Single Band Statistics and Steganalysis Performance

Pham Hai Dang Le, Matthias O. Franz
Institute for Optical Systems (IOS)
HTWG Konstanz University of Applied Sciences
Brauneggerstraße 55, 78462 Konstanz, Germany
$\{$*dangle, mfranz*$\}$*@htwg-konstanz.de*

***Abstract*—Most universal steganalysis techniques use an image model to reconstruct an estimate of the original, unmanipulated cover from the input. Differences between reconstructed and input images are an indication of a steganographic manipulation. In this paper, we analyze the relation between the modeling error of the image model and detection performance in the wavelet domain. Based on the modeling error we define a measure of separability which is highly correlated with detection performance. We find that in uncompressed images only fine scales play a role in steganalysis, whereas separability is spread out over all scales in JPEG images.**

***Keywords*-steganography, steganalysis, universal steganalysis**

I. Introduction

With the rise of digital imagery on the internet, digital images increasingly act as camouflage for steganography. The simplest scenario for detecting a secret message in a harmlessly looking image is the so-called *cover-stego-attack*. Here, the attacker knows both images, the clean image (cover) and the stego image (images with embedded message). The presence of a steganographic manipulation can be detected by simply noticing a difference between cover and stego image. In most realistic scenarios, cover-stego-attacks are not possible because the cover image is not available to the attacker. Steganalysis has to be performed on the basis of the input image alone (*stego-only-attack*). If additionally the embedding algorithm is unknown, we have the scenario of blind or universal steganalysis.

Most universal steganalysis algorithm attempt to simulate the cover-stego scenario by reconstructing the original cover image from the input image at hand. This is possible to a certain extent since there exist considerable dependencies between image elements which can be captured by a suitable image model. Because the embedded message typically is independent of the cover image content, it is to be expected that these dependencies are perturbed by the embedding process. As a consequence, the prediction error of the image model should be smaller in a clean, unmanipulated image than in a stego image. In cases with a large difference between the prediction errors of a stego image and its clean version it is quite likely that the steganographic manipulation can be discovered.

In the present study, we investigate whether such a relation between the prediction error difference and steganalysis performance can be found in a realistic universal steganalysis scenario. We characterize the prediction error in terms of the *explained variance* of the image model in the image. Our hypothesis is that the difference in the explained variances between a stego image and its clean version (referred to as *separability*) correlates with detection rate. As in most universal steganalysis approaches, we do our analysis in the wavelet domain. This allows us to investigate the contribution of single wavelet subbands to overall separability and steganalysis performance.

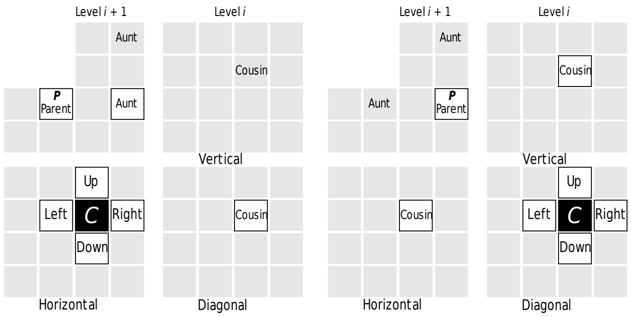
II. Steganalysis Model

Lyu and Farid's steganalysis technique [1] can be divided into three components: (1) the image subband transform, (2) the modeling of the dependency structure of the transform coefficients, and (3) the classification.

In the first step, Lyu and Farid's algorithm transforms an input image from its pixel representation into its wavelet representation using QMF (quadrature mirror filter) wavelets of support size 9 [2]. We used a 3-level wavelet pyramid decomposition which results in 30 subbands for each image (3 levels with 3 orientation subbands and one lowband and 3 color channels). Each subband is indexed according to the scheme depicted in Fig. 3.

The modeling step takes place in the wavelet domain. In order to model the dependency of a wavelet coefficient from its neighborhood, we have to define a neighborhood structure which is shown in Fig. 1. Due to only including the neighboring coefficients from closest orientations on the same scale (hence including horizontal and vertical coefficients for predicting the diagonal subband, but only diagonal coefficients for both the horizontal and vertical subbands), and correspondingly only one (diagonal) or two neighbors (horizontal and vertical) from the coarser scales, neighborhoods in the wavelet representation contain 7 coefficients from the same color channel as well as the corresponding central coefficients from the other color channels (not shown in Fig. 1).

The predictions are computed with linear regression applied to each subband separately, i. e., the magnitude of the central coefficient is obtained as a weighted sum of the magnitudes of its neighboring coefficients greater than a



(a) Horizontal wavelet band (b) Diagonal wavelet band

Figure 1: Neighborhood structure for image modeling: The central coefficient to be predicted is C, the neighbors are highlighted in white (color neighbors not included).

given threshold: It has been shown empirically that only the magnitudes of coefficients are correlated, and the correlation decreases for smaller magnitudes [3]. The weight sets over all subbands thus constitute the image model. In their original approach, Lyu and Farid used standard least-squares regression for this purpose. In our implementation, we use Gaussian Process (GP) regression [4] instead after normalizing all subband coefficients to the interval $[0, 1]$. This approach leads to slightly more robust, but essentially comparable results for the purpose of this study. The GP regression needs an additional model selection step for estimating the noise content in the image, we use Geisser's surrogate predictive probability [4]. It is computed on a subset of the coefficients: The finest scales are subsampled by a factor of 5 and the coarser by a factor of 3, each in both directions. Details on this regression technique can be found in [4]. Each estimator is trained and used for prediction on the same subband. Thus, the training and test set coincide for this application. From the predicted coefficients $\hat{S}$, small coefficients with amplitude below a threshold of $t = 1/255$ are set to zero. For reconstructing complete images, the algebraic signs are transferred from the original to the predicted subband coefficients. The residual r is computed by taking the logarithm of the coefficients of the input image transform S and the predicted coefficients $\hat{S}$ and subsequently subtracting them, hence $r = \log S - \log \hat{S}$.

Next, the four lowest statistical moments, i.e. mean, standard deviation, skewness, and kurtosis, of the subband coefficients (called *marginal statistics* in [1]) and of the subband residuals (called *error statistics*) are computed, again for each color and subband separately. Finally, all these independently normalized statistics serve as feature inputs for a support vector machine [5]. In this study, we use $s = 3$ pyramid levels which results in a 120-dimensional feature vector. The final classification was computed with a 1-norm soft margin non-linear C-SVM using a Gaussian kernel on 2000 images (the training data, 1000 clean and 1000 stego images), and then tested with a set of 1244 examples (622 clean and 622 stego images). We randomly divided the entire set into training and test sets and averaged over 100 splittings, which enabled us to estimate the error of the detection rate on the test set. The choice of the parameter C of the SVM and the width σ of the Gaussian kernel was based on a new paired cross-validation procedure described elsewhere (in preparation). The SVM is tunable in order to adapt the rate of false alarms and the detection rate. We report steganalysis performance at two points of the ROC curve: (1) detection rate at the point of minimum overall error; (2) detection rate at a false alarm rate of 0.01. We prefer to give both values since (1) has a much smaller variance than (2) which is the standard in the literature. Thus, comparisons in (1) are more meaningful (Tables I & II).

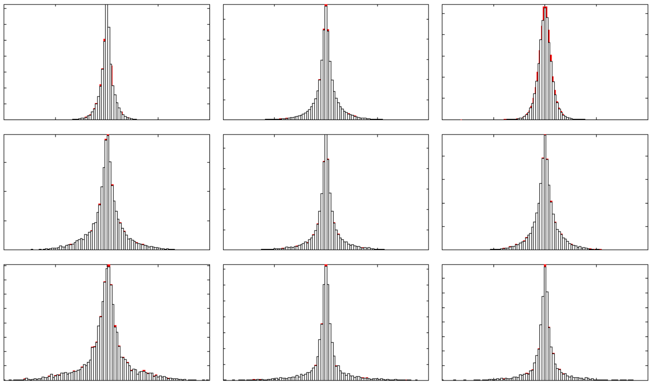


Figure 2: Log prediction error for each scale and orientation of a wavelet pyramid for the green channel of a clean image and its stego version by using OutGuess [6]. Differences are marked with (🟥). From left to right: the horizontal, vertical, and diagonal orientation. From top to bottom: the fine scales on the wavelet level 1 to the coarse scales on the wavelet level 3.

III. Experiments

The starting point of our analysis is the fact that the prediction error of clean and stego images must show significant differences in its wavelet coefficient statistics to be detectable by the subsequent classification stage (Fig. 2). Intuitively, we expect that steganalysis performance should correlate with the degree of the differences between their respective prediction errors. In other words, we expect a higher prediction quality in clean images than in stego images, and the difference between these prediction qualities should correlate with steganalysis performance. To determine the quality of the prediction in a given subband S of the wavelet decomposition, we use the *explained variance* $\mathcal{V}_{\text{expl}}^{S}$ which is computed from the original wavelet coefficients $X^S = \{X(i,j)\}_{i,j \in S}$ in S and their values $\hat{X}^S = \{\hat{X}(i,j)\}_{i,j \in S}$

predicted from the linear image model as

$$\mathcal{V}^{S}_{\mathrm{expl}} \equiv \frac{\mathrm{Var}\left(X^S\right) - \mathcal{V}_{\mathrm{err}}\left(X^S, \hat{X}^S\right)}{\mathrm{Var}\left(X^S\right)}$$

with the error variance $\mathcal{V}_{\mathrm{err}}\left(X^S, \hat{X}^S\right) \equiv \frac{1}{|S|}\sum_{i,j\in S}\left(X^S(i,j) - \hat{X}^S(i,j)\right)^2$. We determine the explained variances $\mathcal{V}^{S,c}_{\mathrm{expl}}$ for S in the clean image and $\mathcal{V}^{S,s}_{\mathrm{expl}}$ in the corresponding stego image, from which we derive a (relatively crude) measure for the difference in the prediction qualities between both images as

$$\Delta\mathcal{V}^{S}_{\mathrm{expl}} = \mathcal{V}^{S,c}_{\mathrm{expl}} - \mathcal{V}^{S,s}_{\mathrm{expl}}.$$

We will refer to this measure as the *separability* of the subband S in the wavelet decomposition. A higher separability (either in a single subband or over all subbands) should generally lead to a higher steganalysis performance, although the reverse is not necessarily true: two pairs of clean and stego images can have similar separabilities, but a different steganalysis performance because our separability measure $\Delta\mathcal{V}_{\mathrm{expl}}$ does not capture all possible differences between the two prediction error distributions.

Both separabilities and steganalysis performance were determined on a database containing (1622) (640 × 480) never compressed PNG color images provided by the German Federal Office for Information Security. From these images, steganograms were generated using three standard algorithms: (1) Straight up ±1 embedding [7]; (2) Straight up LSB replacement; (3) Straight up ternary embedding [8]. The embedding ratios tested were 75, 50, 40, 30 and 25%. Furthermore, we converted the 1622 PNG images into JPEG (with a quality level of 60 to 100) and applied the same three embedding methods as above in the JPEG domain using F5 [9]. Here, the tested embedding ratios were 90, 70, 50, 30 and 10%. In all cases, only the central 256 × 256 region of each image was selected for analysis.

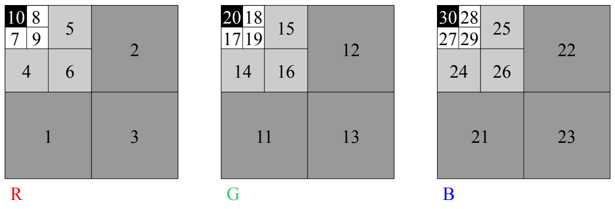


Figure 3: Subbands of the wavelet decomposition in the Fourier domain. Subbands 1-10 are red (R), 11-20 green (G) and 21-30 blue (B), the finest scales have 1, 2, 3 as last digit, medium scales 4, 5, 6, and large scales 7, 8, 9. The lowbands have indices 10, 20 and 30.

In the first experiment, we compared separabilities between subbands to find out whether only a fixed number of subbands shows a high separability, consistently over all images, or whether the most distinctive subband varies from image to image. For this purpose, we identified the subband with the highest separability value for each image and generated a histogram over all PNG and JPEG images separately (Fig. 4).

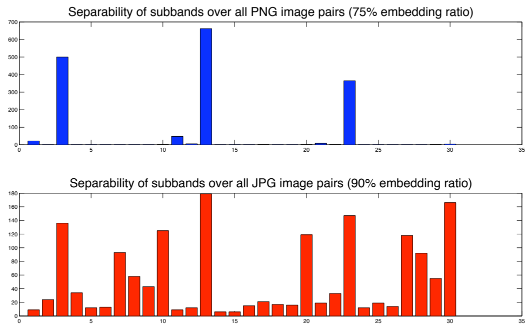


Figure 4: $\Delta\mathcal{V}_{\mathrm{expl}}$ over all PNG (straight-up ±1 embedding at 75% embedding ratio 🟦) and JPEG image pairs (F5-ones with ±1 embedding and 90% embedding ratio 🟥). The abscissa is the index of the wavelet subband shown in Fig. 3. The ordinate gives the number of image pairs in which the corresponding subband has the highest separability.

Fig. 4 shows an interesting difference between both image classes: the subbands of PNG images with highest separability are consistently at the finest scales, with a strong dominance of the diagonal orientation subbands (3, 13, 23), whereas the most distinctive subbands of JPEG images do not show a consistent pattern and are spread over all wavelet bands.

In the second experiment, we investigated whether high separability values lead indeed to a higher steganalysis performance. In addition, we tested whether this correlation is more pronounced with the separability value of the most distinctive subband, or with the average separability over all subbands. The results for the 6 embedding algorithms tested and various embedding ratios are given in Tables I and II. To enhance readability, the highest table entry per image class (PNG or JPEG) and embedding ratio is highlighted with a dark gray background, and the second best with a light gray background.

Table I demonstrates that in the case of PNG images, the separability of the most distinctive subband is a better predictor of steganalysis performance than the average separability over all subbands. This is consistent with our findings in the first experiment where we found that mainly the diagonal fine detail bands contribute most to the overall separability. In the case of JPEG images (Table II), this difference is less clear-cut, as was to be expected from the spread-out separabilities in our first experiment: both separability of the most distinctive subband and average

Table I: Separability of subbands ($\Delta\mathcal{V}_{\text{expl}}$) over all PNG images in comparison to different embedding methods and embedding ratio: P_D (in %) indicates the detection performance, P_{FA} (in %) the false alarm at the point of minimum overall error, $P_D^{0.01}$ (in %) the detection performance at false alarm rate of 1 %, and E_R the embedding rate. Each image pair (clean and stego images) determines max, pos, and **mean**. **mean** (in %) is the mean value of $\Delta\mathcal{V}_{\text{expl}}$ over all 30 bands. max (in %) indicates the highest $\Delta\mathcal{V}_{\text{expl}}$ value over all bands and pos labels the position of max. All three values are obtained by averaging over all image pairs (1622 clean & 1622 stego images).

	Straight-up: plus-minus-1						Straight-up: replacement						Straight-up: ternary					
E_R	max	pos	**mean**	P_D	P_{FA}	$P_D^{0.01}$	max	pos	**mean**	P_D	P_{FA}	$P_D^{0.01}$	max	pos	**mean**	P_D	P_{FA}	$P_D^{0.01}$
75%	3.2460	23	0.6907	99.44	0.4823	99.66	2.7000	23	0.7172	98.81	0.9646	98.58	3.0190	23	0.8337	99.25	0.9646	99.46
50%	1.7880	23	0.3598	97.20	0.9646	90.82	1.4600	23	0.3997	95.21	0.9646	75.59	1.6550	23	0.4673	96.82	0.9646	87.26
40%	1.1050	23	0.2385	93.88	0.9646	63.75	0.9140	23	0.2673	91.00	0.9646	37.19	1.0490	23	0.3130	92.77	0.9646	52.66
30%	0.3950	23	0.0815	82.44	0.9646	7.74	0.8700	10	0.1286	68.72	0.9646	7.30	0.3790	23	0.1345	77.25	0.9646	8.27
25%	0.1450	10	0.0266	60.92	0.8039	2.22	0.3400	10	0.0409	55.53	0.9646	1.96	0.3310	10	0.0421	58.29	0.9646	2.74

Table II: Just as Table I, but for JPEG images: effective change ratios are slightly different.

	F5-ones: plus-minus-1						F5-ones: replacement						F5-ones: ternary					
E_R	max	pos	**mean**	P_D	P_{FA}	$P_D^{0.01}$	max	pos	**mean**	P_D	P_{FA}	$P_D^{0.01}$	max	pos	**mean**	P_D	P_{FA}	$P_D^{0.01}$
90%	3.1880	10	0.3978	95.58	0.9646	88.23	5.2710	30	0.9272	96.32	0.9646	89.61	4.0240	13	0.8687	96.05	0.9646	87.65
70%	2.1960	10	0.3240	95.14	0.9646	82.73	3.0310	23	0.6303	95.62	0.9646	86.05	3.0840	3	0.7025	95.41	0.9646	84.38
50%	1.2690	10	0.2055	95.70	0.9646	86.95	2.3070	23	0.5064	95.67	0.9646	86.56	3.5920	30	0.6032	95.24	0.9646	84.68
30%	0.8350	30	0.2009	95.00	0.9646	85.06	3.9680	20	0.5820	95.50	0.9646	86.45	2.9410	30	0.4573	95.19	0.9646	83.91
10%	0.5330	13	0.1238	94.76	0.9646	82.30	1.5730	30	0.1705	95.27	0.9646	83.72	0.9400	30	0.2112	94.87	0.9646	82.76

separability are strongly correlated among each other, and similarly correlated to steganalysis performance.

IV. Conclusion

Our investigation on the relation between separability and detection performance showed that both quantities are indeed highly correlated. In uncompressed image formats such as PNG or BMP, we found for the tested embedding methods that the relevant wavelet bands for steganalysis performance are located at fine scales. In contrast, images with compression schemes based on the discrete cosine transform (DCT) such as JPEG images did not show a concentration of separability in the fine subbands. Here, separability was spread out across different scales. This can be attributed to the DCT which spreads the bits manipulated by the embedded message over the whole image in the spatial domain.

When building specialized steganalyzers for uncompressed images, our results imply that it is sufficient to investigate the fine scales of the wavelet transform. In a more general setting, our separability metric could be used to construct an adaptive image basis for a given type of embedding during training. In wavelet packet or local cosine bases, for instance, the decision on whether a certain subband should be further decomposed can be based on the separability of the subband.

References

[1] S. Lyu and H. Farid, "Steganalysis using higher-order image statistics," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 1, pp. 111–119, 2006.

[2] E. P. Simoncelli and E. H. Adelson, "Subband transforms," in *Subband Image Coding*, J. W. Woods, Ed. Norwell, MA, USA: Kluwer Academic Publishers, 1990.

[3] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Transactions on Image Processing*, vol. 8, no. 12, pp. 1688–1701, December 1999.

[4] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, T. Dietterich, Ed. Cambridge, MA, USA: MIT Press, January 2006.

[5] B. Schölkopf and A. J. Smola, *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[6] N. Provos and N. Provos, "Defending against statistical steganalysis," in *10th USENIX Security Symposium*, 2001, pp. 323–335.

[7] T. Holotyak, J. J. Fridrich, and D. Soukal, "Stochastic approach to secret message length estimation in ±k embedding steganography," in *Security, Steganography, and Watermarking of Multimedia Contents*, 2005, pp. 673–684.

[8] J. J. Fridrich, P. Lisonek, and D. Soukal, "On steganographic embedding efficiency," in *Information Hiding*, 2006, pp. 282–296.

[9] A. Westfeld, "F5 - a steganographic algorithm: High capacity despite better steganalysis," in *4th International Workshop on Information Hiding*. Springer-Verlag, 2001, pp. 289–302.