic of this paper.

[4] The 'd2' scan involves residuals $r_{ij}, r_{i,j+1}, r_{i+1,j+2}, r_{i+1,j+3}$ and three more horizontally and vertically flipped versions.

$E \backslash M$	10	20	50	100	169	350	500	1000
10	.4260	.4262	.4260	.4259	.4268	.4259	.4260	.4263
20	-	.3893	.3893	.3899	.3894	.3887	.3899	.3894
50	-	-	.3142	.3145	.3144	.3135	.3144	.3136
100	-	-	-	.2885	.2882	.2879	.2886	.2879
169	-	-	-	-	.2754	.2757	.2757	.2752
350	-	-	-	-	-	.2707	.2683	.2695
500	-	-	-	-	-	-	.2662	.2653
1000	-	-	-	-	-	-	-	.2739

Table 1: $\overline{P}_{\mathrm{E}}$ for KB residual ($D = 169$) with S-UNIWARD at 0.4 bpp when training the map φ with exp-Hellinger kernel on M randomly selected images and retaining E dimensions, $M, E \in \{10, 20, 50, 100, 169, 350, 500, 1000\}$. The statistical spread in the form of sample standard deviation ranges between 0.0012 and 0.0039. The values below the main diagonal are not achievable because $E \leq M$.

4.2 Boosting submodels

As our first experiment, we investigated the effect of the parameters M and E on detection error $\overline{P}_{\mathrm{E}}$. For brevity, we only report the result with the KB residual and S-UNIWARD at 0.4 bpp with the exponential Hellinger kernel for the non-linear map φ. Table 1 shows the detection error for different combinations of E and M. There appears to be a benefit in retaining more coordinates E than the original feature dimensionality D. With $M = 500$ training images, retaining $E = 500$ coordinates rather than $D = 169$ leads to about 1% improvement. There does not seem to be any benefit in using more images for training or retaining more than 500 coordinates. We note that our goal was to boost the detection accuracy without increasing the feature dimensionality. Inspecting the row in Table 1 corresponding to $E = D = 169$, the number of training images does not have a major effect on detection accuracy as long as $M \geq D$.

Table 2 shows the results for all kernels introduced in Section 2 with two submodels and two steganographic techniques at payload 0.4 bpp for $M = 500$ and $E = 500$. The first row of the table is the performance obtained using the original features as they appear in SRM. The second row shows the results with simply square-rooting the features. The third row contains detection errors obtained using Gaussian SVM, again averaged over ten 50/50 database splits.

As noted in the previous section, the square rooted features correspond to the Hellinger kernel when using the known explicit map instead of the Nyström approximation. It is comforting to discover that these results match those obtained using Nyström approximation with this kernel (row four) as they should. Inspecting the remaining rows, it is very clear that the exponential kernels are superior to the non-exponential ones and they also have quite similar performance. The boost they provide w.r.t. the original features (row 1) is up to 3.5% for the 'minmax22h' submodel for S-UNIWARD. Also, it is quite apparent that simply square-rooting the features (Hellinger kernel) is far from the best option. For the KB residual, explicit maps with exponential kernels match the result obtained using G-SVM. For the 'minmax22h' residual, the G-SVM outperforms explicit maps by about 1%. Because of the larger complexity associated with the chi-square and Jensen–Shannon kernels, we selected the exponential Hellinger kernel for all experiments with rich models in the next section.

	Kernel	minmax22h (S-UNI)	KB (S-UNI)	KB (HILL)
1	Original	0.3293±0.0030	0.2933±0.0028	0.3281±0.0027
2	Square root	0.3150±0.0032	0.2812±0.0028	0.3228±0.0020
3	G-SVM	0.2841±0.0018	0.2618±0.0023	0.3026±0.0016
4	Hellinger	0.3252±0.0038	0.2827±0.0035	0.3242±0.0024
5	Exp-Hellinger	0.2984±0.0030	0.2626±0.0024	0.3026±0.0030
6	Linear	0.3317±0.0030	0.2837±0.0029	0.3378±0.0028
7	Exp-linear	0.3052±0.0030	0.2603±0.0029	0.3030±0.0039
8	Chi-square	0.3063±0.0030	0.2685±0.0035	0.3106±0.0035
9	Exp-chi-square	0.2991±0.0023	0.2619±0.0024	0.3045±0.0019
10	JS	0.3076±0.0039	0.2696±0.0033	0.3103±0.0026
11	Exp-JS	0.2950±0.0023	0.2611±0.0022	0.3017±0.0016

Table 2: $\overline{P}_{\mathrm{E}}$ for various kernels for S-UNIWARD and HILL at 0.4 bpp for KB co-occurrence (dim 169) and 'minmax22h' (dim 101) submodel of the SRM. The abbreviation 'JS' stands for the Jensen–Shannon kernel.

5. EXTENSION TO RICH MODELS

The purpose of this section to extend the proposed approach to high-dimensional rich models. Since the complexity of training the non-linear map is $\mathcal{O}(DM^2 + M^3)$, see Section 3.3, it would not be feasible to train the map for the entire rich feature vector. Instead, we learn the map for each submodel of the rich model separately. Furthermore, in order not to increase the feature dimensionality, we keep the number of retained coordinates $E = D$. We do so despite the benefit of using $E > D$ (see Table 1) because, when richified, we did not see any benefit of inflating the dimensionality of the entire feature vector.[5]

Given a set of N_{trn} cover images and the same amount of the corresponding stego images for training the entire detector, we randomly reserved $M < N_{trn}$ cover images for training the maps for all submodels. The classifier is next trained on all N_{trn} images, including the images used for training the map. Including the M images for classifier training is unlikely to lead to over training because the map training is not informed about the stego class. Moreover, we determined experimentally that as few as $M = 350$ cover images are sufficient for the map training, which is only a small part of the training set ($N_{trn} = 5,000$ for datasets derived from BOSSbase 1.01). Finally, we did carry out comparative tests in which we only trained on $N_{trn} - M$ images, which resulted in similar detection errors within their statistical spread.

We note that the map φ has the following data structures as its parameters:[6] 1) the set of M cover features $\mathbf{x}^{(i)}$, $i = 1, \ldots, M$, which is an $M \times D$ dimensional array, 2) the set of E eigen-vectors ϕ_a, $a = 1, \ldots, E$, stored as an $E \times D$ array.

In our experiments, we tested four state-of-the-art steganographic methods embedding in the spatial domain: WOW [17], S-UNIWARD [21], HILL [26], and MiPOD [31]. For each payload, a separate binary classifier implemented with the FLD-ensemble [25] was trained on the original features and on the transformed features. For each split of the database into a training and testing set, the map φ was retrained on a different subset of the training set. The empirical security was measured as the total detection error P_{E} (10) averaged over ten 50/50 database splits. In all experiments, we used $E = D$. Based on the experiments with individual submodels in the previous section, we tested only one kernel, the exponential Hellinger. We remind that each feature vector was L_1-normalized.

We first report the results for the maxSRMd2 feature set [10] on BOSSbase 1.01. Table 3 shows $\overline{P}_{\mathrm{E}}$ as a function of payload for the original maxSRMd2, its square rooted version, and the transformed version using exp-Hellinger on BOSSbase 1.01. To better contrast the improvement in detection, in Figure 1 we show the difference between the detection error of the original features, $\overline{P}_{\mathrm{E}}^{(orig)}$, and the error obtained using square-rooting, $\overline{P}_{\mathrm{E}}^{(sqrt)}$, and with exp-Hellinger, $\overline{P}_{\mathrm{E}}^{(exp-H)}$. The difference is expressed in percents (multiplied by 100). The results indicate that a consistent detection boost is obtained across all four embedding algorithms. The biggest boost was obtained for WOW and the smallest for MiPOD and HILL. The square rooting is not as effective as the transform obtained with the exponential Hellinger kernel.

At this point, we note that when applying the non-linear map to the SRM feature set we observed a gain that was very similar to that of the maxSRMd2 set, which is why we do not report these results here. In Section 5.1 below, we comment on other rich feature sets currently used in steganalysis.

As our second batch of experiments, we used the Spatio-Color Rich Model with $q = 1$ (SCRMQ1) [14] feature set of dimensionality 18,157. It is a merger of the SRMQ1, which is a subset of the SRM and the Color Rich Model (CRM) formed by three-dimensional co-occurrences of residuals across three color channels. The image source was the same as in [14], a color version of BOSSbase prepared as follows. Starting with the full-resolution raw images, we converted them using the same script that was used for creating the BOSSbase with the following modifications. The output of 'ufraw' (ver. 0.18 with 'dcraw' ver. 9.06) was changed to the color ppm format instead of the pgm grayscale. Also, all calls of 'convert' used ppm for the output as well as for resizing so that the smaller image dimension was 512 and for central cropping to 512×512. As in the original script, the resizing algorithm uses the Lanczos kernel. We thus obtained 10,000 true color 512×512 ppm images. This version of color BOSSbase will be called 'BOSSbaseColor'.

The above four steganographic algorithm were applied by color channels and the same relative payload was embedded in each channel. The complete results are listed in Table 4. The non-linear map boosts detection to a different degree depending on the steganographic method and payload. The largest gain of almost 4% is observed for WOW.

5.1 Application to other rich feature sets

In this section, we comment on our experience with applying non-linear maps to other types of rich models. Modern embedding algorithms for JPEG images (J-UNIWARD [21] and UED [15, 16]) are currently best detected with phase-aware rich models [19, 20] formed by histograms of noise residuals split by their location with respect to the location of the 8×8 pixel grid used for compression. In particular, the so-called Gabor Filter Residuals (GFR) [32] made aware of the selection channel [8] appear among the best. Experiments with this feature set on selected payloads on J-

[5] Experiments on selected payloads with the tested stego algorithms with $E = 500$ retained coordinates for each submodel resulted in statistically insignificant deviations from $E = D$ (not shown in this paper).

[6] For simplicity, we assume that the M training examples were selected as the first M features from the training set.

	Payload (bits per pixel)					
S-UNI	0.05	0.1	0.2	0.3	0.4	0.5
maxSRMd2	0.4168±0.0024	0.3652±0.0008	0.2919±0.0023	0.2374±0.0023	0.1917±0.0042	0.1569±0.0035
Square root	0.4177±0.0033	0.3588±0.0025	0.2851±0.0034	0.2276±0.0021	0.1785±0.0033	0.1433±0.0026
exp-Hellinger	0.4178±0.0020	0.3608±0.0033	0.2803±0.0027	0.2181±0.0028	0.1720±0.0020	0.1348±0.0025
HILL						
maxSRMd2	0.4246±0.0040	0.3742±0.0022	0.3105±0.0033	0.2580±0.0033	0.2196±0.0039	0.1815±0.0033
Square root	0.4188±0.0030	0.3669±0.0032	0.3007±0.0025	0.2512±0.0036	0.2116±0.0026	0.1736±0.0030
exp-Hellinger	0.4191±0.0022	0.3653±0.0024	0.2974±0.0028	0.2451±0.0024	0.2004±0.0019	0.1649±0.0031
MiPOD						
maxSRMd2	0.4427±0.0026	0.3949±0.0031	0.3246±0.0034	0.2709±0.0027	0.2272±0.0037	0.1865±0.0029
Square root	0.4401±0.0028	0.3926±0.0047	0.3185±0.0022	0.2635±0.0027	0.2209±0.0036	0.1818±0.0022
exp-Hellinger	0.4426±0.0032	0.3911±0.0038	0.3148±0.0026	0.2568±0.0024	0.2104±0.0028	0.1720±0.0031
WOW						
maxSRMd2	0.3574±0.0024	0.2984±0.0020	0.2331±0.0018	0.1907±0.0028	0.1559±0.0024	0.1279±0.0030
Square root	0.3492±0.0021	0.2854±0.0033	0.2140±0.0031	0.1702±0.0026	0.1375±0.0020	0.1118±0.0033
exp-Hellinger	0.3470±0.0024	0.2820±0.0024	0.2094±0.0025	0.1645±0.0031	0.1310±0.0028	0.1068±0.0032

Table 3: Detection error $\overline{P}_{\mathrm{E}}$ for four steganographic schemes and five payloads in bpp on BOSSbase 1.01 with FLD-ensemble trained with maxSRMd2 features, their square rooted form, and transformed using exponential Hellinger kernel (by submodels).

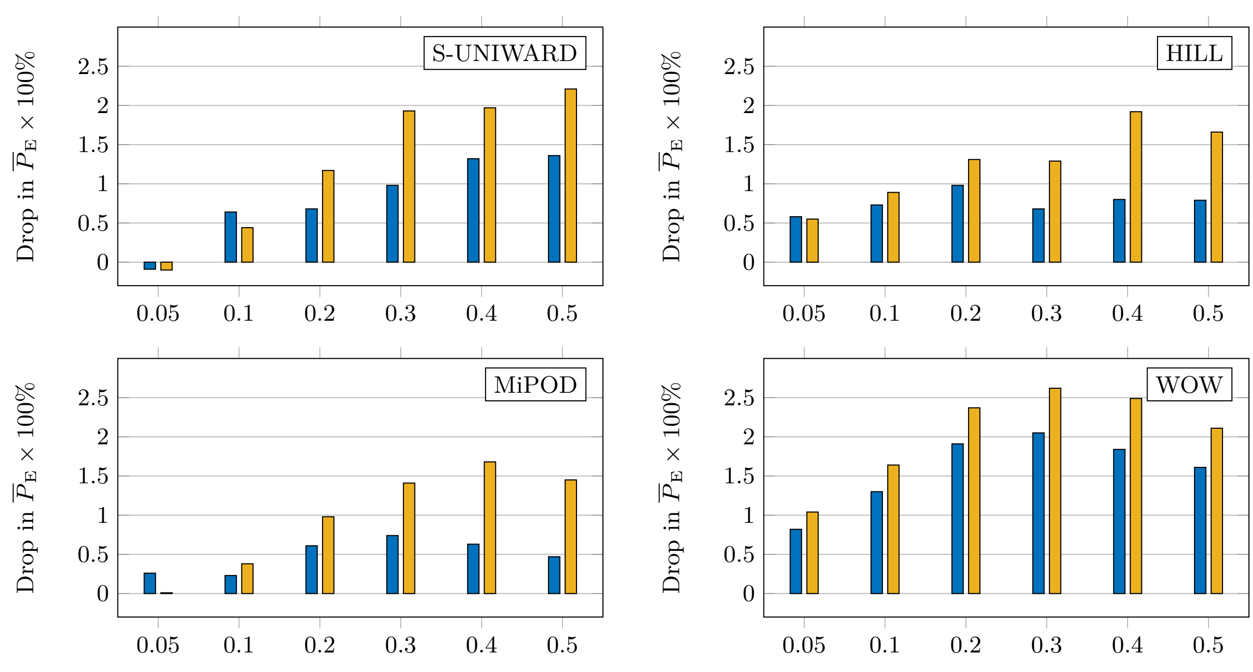

Figure 1: Drop in detection error $\overline{P}_{\mathrm{E}} \times 100\%$ with respect to the original maxSRMd2 feature set as a function of payload in bpp. Blue: $\left(\overline{P}_{\mathrm{E}}^{(orig)} - \overline{P}_{\mathrm{E}}^{(sqrt)}\right) \times 100$, yellow: $\left(\overline{P}_{\mathrm{E}}^{(orig)} - \overline{P}_{\mathrm{E}}^{(exp-H)}\right) \times 100$.

	Payload (bpp per channel)					
S-UNI	0.05	0.1	0.2	0.3	0.4	0.5
SCRMQ1	0.4549±0.0022	0.3939±0.0026	0.2977±0.0027	0.2216±0.0016	0.1710±0.0032	0.1306±0.0038
Square root	0.4499±0.0028	0.3853±0.0029	0.2885±0.0023	0.2154±0.0017	0.1630±0.0027	0.1230±0.0032
exp-Hellinger	0.4487±0.0047	0.3789±0.0031	0.2761±0.0037	0.2016±0.0017	0.1461±0.0033	0.1067±0.0028
HILL						
SCRMQ1	0.4699±0.0024	0.4227±0.0029	0.3288±0.0022	0.2528±0.0017	0.1967±0.0024	0.1558±0.0043
Square root	0.4586±0.0035	0.4021±0.0031	0.3130±0.0022	0.2416±0.0036	0.1896±0.0025	0.1497±0.0022
exp-Hellinger	0.4520±0.0032	0.3904±0.0044	0.2927±0.0026	0.2212±0.0031	0.1724±0.0027	0.1332±0.0022
MiPOD						
SCRMQ1	0.4557±0.0029	0.4034±0.0029	0.3081±0.0031	0.2397±0.0042	0.1872±0.0045	0.1476±0.0026
Square root	0.4477±0.0022	0.3904±0.0021	0.3006±0.0018	0.2317±0.0028	0.1812±0.0029	0.1439±0.0030
exp-Hellinger	0.4485±0.0031	0.3802±0.0032	0.2839±0.0014	0.2133±0.0034	0.1633±0.0040	0.1253±0.0034
WOW						
SCRMQ1	0.4507±0.0009	0.3975±0.0033	0.2997±0.0033	0.2283±0.0021	0.1793±0.0046	0.1365±0.0036
Square root	0.4367±0.0040	0.3700±0.0042	0.2750±0.0029	0.2092±0.0021	0.1641±0.0011	0.1263±0.0027
exp-Hellinger	0.4296±0.0033	0.3600±0.0029	0.2618±0.0020	0.1936±0.0022	0.1468±0.0019	0.1129±0.0018

Table 4: Detection error $\overline{P}_{\mathrm{E}}$ for four steganographic schemes and five payloads in bits per pixel per color channel on color version of BOSSbase with FLD-ensemble trained with SCRMQ1 features, their square rooted form, and transformed using exponential Hellinger kernel (by submodels).

UNIWARD, however, indicated no benefit of square-rooting the GFR features (this is equivalent to using the Hellinger kernel). Also, we did not observe any boost when applying a non-linear transformation to the projection SRM (PSRM) [18]. The PSRM as well as the phase-aware features share one aspect that is different from rich models such as SRM, maxSRMd2, and SCRMQ1. The former are computed as first-order statistics (histograms) rater than high-dimensional co-occurrences. Histograms are generally much better populated than co-occurrences and the differences among the populations of individual bins are much smaller. Features formed as collections of well-populated histograms do not seem to benefit from non-linear transformation investigated in this paper.

There exist co-occurrence based rich models for the JPEG domain, such as JRM [24] and CC-C300 [23] formed by many two-dimensional co-occurrences from DCT coefficients and their differences. However, as reported in these papers, the decision boundary between cover and stego features within these representations is almost linear because of the direct relationship between the embedding domain and the domain in which the steganalysis representation is built. According to our experiments with the JRM on J-UNIWARD and nsF5 [13], square-rooting the features before classification with the FLD-ensemble has no effect on detection accuracy.

5.2 Rich model compactification

In this section, we study whether the transform φ is a useful tool for compactifying the rich model by simply retaining fewer coordinates in each transformed submodel. Compactification of rich models is a topic that has already been addressed in [12], where a greedy forward feature selection method has been applied to SRM submodels. It has also been investigated within the context of unsupervised detectors, where high-dimensional models could not be applied [29]. We stress that retaining a smaller subset of transformed dimensions is similar in spirit to applying a regular PCA to cover features and, as such, has obvious limitations because of the absence of feedback from the embedding scheme. Thus, it is unlikely to provide compactification ratios similar to approaches that consider both cover and stego features, such as calibrated least squares (CLS) [29]. On the other hand, the compactification only depends on the cover source, which makes the approach potentially useful for unsupervised universal steganalysis.

As can be seen from Table 1, for an individual SRM submodel the detection error increases quite rapidly with the decreased number of retained coordinates. On the other hand, differences in the performance of individual submodels usually do not scale directly to the rich model as it is likely that the submodels "compensate for each other weaknesses." Thus, the entire rich model may still perform rather well when compacted. Figure 2 confirms this hypothesis, showing the detection error $\overline{P}_{\mathrm{E}}$ as a function of the number of retained coordinates. Even when retaining only 10% of the coordinates, $E = 0.1 \times D$, there still appears to be a small gain in detection accuracy w.r.t. the original maxSRMd2 feature.

6. CONCLUSION

Supervised detectors of steganography are currently built using classifiers trained on high-dimensional rich models. The excessive training complexity associated with large training sets and high-dimensional features forced steganalysts to adopt simple(r) machine learning paradigms, such as the popular FLD-ensemble and its linearized versions, potentially thus losing on detection accuracy that could be obtained with more powerful non-linear classifiers, such as kernelized support vector machines. In this paper, we investigate the possibility to boost steganalysis with simple classifiers by non-linearly transforming the features. The transformation is learned on a small set of cover features with the constraint that the dot products of mapped features approximate the output of a specific kernel, a task equivalent to kernelized PCA. The feature transformation can be in-

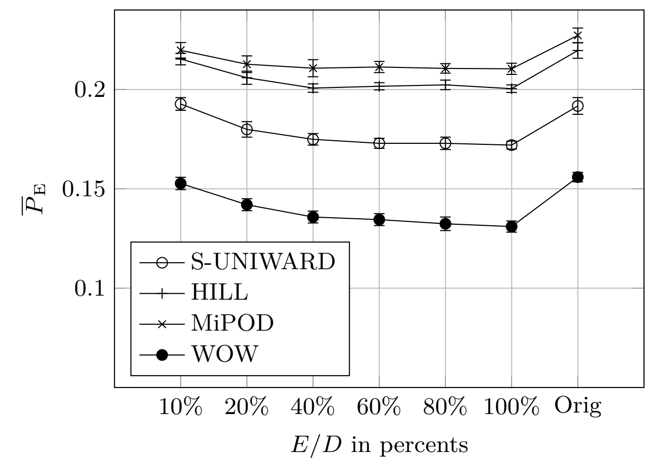

Figure 2: Detection error $\overline{P}_{\mathrm{E}}$ as a function of the relative number of retained coordinates, E/D. Tested payload 0.4 bpp, exp-Hellinger kernel.

terpreted as a different way of measuring distances in the feature space. Retaining only a subset of transformed coordinates corresponding to the largest eigenvalues, the general version of the transformation is obtained using Nyström approximation. The approach is scaled up to the full spatial rich model by learning the transformation separately for each submodel in order to keep the computational complexity low.

Exponential forms of the linear, Hellinger (Bhattacharyya), chi-square, and Jensen–Shannon kernels provide similar performance and substantially improve upon the original (non-transformed) form of the features. A consistent gain between 2–4% was observed for the selection-channel-aware maxSRMd2 features as well as the Spatio-Color Rich Model for steganalysis of color images. The detection improvement varies across steganographic methods and payloads. Learning the transformation is a relatively low-cost task that only needs to be executed once for a given cover source. In particular, the transformation does not depend on the steganographic method and the payload. By retaining fewer dimensions in each SRM submodel, it is possible to compactify the rich descriptor by a factor of 10 without losing the detection performance of the original (non-transformed) feature vector. This could be useful for unsupervised universal steganalysis detectors.

We wish to point out that the non-linear transformation seems effective only for features built as high-dimensional co-occurrences, such as the SRM, maxSRM, and SCRMQ1. In particular, it does not bring any improvement for "dense" features built as histograms spanning a few bins, such as JPEG-phase-aware features [32, 20, 19] and the projection spatial rich model [18]. We hypothesize that it is because the populations of co-occurrence bins are typically highly imbalanced while the bins in histograms are more evenly populated, making the effect of the non-linearity negligible.

7. ACKNOWLEDGMENTS

The work on this paper was supported by Air Force Office of Scientific Research under the research grant number FA9950-12-1-0124. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of AFOSR or the U.S. Government. The authors would like to thank anonymous reviewers for their insightful comments.

8. REFERENCES

[1] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2911–2918, June 2012.

[2] O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 15(5):1155–1178, 2007.

[3] L. Chen, Y.-Q. Shi, and P. Sutthiwan. Variable multi-dimensional co-occurrence for steganalysis. In *Digital Forensics and Watermarking, 13th International Workshop, IWDW*, volume 9023, pages 559–573, Taipei, Taiwan, October 1–4 2014. Springer.

[4] L. Chen, Y.Q. Shi, P. Sutthiwan, and X. Niu. A novel mapping scheme for steganalysis. In Y.Q. Shi, H.-J. Kim, and F. Perez-Gonzalez, editors, *International Workshop on Digital Forensics and Watermaking*, volume 7809 of *LNCS*, pages 19–33. Springer Berlin Heidelberg, 2013.

[5] R. Cogranne and J. Fridrich. Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory. *IEEE Transactions on Information Forensics and Security*, 10(2):2627–2642, December 2015.

[6] R. Cogranne, V. Sedighi, T. Pevný, and J. Fridrich. Is ensemble classifier needed for steganalysis in high-dimensional feature spaces? In *IEEE International Workshop on Information Forensics and Security*, Rome, Italy, November 16–19 2015.

[7] D. Cozzolino, G. Poggi, and L. Verdoliva. Splicebuster: a new blind image splicing detector. In *IEEE International Workshop on Information Forensics and Security*, Rome, Italy, November 16–19 2015.

[8] T. Denemark, M. Boroumand, and J. Fridrich. Steganalysis features for content-adaptive JPEG steganography. *IEEE Transactions on Information Forensics and Security*, June 2016. To appear.

[9] T. Denemark and J. Fridrich. Improving selection-channel-aware steganalysis features. In A. Alattar and N. D. Memon, editors, *Proceedings IS&T, Electronic Imaging, Media Watermarking, Security, and Forensics 2016*, San Francisco, CA, February 14–18, 2016.

[10] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich. Selection-channel-aware rich model for steganalysis of digital images. In *IEEE International Workshop on Information Forensics and Security*, Atlanta, GA, December 3–5, 2014.

[11] T. Filler, T. Pevný, and P. Bas. BOSS (Break Our Steganography System). http://www.agents.cz/boss, July 2010.

[12] J. Fridrich and J. Kodovský. Rich models for steganalysis of digital images. *IEEE Transactions on*

Information Forensics and Security, 7(3):868–882, June 2011.

[13] J. Fridrich, T. Pevný, and J. Kodovský. Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities. In J. Dittmann and J. Fridrich, editors, *Proceedings of the 9th ACM Multimedia & Security Workshop*, pages 3–14, Dallas, TX, September 20–21, 2007.

[14] M. Goljan, R. Cogranne, and J. Fridrich. Rich model for steganalysis of color images. In *Sixth IEEE International Workshop on Information Forensics and Security*, Atlanta, GA, December 3–5, 2014.

[15] L. Guo, J. Ni, and Y.-Q. Shi. An efficient JPEG steganographic scheme using uniform embedding. In *Fourth IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5, 2012.

[16] L. Guo, J. Ni, and Y. Q. Shi. Uniform embedding for efficient JPEG steganography. *IEEE Transactions on Information Forensics and Security*, 9(5), 2014.

[17] V. Holub and J. Fridrich. Designing steganographic distortion using directional filters. In *Fourth IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5, 2012.

[18] V. Holub and J. Fridrich. Random projections of residuals for digital image steganalysis. *IEEE Transactions on Information Forensics and Security*, 8(12):1996–2006, December 2013.

[19] V. Holub and J. Fridrich. Low-complexity features for JPEG steganalysis using undecimated DCT. *IEEE Transactions on Information Forensics and Security*, 10(2):219–228, Feb 2015.

[20] V. Holub and J. Fridrich. Phase-aware projection model for steganalysis of JPEG images. In A. Alattar and N. D. Memon, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2015*, volume 9409, San Francisco, CA, February 8–12, 2015.

[21] V. Holub, J. Fridrich, and T. Denemark. Universal distortion design for steganography in an arbitrary domain. *EURASIP Journal on Information Security, Special Issue on Revised Selected Papers of the 1st ACM IH and MMS Workshop*, 2014:1, 2014.

[22] A. D. Ker and T. Pevný. The steganographer is the outlier: Realistic large-scale steganalysis. *IEEE Transactions on Information Forensics and Security*, 9(9):1424–1435, September 2014.

[23] J. Kodovský and J. Fridrich. Steganalysis in high dimensions: Fusing classifiers built on random subspaces. In A. Alattar, N. D. Memon, E. J. Delp, and J. Dittmann, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security and Forensics III*, volume 7880, pages OL 1–13, San Francisco, CA, January 23–26, 2011.

[24] J. Kodovský and J. Fridrich. Steganalysis of JPEG images using rich models. In A. Alattar, N. D. Memon, and E. J. Delp, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2012*, volume 8303, pages 0A 1–13, San Francisco, CA, January 23–26, 2012.

[25] J. Kodovský, J. Fridrich, and V. Holub. Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on Information Forensics and Security*, 7(2):432–444, 2012.

[26] B. Li, M. Wang, and J. Huang. A new cost function for spatial image steganography. In *Proceedings IEEE, International Conference on Image Processing, ICIP*, Paris, France, October 27–30, 2014.

[27] I. Lubenko and A. D. Ker. Steganalysis with mismatched covers: Do simple classifiers help. In J. Dittmann, S. Katzenbeisser, and S. Craver, editors, *Proc. 13th ACM Workshop on Multimedia and Security*, pages 11–18, Coventry, UK, September 6–7 2012.

[28] F. Perronnin, J. Sanchez, and Yan Liu. Large-scale image categorization with explicit data embedding. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2297–2304, June 2010.

[29] T. Pevný and A. D. Ker. The challenges of rich features in universal steganalysis. In A. Alattar, N. D. Memon, and C. Heitzenrater, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2013*, volume 8665, pages 0M 1–15, San Francisco, CA, February 5–7, 2013.

[30] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 2001.

[31] V. Sedighi, R. Cogranne, and J. Fridrich. Content-adaptive steganography by minimizing statistical detectability. *IEEE Transactions on Information Forensics and Security*, 11(2):221–234, 2016.

[32] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang. Steganalysis of adaptive JPEG steganography using 2D Gabor filters. In P. Comesa na, J. Fridrich, and A. Alattar, editors, *3rd ACM IH&MMSec. Workshop*, Portland, Oregon, June 17–19, 2015.

[33] W. Tang, H. Li, W. Luo, and J. Huang. Adaptive steganalysis against WOW embedding algorithm. In A. Uhl, S. Katzenbeisser, R. Kwitt, and A. Piva, editors, *2nd ACM IH&MMSec. Workshop*, pages 91–96, Salzburg, Austria, June 11–13, 2014.

[34] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492, March 2012.