Random Projections of Residuals as an Alternative to Co-occurrences in Steganalysis

Vojtěch Holub, Jessica Fridrich, and Tomáš Denemark

Department of ECE, SUNY Binghamton, NY, USA

ABSTRACT

Today, the most reliable detectors of steganography in empirical cover sources, such as digital images coming from a known source, are built using machine-learning by representing images with joint distributions (co-occurrences) of neighboring noise residual samples computed using local pixel predictors. In this paper, we propose an alternative statistical description of residuals by binning their random projections on local neighborhoods. The size and shape of the neighborhoods allow the steganalyst to further diversify the statistical description and thus improve detection accuracy, especially for highly adaptive steganography. Other key advantages of this approach include the possibility to model long-range dependencies among pixels and making use of information that was previously underutilized in the marginals of co-occurrences. Moreover, the proposed approach is much more flexible than the previously proposed spatial rich model, allowing the steganalyst to obtain a significantly better trade off between detection accuracy and feature dimensionality. We call the new image representation the Projection Spatial Rich Model (PSRM) and demonstrate its effectiveness on HUGO and WOW – two current state-of-the-art spatial-domain embedding schemes.

Keywords: Steganalysis, co-occurrence, residual, projection, classification, PSRM

1. INTRODUCTION

Digital images exhibit short-range but quite complex statistical dependencies among individual elements – pixels or transform coefficients. These dependencies are modified by steganographic embedding changes, which can be well modeled as a low-amplitude noise that may be adaptive to the host image content. A common way to detect steganography is to first suppress the content and model only the noise component (the noise residual), which has the effect of increasing the signal to noise ratio between the stego signal and the host media and obtaining a signal with a much smaller dynamic range. The residual is typically obtained by subtracting from each pixel its prediction based on its immediate neighborhood without utilizing the pixel value itself.

Once the noise residual is obtained, it is represented in a more compact statistical form* usually as a sample joint probability distribution (co-occurrence matrix) of adjacent residual samples. In the past, authors have repeatedly convincingly demonstrated that higher-order co-occurrences are generally better capable of detecting steganographic changes [2, 3, 10], where the optimal value of the order is determined by the cover source. In general, sources with longer-range dependencies will likely be better detected with higher order co-occurrences. Since the dimensionality of the co-occurrence increases exponentially with its order, steganalysts had to quantize the residual, and sometimes quite harshly, to obtain a reasonably low-dimensional and statistically significant descriptor for subsequent machine learning.

Modern content-adaptive steganographic embedding schemes constrain their embedding changes to textures and edges, where the values of the noise residuals are typically large and end up in the boundary bins (marginals) of the co-occurrences, which limits the detection accuracy one can achieve in practice. To somehow alleviate this problem, several authors independently proposed to employ an entire family of noise residuals [2, 3] to increase the chances of detecting embedding changes in complex content.

To the best knowledge of the authors of this article, the current state-of-the-art image descriptor for steganalysis of digital images in the spatial domain is the Spatial Rich Model (SRM) [2], which contains 39 (106)

E-mail: {vholub1,fridrich,tdenema1}@binghamton.edu; `http://dde.binghamton.edu`

*This can be interpreted as a user-controlled ad hoc dimensionality reduction.

submodels in the form of co-occurrence matrices compactified using symmetries of natural scenes. Since the number of elements in each co-occurrence increases exponentially with its order, rich models tend to be high dimensional. Fortunately, with increasing dimensionality, the decision boundary separating cover and stego features tends to be more linear, which permits using simpler machine learning paradigms, including the ensemble [7] or on-line classifiers, such as the perceptron and its variants [8]. Although training in high-dimensional spaces is no longer an obstacle, the high dimensionality complicates the execution of other tasks, such as feature selection and optimization of the feature space parameters. Moreover, in practical applications feature extraction starts dominating the computational burden, making real-time detection of steganography more difficult.

In this article, we explore alternative statistical descriptors of noise residuals. Instead of forming co-occurrences of neighboring residual samples, we project them on random directions and use the first-order statistic (the histogram) of the projections as steganalysis features. This brings several advantages over the representation based on co-occurrences. Since the projections are one-dimensional, it is possible to employ finer binning and larger thresholds to make use of information contained in the tails affected by content-adaptive steganography. Thus, we expect a boost in detection accuracy especially for highly-adaptive embedding schemes, such as WOW [4]. Second, by using very large projection neighborhoods one can potentially model long-range dependencies among pixels. Third, by selecting many different neighborhood shapes, the statistical description can be further diversified, which is likely to improve the detection accuracy. Finally, a much greater design flexibility is obtained since the size and shape of the projection neighborhoods, the number of projection vectors, as well as the histogram bins can be incrementally adjusted to achieve a desired trade-off between detection accuracy and feature dimensionality.

After introducing notation and basic concepts as well as the common core of all experiments, in Section 3, we briefly describe the local pixel predictors that will be used for computing noise residuals.[†] Then, in Section 4 we introduce the concepts of a projection neighborhood and projection vectors and explain the procedure for obtaining statistical image descriptors based on random projections of residuals. An initial exploratory analysis with selected residuals is carried out in Section 5. Here, we investigate how the number, size, and shape of projection neighborhoods and the number of projection vectors affect the detection accuracy. It will become apparent in this section already, that for content-adaptive algorithms descriptors based on projections can outperform co-occurrences at the same feature dimensionality by a quite significant margin. In Section 6, we scale up the approach to an entire rich model, which we call PSRM (Projection SRM), and in Section 7 we test its ability to detect steganalysis on HUGO [11] and WOW [4] on a range of payloads. Section 8 contains a summary and an outline of future effort.

2. PRELIMINARIES

High-dimensional arrays, matrices, and vectors will be typeset in boldface and their individual elements with the corresponding lower-case letters in italics. The calligraphic font is reserved for sets. The symbols $\mathbf{X} = (x_{ij}) \in \mathcal{X} = \mathcal{I}^{n_1 \times n_2}$ and $\mathbf{Y} = (y_{ij}) \in \mathcal{X}$, $\mathcal{I} = \{0, \dots, 255\}$, will always represent pixel values of 8-bit grayscale cover and stego images with $n = n_1 \times n_2$ pixels. For a set of L centroids, $\mathcal{Q} = \{q_1, \dots, q_L\}$, $0 \le q_1 \le \dots \le q_L$, a scalar quantizer is defined as $Q_{\mathcal{Q}}(x) \triangleq \arg\min_{q \in \mathcal{Q}} |x - q|$. The Iverson bracket $[S] = 1$ when the statement S is true and 0 otherwise. The vector operator "$\cdot$" stands for a dot product between two vectors.

2.1 Common core of all experiments

All experiments in this paper are carried out on BOSSbase 1.01 [1]. This database contains 10,000 images acquired by eight digital cameras in the RAW format (CR2 or DNG) and subsequently processed by converting to 8-bit grayscale, resizing, and cropping to the size of 512×512 pixels. The script for this processing is also available from the BOSS competition web site.

The classifiers we use are all instances of the ensemble proposed in [7] and available from `http://dde.binghamton.edu/download/ensemble`. They employ Fisher linear discriminants as base learners trained on random subspaces of the feature space. The out-of-bag (OOB) estimate of the testing error, E_{OOB}, on bootstrap samples of the training set is used to automatically determine the random subspace dimensionality and the

[†]They are essentially all taken from [2].

number of base learners as described in [7]. We also use E_{OOB} to report the detection performance since it is an unbiased estimate of the testing error. We train a separate classifier for each image source model, embedding method, and payload.

3. NOISE RESIDUALS

As explained in the introduction, we utilize an existing set of noise residuals incorporated in the SRM [2] but represent them using a different statistical descriptor. Thus, we limit this section to be rather brief and refer the reader to the original publication for more details.

Denoting an estimate of the cover image pixel x_{ij} from the neighborhood $\mathcal{N}(\mathbf{Y}, i, j)$ of pixel y_{ij} as $\theta(\mathcal{N}(\mathbf{Y}, i, j))$, the noise residual is,

$$\mathbf{Z} = (z_{ij}), \quad z_{ij} = \theta(\mathcal{N}(\mathbf{Y}, i, j)) - y_{ij}. \tag{1}$$

Most pixel predictors are realized as shift-invariant finite-impulse response linear filters captured by a kernel matrix. For example, the kernels

$$\mathbf{K}_3 = \frac{1}{4}\begin{pmatrix} -1 & 2 & -1 \\ 2 & 0 & 2 \\ -1 & 2 & -1 \end{pmatrix}, \quad \mathbf{K}_5 = \frac{1}{12}\begin{pmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & 0 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{pmatrix}, \tag{2}$$

proposed in [5] and in [6] predict the value of the central pixel from its local 3×3 and 5×5 neighborhoods using the operation of convolution, which gives the residual the following form:

$$\mathbf{Z} = \mathbf{K} * \mathbf{Y} - \mathbf{Y}. \tag{3}$$

Other kernel predictors involve pixels arranged only in a horizontal (or vertical) direction derived from local constant, linear, and quadratic models of local image patches. A few selected examples of residuals obtained using such polynomial predictors are listed in Table 1.

Local model	s	Horizontal residual $\mathbf{Z}^{(\mathrm{h})} = (z_{ij}^{(\mathrm{h})})$
Constant	2	$y_{i,j+1} - y_{ij}$
Linear	3	$(y_{i,j-1} + y_{i,j+1})/2 - y_{ij}$
Quadratic	4	$(y_{i,j-1} + 3y_{i,j+1} - y_{i,j+2})/3 - y_{ij}$

Table 1. Examples of directional (horizontal) residuals with support size s based on locally polynomial image models.

Numerous other residuals can be formed by taking the minimum (maximum) of the output of two or more linear residuals, e.g., one in the horizontal and one in the vertical direction, $z_{ij}^{(\min)} = \min\{z_{ij}^{(\mathrm{h})}, z_{ij}^{(\mathrm{v})}\}$, $z_{ij}^{(\max)} = \max\{z_{ij}^{(\mathrm{h})}, z_{ij}^{(\mathrm{v})}\}$ [2].

The traditional approach based on co-occurrences now continues with quantizing $\mathbf{Z}$ to a set of centroids $\mathcal{Q} = \{-Tq, (-T+1)q, \ldots, Tq\}$, where $T > 0$ is an integer threshold and $q > 0$ is a quantization step:

$$\mathbf{R} = (r_{ij}), \quad r_{ij} \triangleq Q_{\mathcal{Q}}(z_{ij}), \tag{4}$$

Having obtained the quantized residual $\mathbf{R}$, a co-occurrence matrix of Dth order is formed from D neighboring values of r_{ij} from the entire image. To keep the co-occurrence bins well populated, both D and T are usually kept low. For example, in [2] the authors used $D = 4$, $T = 2$, and $q \in \{1, 2, 3\}$, and formed two versions of a SRM – one with a single quantization step $q = 1$ of dimensionality $12,753$ (SRMQ1) and the full SRM with all three quantizations of dimensionality $34,671$. By choosing such a small value of T and limiting $D = 4$, neither version of the SRM can capture long-range dependencies among residuals. The potentially useful information contained

in the tails of the distribution of z_{ij} is also lost, which limits the detection accuracy of highly adaptive schemes. Since the co-occurrence dimensionality is $(2T+1)^D$, changing the parameters T and D gives the steganalyst rather limited options to control the dimensionality.

In contrast, in this paper we avoid quantizing the residual and, instead, project $\mathbf{Z}$ on random directions and quantize the scalar projections. The details of this procedure are described next.

4. RANDOM PROJECTIONS OF RESIDUALS

To motivate our approach, we provide the following observations collected through extensive experiments on digital images during the past two years. First, it appears that the centroid set $\mathcal{Q}$ in the quantizer $Q_{\mathcal{Q}}$ affects the detection accuracy only marginally [2, 9]. This suggests that the detection utilizes mainly the *shape* of the distribution rather than the individual bins. Thus, the information content in the co-occurrence may be well captured with a parametric model. However, finding a good parametric model for high-dimensional (e.g., 4D or higher) co-occurrences seems rather difficult. Neighboring residual samples appear anti-correlated, and the dependencies spread across all four neighboring samples in a complex manner. Moreover, the model will likely strongly depend on the cover source and post-processing (e.g., resizing). Consequently, we made the decision to keep the representation by binning as it is more universal.

The idea proposed in this article is to represent the joint distribution of neighboring residuals by projecting them on one-dimensional lines, where one can afford finer binning. If the joint distribution of k neighboring residual samples is modeled using a probability density function $\rho(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^k$, forming a histogram of residuals projected on vector $\mathbf{v} \in \mathbb{R}^k$, $\mathbf{v} \neq 0$, is equivalent to representing ρ using its integrals over $k-1$ dimensional hyperplanes $\sigma(\mathbf{x})$ with normal vector $\mathbf{v}$:

$$h(y; \rho, \mathbf{v}) = \int\limits_{\mathbf{x}\cdot\mathbf{v}\in y} \rho(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}), \quad y \in \mathbb{R}. \tag{5}$$

By choosing a continuum of such vectors $\mathbf{v}$, we are reminded of the Radon transform of ρ. Our task, however, is not to reconstruct ρ from its integrals (5) but to perform classification in the projected space. The projections are essentially a different type of marginalization than the truncation used in co-occurrences. Intuitively, when selecting sufficiently many projection vectors $\mathbf{v}$, we improve our ability to distinguish between the distributions of residuals for cover and stego images. An important potential advantage of projections is that we retain more information about the tails, where content-adaptive steganographic methods restrict their embedding changes.

Having provided a motivation for our approach, we now describe in detail the proposed alternative statistical descriptor of noise residuals by projecting neighboring residual samples onto a set of random vectors. To this end, we introduce a few concepts. A projection neighborhood $\mathcal{P}$ is defined as a connected set of adjacent pixels, such as four horizontally neighboring pixels, four diagonally neighboring pixels, a square of 2×2 adjacent pixels, etc. By $\mathbf{Z}_{\mathcal{P}}$, we understand the residual $\mathbf{Z}$ constrained to the projection neighborhood $\mathcal{P}$.

For each projection neighborhood $\mathcal{P}$, we furthermore define the following set of possible projection vectors $\mathcal{V}_{\mathcal{P}} = \{-2, -1, 0, 1, 2\}^{|\mathcal{P}|}$.[‡] We choose the projection vectors $\mathbf{v} \in \mathcal{V}_{\mathcal{P}}$ uniformly pseudo-randomly. Moreover, each element of $\mathbf{v}$ is uniquely associated with an element of $\mathcal{P}$ (and thus $\mathbf{Z}_{\mathcal{P}}$ as well) in some arbitrary but fixed manner (e.g., in a row-by-row fashion). For a given projection vector $\mathbf{v} \in \mathcal{V}_{\mathcal{P}}$, we denote

$$\mathcal{P}(\mathbf{v}) = \{\mathbf{Z}_{\mathcal{P}} \cdot \mathbf{v} | \mathbf{Z}_{\mathcal{P}} \subset \mathbf{Z}\} \tag{6}$$

the set of projections of $\mathbf{v}$ on such $\mathbf{Z}_{\mathcal{P}}$ that "fit into $\mathbf{Z}$." This operation can be implemented as a convolution of $\mathbf{Z}$ with a "projection kernel" $\mathbf{K}(\mathcal{P}, \mathbf{v})$ of the same shape as $\mathcal{P}$ with $\mathbf{v}$ as its elements. For example, for $\mathcal{P}$ shaped as a 2×2 square and $\mathbf{v} = (-1, 0, 1, -2)$, assuming a row-by-row mapping between $\mathcal{P}$ and $\mathbf{v}$, the corresponding kernel is

$$\mathbf{K}(\mathcal{P}, \mathbf{v}) = \begin{pmatrix} -1 & 0 \\ 1 & -2 \end{pmatrix}. \tag{7}$$

[‡]Note that we limited the elements of projection vectors to small integers.

One can obtain more robust statistical descriptors by utilizing the fact that statistical properties of natural images should not change with direction or mirroring. The way these symmetries should be incorporated depends on whether the residual is directional or non-directional. For example, the residuals obtained using the kernels $\mathbf{K}_3$ and $\mathbf{K}_5$ are non-directional, while those listed in Table 1 are directional. For non-directional residuals, one can merge the projections obtained using the following operations: horizontal mirroring of the kernel, $\overleftrightarrow{\mathbf{K}}$, vertical mirroring, $\mathbf{K}\updownarrow$, the version rotated by 180°, $\mathbf{K}^{\circlearrowleft}$:

$$\overleftrightarrow{\mathbf{K}}(\mathcal{P},\mathbf{v}) = \begin{pmatrix} 0 & -1 \\ -2 & 1 \end{pmatrix}, \qquad \mathbf{K}\updownarrow(\mathcal{P},\mathbf{v}) = \begin{pmatrix} 1 & -2 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{K}^{\circlearrowleft}(\mathcal{P},\mathbf{v}) = \begin{pmatrix} -2 & 1 \\ 0 & -1 \end{pmatrix}, \tag{8}$$

and their transposed versions (total eight projection kernels). For directional residuals, one can merge the projections only when properly combining the kernel transformations with the residuals. For example, one can combine the outputs of $\mathbf{K}$, $\overleftrightarrow{\mathbf{K}}$, $\mathbf{K}\updownarrow$, and $\mathbf{K}^{\circlearrowleft}$ applied to the a residual with a horizontal direction and the outputs of all four transposed kernels applied to the vertical version of the residual, which gives us again eight combinations.

The distribution of residuals obtained using a linear filter as in (3) will in general be zero mean and symmetrical about zero. Thus, the distribution of residual projections will also be symmetrical about zero, where its maximum value will also be located. The symmetry of the distribution allows us to work with absolute values of projections and thus decrease the number of bins or, alternatively, choose a higher truncation threshold. In particular, we select a set of quantization centroids

$$\mathcal{Q}_{T,q} = \{q/2, 3q/2, \ldots, (2T+1)q/2\}, \tag{9}$$

which corresponds to a quantizer with $T+1$ bins, T of them with width q, and one spanning the interval $[(2T+1)q/2, \infty)$ – the marginal bin. Working with absolute values of the projections is equivalent to stating that the marginal of residuals of natural images are symmetrical, which is a reasonable assumption. In fact, if a steganographic scheme violated this symmetry, the absolute values of projections would be unable to detect this artifact. However, an embedding scheme creating such an asymmetry would be fundamentally flawed and one could likely construct very accurate targeted quantitative attacks by utilizing this symmetry violation. (A good example is the Jsteg algorithm [12].)

Formally, for a quantizer centroid set $\mathcal{Q}$, the histogram for projection neighborhood $\mathcal{P}$ and projection vector $\mathbf{v}$ is:

$$\mathbf{h}(l; \mathcal{Q}, \mathcal{P}, \mathbf{v}) = \sum_{\mathbf{Z}_{\mathcal{P}} \subset \mathbf{Z}} [Q_{\mathcal{Q}}(|\mathbf{Z}_{\mathcal{P}} \cdot \mathbf{v}|) = l], \quad l \in \mathcal{Q}. \tag{10}$$

Since the distribution of residuals obtained using the operations min (max) is not centered at zero, we cannot use absolute values and, instead, work with an expanded set of centroids:

$$\mathcal{Q}^{(\mathrm{x})}_{T,q} = \mathcal{Q}_{T,q} \cup \{-\mathcal{Q}_{T,q}\}, \tag{11}$$

which has double the cardinality of $\mathcal{Q}_{T,q}$. Because for any finite $\mathcal{R} \subset \mathbb{R}$, $\min \mathcal{R} = -\max\{-\mathcal{R}\}$, we can merge the projections of residuals $\mathbf{Z}^{(\min)}$ and $\mathbf{Z}^{(\max)}$ into one histogram:

$$\mathbf{h}(l; \mathcal{Q}^{(\mathrm{x})}, \mathcal{P}, \mathbf{v}) = \sum_{\mathbf{Z}_{\mathcal{P}} \subset \mathbf{Z}} [Q_{\mathcal{Q}^{(\mathrm{x})}}(\mathbf{Z}^{(\min)}_{\mathcal{P}} \cdot \mathbf{v}) = l] + [Q_{\mathcal{Q}^{(\mathrm{x})}}(-\mathbf{Z}^{(\max)}_{\mathcal{P}} \cdot \mathbf{v}) = -l], \quad l \in \mathcal{Q}^{(\mathrm{x})}. \tag{12}$$

5. EXPLORATORY ANALYSIS

In this section, we carry out a few selected experiments to probe the ability of residual projections to detect steganography and show the potential of this approach. The steganographic algorithm chosen for the experiments is WOW [4] as we expect a bigger improvement for this algorithm than HUGO due to its stronger content adaptivity.

The first experiment was carried out with non-directional residuals obtained using the $\mathbf{K}_3$ and $\mathbf{K}_5$ predictors (the submodel 'SQUARE11' in [2]). In the SRM, the dimensionality of the co-occurrence for each residual after

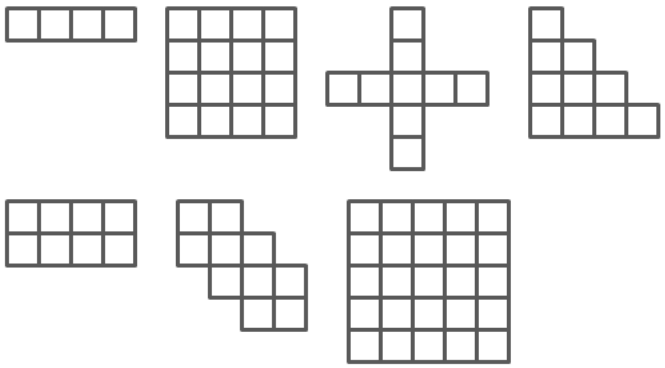

Figure 1. Seven types of projection neighborhoods $\mathcal{P}$ used in our investigative experiments in Section 5: 1×4, 4×4, cross, stairs, 2×4, thick diagonal, and 5×5 square.

k	1	2	3	4	5	6	7
$N_{\mathcal{P}}$	34	17	11	8	6	5	4
$2(T+1)kN_{\mathcal{P}}$	340	340	330	320	300	300	280

Table 2. Dimensionality of the feature vector formed by k projection neighborhoods, each with $N_{\mathcal{P}}$ projection vectors.

symmetrization was 169, giving the dimension of 338 for this submodel consisting of the union of both of these residuals. For a fair comparison between the co-occurrence and the projection descriptors, we need to keep approximately the same dimensionality. This means that the number of projection neighborhoods $\mathcal{P}$, N_p, the number of pseudo-randomly chosen projection vectors for neighborhood $\mathcal{P}_k$, $N_{\mathcal{P}_k}$, and the number of quantization bins $T+1$ must satisfy

$$2(T+1)\sum_{k=1}^{N_p} N_{\mathcal{P}_k} \approx 338. \tag{13}$$

The factor of 2 reflects the fact that we work with two residuals (2) at the same time.

For the projections, we fixed $T = 4$ for $\mathcal{Q}_{T,q}$ (9), which gave us $T+1 = 5$ bins to represent the histogram for each projection neighborhood and vector. Since the projection is a linear combination of weakly dependent $|\mathcal{P}|$ residuals with coefficients $(v_1, \ldots, v_{|\mathcal{P}|}) = \mathbf{v}$, the variance of the projection is approximately $||\mathbf{v}||_2^2$. Thus, we set $q = ||\mathbf{v}||_2$.

Since there are $\binom{N_p}{k}$ possibilities to select $k = 1, \ldots, N_p$ neighborhoods, we ran our test for all $2^{N_p} - 1$ combinations of N_p hand-designed projection neighborhoods displayed in Figure 1. Furthermore, for each k we kept the number of projections $N_{\mathcal{P}_k}$ the same across all neighborhoods. From (13), we determined the following number of projection vectors that would give us the total dimension of approximately 338 for the union of both residuals (see Table 2).

Since for each k, we have $\binom{N_p}{k}$ different values of the E_{OOB}, in Figure 2 (left) we display only the mean, minimum, and maximum E_{OOB} values taken over all $\binom{N_p}{k}$ values. The dashed line shows the performance of the union of both residuals when a co-occurrence is used for their representation. The improvement one can obtain by representing residuals using projections is almost 2% and is statistically significant. It appears that combining $2-4$ projection neighborhoods gives the best results over using more neighborhoods (as we are forced to decrease the number of projection vectors). In particular, the lowest detection error of 26.29% was obtained by combining the 1×4, 5×5, and cross neighborhoods.

We ran another experiment of this type for the directional residual using the predictor shown in the third row of Table 1 (it corresponds to the submodel named '3rdspam14hv' in [2]). We were again combining two residuals – one obtained with a horizontal predictor and one with its vertical equivalent. Thus, the co-occurrence representation had the same dimensionality of 338. The results are graphically displayed in Figure 2 (right). Here, the improvement over the co-occurrence representation is even more striking (almost 4%). The best combination of neighborhoods with $E_{\mathrm{OOB}} = 27.31\%$ was obtained for the 1×4 and the stairs neighborhoods.

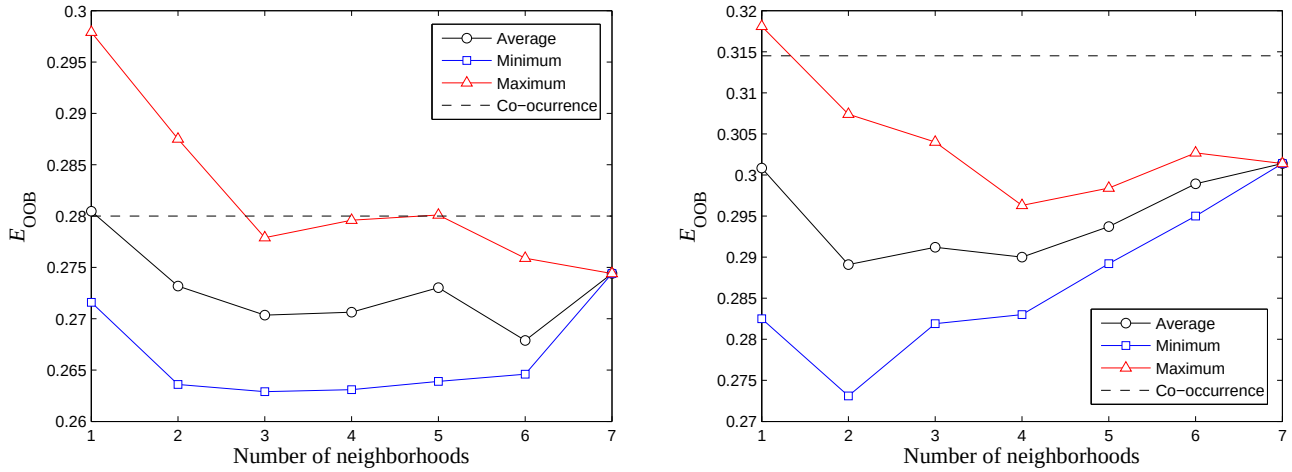

Figure 2. Average, minimum, and maximum detection error E_{OOB} for 1–7 projection neighborhoods (shown in Figure 1) with the number of histogram bins always chosen to obtain feature dimensionality of approximately 338. Left: union of residuals obtained using $\mathbf{K}_3$ and $\mathbf{K}_5$. Right: residuals obtained using the predictor shown in the third row of Table 1 and its vertical version. Steganographic algorithm: WOW at 0.4 bpp.

6. PROJECTION SPATIAL RICH MODEL (PSRM)

In this section, we scale up our approach to a larger number of projection neighborhoods and all residuals from the SRM. As in the previous section, we fixed $T = 4$ and $q = ||\mathbf{v}||_2$ for the set of centroids for the quantizer, which gave us five bins for quantizing residuals symmetrical about zero and 10 bins for residuals of the type min-max.

To increase the diversity of the PSRM, we augmented the seven projections neighborhoods from Figure 1 with four more shapes (see Figure 4). Since it would be computationally infeasible to test all possible combinations of neighborhoods as in the previous section, and also because we want to select a form of PSRM that is heuristically justifiable and not tailored to a specific algorithm, we heuristically assembled the following five neighborhood combinations (the numbers in brackets stand for the number of projection neighborhoods in each combination):

1. All (11). This choice corresponds to maximal diversity across projection neighborhoods.
2. Diverse (6) are formed by 1×4, 1×8, 3×3, 5×5, cross, and thick diagonal. The motivation here was to select a smaller number of diverse neighborhoods.
3. Subsets of 4×4 (6). This set contains the following subsets of the 4×4 neighborhood: 1×4, 2×4, 4×4, thick diagonal, and stairs.
4. Sparse (5). Here, we selected five neighborhood shapes that are "sparse": 1×4, 1×8, diagonal, cross, and 3×3.
5. Diverse (3) contains 1×4, 3×3, and the diagonal.

As in the previous section, for each combination of projection neighborhoods, we gradually increased the number of projection vectors and performed steganalysis of WOW at 0.4 bpp. The OOB error estimate, E_{OOB}, is plotted as a function of the model dimensionality in Figure 3. By observing the figure, we make the following conclusions. First, it clearly pays to diversify across the neighborhood shapes even at the price of not having too many projection vectors. For example, at the dimensionality of the SRMQ1, 12,753, the PSRM version 'All (11)' used only three projection vectors per projection neighborhood and residual, yet it achieved the best overall performance. The 'Diverse (6)' uses approximately half of the projection neighborhoods, which allows doubling the number of projection vectors. Its performance is comparable to 'All (11)'. 'Subsets of 4×4 (6)' and 'Sparse

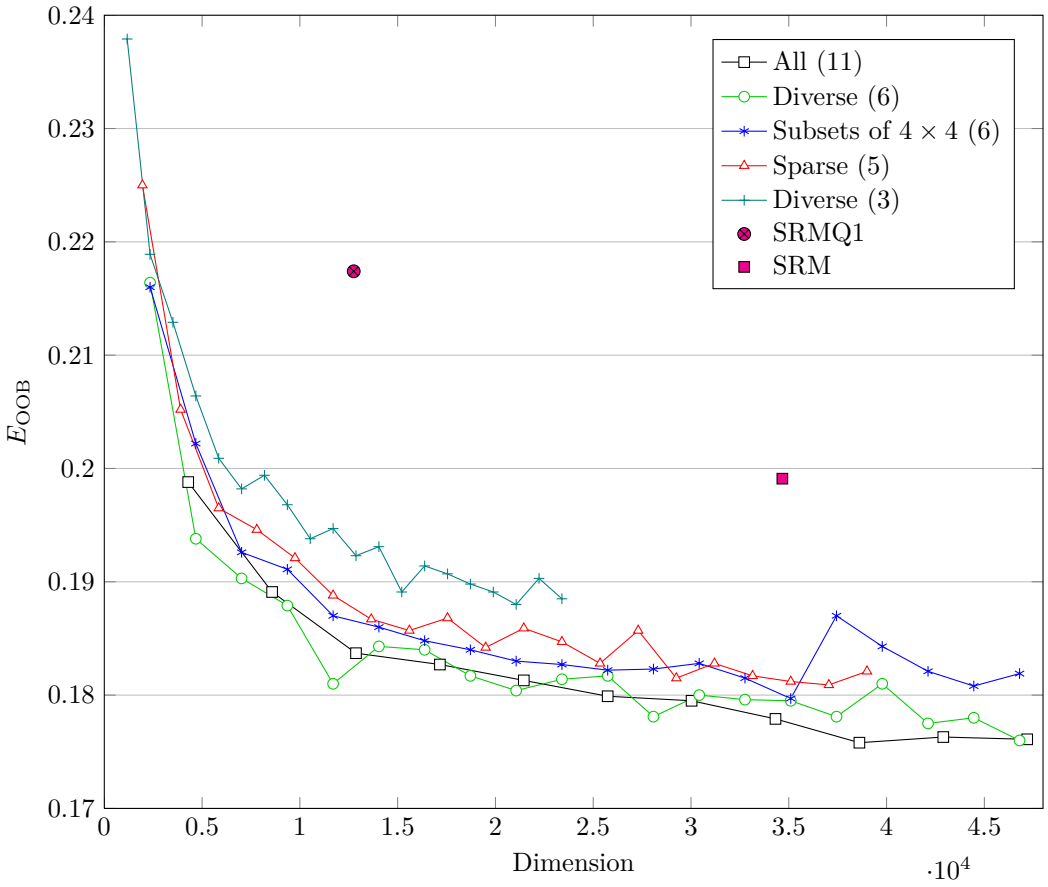

Figure 3. Detection error E_{OOB} as a function of model dimensionality for five different combinations of projection neighborhoods described in the text. The performance of the SRM and its downscaled version SRMQ1 is also plotted for comparison.

(5)' performed slightly worse and the 'Diverse (3)' was the worst combination overall. This indicates that, for a fixed dimensionality, there is a trade-off between the number of projection neighborhoods and the number of projection vectors. Rather than increasing the number of projection vectors, it is better to diversify across the neighborhoods.

All five PSRM models clearly outperformed the SRMQ1 and SRM. At the dimensionality of the SRMQ1, the 'All (11)' version of PSRM detected WOW with an error smaller by almost 4%, which is a very significant improvement. Also, the detection error of all PSRM versions saturates much quicker than for the SRM. (Observe the difference in detection error between the SRMQ1 and SRM and the 'All (11)' version of PSRM for the corresponding dimensionality.) Another way to put this is that the PSRM is able to achieve the same detection error as the SRMQ1 (SRM) with dimensionality that is 5–7 times smaller. This is again a very significant decrease.

The complexity of the model as well as limited time before finalizing this paper prevented the authors from exploring other potentially fruitful alternatives, such as using a coarser projection quantization to further increase the number of projection neighborhoods / projection vectors. The authors plan to investigate this direction in their future work.

7. EXPERIMENTS

In this section, we report the detection errors for four rich models and two embedding algorithms (HUGO and WOW) across relative payloads 0.05, 0.1, 0.2, ..., 0.5 bpp (bits per pixel).

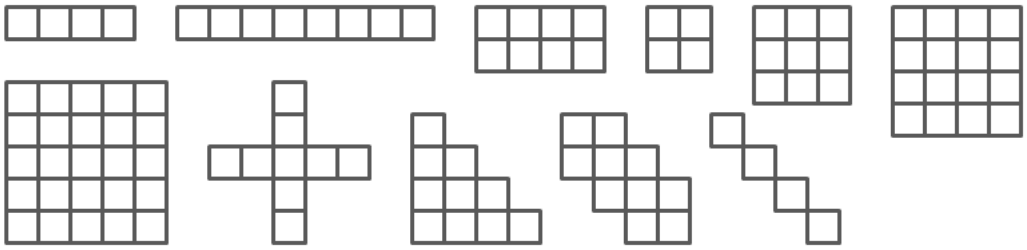

Figure 4. Eleven types of projection neighborhoods $\mathcal{P}$ used in the PSRM: 1×4, 1×8, 2×4, 2×2, 3×3, 4×4, 5×5, cross, stairs, thick diagonal, and diagonal.

Model	SRM	SRMQ1	PSRM8	PSRM3
Dimension	34,671	12,753	34,320	12,870

Table 3. Dimensionalities of four models used to steganalyze HUGO and WOW.

HUGO [11] embeds by minimizing an embedding distortion defined as a weighted norm between the features of the cover and stego image in the SPAM feature space [10]. We used the HUGO embedding simulator [1] with default settings $\gamma = 1$, $\sigma = 1$, and the switch --T with $T = 255$ to remove the weakness reported in [6].

The WOW embedding algorithm [4] uses a bank of directional high-pass filters (8-tap Daubechies wavelet first-level undecimated decomposition) to compute directional residuals to assess the local image content around each pixel along multiple different directions. It obtains the pixel costs in the following manner. First, for every subband and every pixel it computes the so-called embedding suitabilities, which are sums of weighted changes of wavelet coefficients. Then, the suitabilities are aggregated using a reciprocal Hölder norm to obtain costs with the property that if at least one suitability is zero (or very small), the embedding cost is infinite (very large). Since WOW is highly adaptive, it better resists steganalysis using rich models than HUGO.

The four models used for steganalysis are the SRM, SRMQ1, and the 'All (11)' version of PSRM with the number of projection vectors chosen to approximately match those of the SRM and SRMQ1 (3 and 8). The exact dimensionalities appear in Table 3. The abbreviations PSRM3 and PSRM8 were chosen to show the number of projection vectors per neighborhood and residual in the corresponding PSRM.

	WOW				HUGO			
bpp	SRM	PSRM8	SRMQ1	PSRM3	SRM	PSRM8	SRMQ1	PSRM3
0.05	44.72	44.38	45.76	45.12	43.55	42.38	44.75	42.95
0.1	39.58	38.15	41.32	39.07	36.51	35.13	37.55	35.51
0.2	31.17	29.14	33.16	29.68	25.42	24.44	26.76	24.64
0.3	25.36	22.53	26.91	23.08	17.92	16.48	19.30	17.13
0.4	19.91	17.79	21.74	18.37	12.78	11.64	13.37	12.09
0.5	16.36	13.87	17.59	14.26	8.56	8.20	9.43	8.40

Table 4. E_{OOB} for WOW and HUGO using four rich models.

Table 4 shows the improvement in detection error when using the PSRM over SRM. As one may expect, the improvement is larger for WOW than for HUGO because WOW is more selective about placing its embedding changes – it tries to avoid clean edges. Overall, WOW remains markedly more secure than HUGO. The improvement is also larger when both algorithms are forced to embed larger payloads. For payloads larger than 0.2 bpp, PSRM outperforms SRM (at both dimensionalities) by about 3%, which is a quite significant improvement. On the other hand, for HUGO, the detection error decreased by a smaller amount – about $1-2\%$ depending on the payload.

The results of these experiments confirm our reasoning presented in the introduction and in Section 4: The PSRM is a more diverse model that can better detect embedding changes that affect the tails of noise residuals.

8. CONCLUSIONS

In this paper, we proposed an alternative descriptor of noise residuals by projecting them on random directions and using the first-order statistic of projections for the representation. We used the same residuals that were previously proposed for the spatial rich model (SRM) and merely replaced the co-occurrences of quantized residuals with the histograms of their random projections. The resulting model is called the Projection Spatial Rich Model (PSRM).

By experimenting with different projection neighborhoods, we determined that diversification across the neighborhoods significantly boosts detection performance especially for highly adaptive steganographic schemes. This is because the projections can better detect changes that affect mostly the residuals' tails.

The advantage of the PSRM over SRM can be interpreted in two different ways. First, for the same model dimensionality, the PSRM exhibits a smaller detection error than SRM (smaller by about 3% for WOW and $1-2\%$ for HUGO, depending on the payload). Second, for a chosen detection error, the model dimensionality needed to achieve the error is much smaller for PSRM than for SRM. The detection error falls of and saturates with the number of projection vectors much faster than when adding residuals in the SRM. In particular, a given detection accuracy of SRM was reached with a 5–7 times smaller dimension of the PSRM.

Finally, the new model offers greater design flexibility (dimensionality versus accuracy trade-off) as the steganalyst can easily fine-tune the number of projection neighborhoods and the projection vectors to an almost any desired number.

In the future, the authors contemplate exploring further diversification of the PSRM across more neighborhoods by utilizing a coarser projection quantization. Also, the residual projections can be viewed in a more general light as simply outputs of a very large kernel – they are essentially linear combinations of shifted kernels. From this point of view, the PSRM actually collects *first-order statistics* of very complex residuals with very large supports. It is certainly worth investigating whether better performance might be obtained by designing the large kernels directly rather than via linear combinations of outputs of shifted kernels.

9. 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.

REFERENCES

1. P. Bas, T. Filler, and T. Pevný. Break our steganographic system – the ins and outs of organizing BOSS. In T. Filler, T. Pevný, A. Ker, and S. Craver, editors, *Information Hiding, 13th International Conference*, volume 6958 of Lecture Notes in Computer Science, pages 59–70, Prague, Czech Republic, May 18–20, 2011.
2. 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.
3. G. Gül and F. Kurugollu. A new methodology in steganalysis : Breaking highly undetactable steganograpy (HUGO). In T. Filler, T. Pevný, A. Ker, and S. Craver, editors, *Information Hiding, 13th International Conference*, Lecture Notes in Computer Science, pages 71–84, Prague, Czech Republic, May 18–20, 2011.
4. 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.
5. A. D. Ker and R. Böhme. Revisiting weighted stego-image steganalysis. In E. J. Delp, P. W. Wong, J. Dittmann, and N. D. Memon, editors, *Proceedings SPIE, Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, volume 6819, pages 5 1–5 17, San Jose, CA, January 27–31, 2008.

6. J. Kodovský, J. Fridrich, and V. Holub. On dangers of overtraining steganography to incomplete cover model. In J. Dittmann, S. Craver, and C. Heitzenrater, editors, *Proceedings of the 13th ACM Multimedia & Security Workshop*, pages 69–76, Niagara Falls, NY, September 29–30, 2011.
7. 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.
8. I. Lubenko and A. D. Ker. Going from small to large data sets in steganalysis. In A. Alattar, N. D. Memon, and E. J. Delp, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics of Multimedia XIV*, volume 8303, pages 0M 1–10, San Francisco, CA, January 23–26, 2012.
9. T. Pevný. Co-occurrence steganalysis in high dimension. In A. Alattar, N. D. Memon, and E. J. Delp, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics of Multimedia XIV*, volume 8303, pages 0B 1–13, San Francisco, CA, January 23–26, 2012.
10. T. Pevný, P. Bas, and J. Fridrich. Steganalysis by subtractive pixel adjacency matrix. *IEEE Transactions on Information Forensics and Security*, 5(2):215–224, June 2010.
11. T. Pevný, T. Filler, and P. Bas. Using high-dimensional image models to perform highly undetectable steganography. In R. Böhme and R. Safavi-Naini, editors, *Information Hiding, 12th International Conference*, volume 6387 of Lecture Notes in Computer Science, pages 161–177, Calgary, Canada, June 28–30, 2010. Springer-Verlag, New York.
12. D. Upham. Steganographic algorithm JSteg. Software available at `http://zooid.org/~paul/crypto/jsteg`.