https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-535

Algorithm Mismatch in Spatial Steganalysis

***Stephanie Reinders*[+]*, Li Lin*[+]*, Yong Guan*[†]*, Min Wu*[⋆]*, Jennifer Newman*[+]**
[+] ***Department of Mathematics,*** [†] ***Department of ECPE, Iowa State University, Ames, Iowa, USA,***
[⋆] ***Department of ECE, University of Maryland, College Park, MD, USA***

Abstract

The number and availability of stegonographic embedding algorithms continues to grow. Many traditional blind steganalysis frameworks require training examples from every embedding algorithm, but collecting, storing and processing representative examples of each algorithm can quickly become untenable. Our motivation for this paper is to create a straight-forward, non-data-intensive framework for blind steganalysis that only requires examples of cover images and a single embedding algorithm for training. Our blind steganalysis framework addresses the case of algorithm mismatch, where a classifier is trained on one algorithm and tested on another, with four spatial embedding algorithms: LSB matching, MiPOD, S-UNIWARD and WOW.

We use RAW image data from the BOSSbase database and and data collected from six iPhone devices. Ensemble Classifiers with Spatial Rich Model features are trained on a single embedding algorithm and tested on each of the four algorithms. Classifiers trained on MiPOD, S-UNIWARD and WOW data achieve decent error rates when testing on all four algorithms. Most notably, an Ensemble Classifier with an adjusted decision threshold trained on LSB matching data achieves decent detection results on MiPOD, S-UNIWARD and WOW data.

Introduction

Steganography is the practice of hiding a message, called a *payload*, in an innocent looking object, called a *cover*. The goal of steganography is to hide the payload in such a way that a casual observer will be unaware a secret message is being sent. Digital image stegaphography hides payloads of text, images, or other data in digitial images. Image steganalysis is the analysis of an image for steganography content, and is typically accomplished with machine learning or signature-based detection.

In *targeted steganalysis* the steganalyst assumes knowledge of the particular embedding algorithm used. In contrast, *blind steganalysis* or *blind detection* assumes no, or little, knowledge of the embedding algorithm. "The goal of blind steganalysis is to detect any steganographic method irrespective of its embedding mechanism." [1] The case of *algorithm mismatch*, where a classifier is used to detect embedding algorithms not used in training, is crucial to performing blind detection in the real world because a steganalyst is unlikely to know which embedding algorithm was used. The blind detection framework that we present in this paper focuses on algorithm mismatch.

Kong, Feng, Li and Guo address the algorithm mismatch problem on JPEG image data through domain adaptation techniques [2]. They show that using their iterative, non-linear feature transformation a classifier trained on covers and a single embedding algorithm achieves decent detection rates on unseen embedding algorithms.

Pevný and Fridrich construct several classifiers capable of blind detection by showing the classifiers examples of as many algorithms as possible [3, 4, 5, 6]. However, as the number of stego algorithms increases, the steganalyst will find it increasingly challenging to collect, store, and process examples of every possible algorithm.

We address the algorithm mismatch problem in the spatial domain with four embedding algorithms: LSB matching, MiPOD [7], S-UNIWARD [8] and WOW [9]. We devise a blind classification framework, consisting of a single binary classifier, that does not require feature transformation and only requires training examples from covers and a single embedding algorithm. We perform *algorithm mismatch experiments* where we train an Ensemble Classifier [10] with Spatial Rich Model [11] features on one of the four embedding algorithms and test on all four algorithms. Our results show that MiPOD, S-UNIWARD, and WOW trained classifiers achieve decent detection rates when testing all four embedding algorithms. Furthermore, an LSB matching trained Ensemble Classifier with an adjusted decision threshold is able to achieve decent detection rates when testing MiPOD, S-UNIWARD and WOW image data.

The Prior Art section contains a more in-depth summary of previous work in the area of blind detection. We describe the datasets used and the structure of our algorithm mismatch experiments in the Methods section. The Results section details the results of our experiments. We summarize our findings and explain potential avenues for future research in the Conclusions and Future Work section.

Prior Art

In this section we summarize the *multi-classifier*, *one-class classifier*, *one-against-all classifier*, and *domain adaptation* approaches to blind detection and explain how our approach differs from them.

The term blind steganalysis or blind detection is used in two related but different ways in the literature. In some cases, the term is used to refer to steganalysis frameworks that aren't constructed for a specific embedding algorithm. As an example, the quantitative steganalyzer introduced by Pevný, Fridrich and Ker [12] is a blind framework in the sense that it isn't built for any specific algorithm. Many feature sets are not specialized to a specific embedding algorithm, but are suitable for many algorithms. Steganalysis frameworks that use such feature sets are occasionally referred to as blind steganalysis frameworks in the literature [13, 14]. The term blind steganalysis is also used to refer to the act of detecting stego images when the embedding algorithm is unknown. We use this meaning of the term in this paper.

The blind steganalysis framework we present in this paper addresses a specific blind steganalysis problem, the algorithm

mismatch problem, where classification is done on unseen embedding algorithms. For the remainder of this section, we focus on prior art that addresses the algorithm mismatch problem.

One approach to blind detection trains a group of binary classifiers, called a *multi-classifier*, on a wide variety of embedding algorithms [3, 4, 5, 6]. More specificially, a binary classifier is trained for each possible pair of n classes, and the multi-classifier is the collection of these $\binom{n}{2}$ binary classifiers. The multi-classifier assigns a test image to one of n classes by asking each binary classifier to vote on the class of the test image. The class with the most votes is chosen as the winner. Cover is one class and each stego embedding algorithm is its own class. While, Pevný and Fridrich show that their multi-classifier achieves good detection results, even on an unseen algorithm, this approach is data-intensive as it requires training examples of $n-1$ stego algorithms. Our approach differs in that it only needs training examples from a single algorithm.

Another approach uses a *one-against-all classifier*, a single binary classifier trained on cover images and a wide variety of stego algorithms [1]. The one-against-all classifier and the multi-classifier both operate on the theory that if the classifier is shown a sufficient sample of stego algorithms to effectively represent the stego space, the classifier will be able to detect unseen stego algorithms. A third approach trains a *one-class classifier* only on cover images and uses anomaly detection to identify stego images [1]. This approach attempts to sufficiently represent the cover space in such a manner that the classifier can recognize stegos of any algorithm as not belonging to the cover space. Pevný and and Fridrich found that the one-against-all classifier could perform poorly on unseen algorithms and the one-class classifier had lower overall accuracy compared with the multi-classifier. Unlike the one-against-all classifier that is trained on many stego algorithms and the one-class classifier that isn't trained on any stego algorithms, our proposed classifier is trained on covers and a single stego algorithm.

Unsupervised learning has been used in several works for blind detection [15, 16] and don't require training or knowledge of the embedding algorithm.

Kong, Feng, Li and Guo apply domain adaptation to the algorithm mismatch problem on JPEG image data [2]. Algorithm-mismatch, cover-source mismatch, or other factors could cause the training features in the source domain and the test features in the target domain to have different distributions. If a classifier is trained on features with one distribution, it might not perform well on test features of a different distribution. Kong, Feng, Li and Guo address the sitution where the steganalyst has access to examples of labeled covers and stegos in the source domain and unlabeled covers and stegos in the target domain. They apply a feature transformation as a two-step process to make the features of the training set similar to the features of the test set in the target domain. First, the features of the source domain are transformed so that the joint expectations and the standard deviations of the source domain and the target domain are the same. Then as a second step, the source domain features are further transformed to minimize the maximum mean discrepancy between the marginal and conditional distributions. They show the success of their method by applying the feature transformation to a test set of covers and stegos from a single embedding algorithm, then training a classifier on the transformed features, and testing the classifier on covers and stegos from a different embedding algorithm. They compare the classification results with a classifier trained on covers and a single embedding algorithm and tested on covers and that same embedding algorithm. While Kong, Feng, Li and Guo focus on the DCT domain, we address the algorithm mismatch problem in the spatial domain. Our approach does not require feature transformation: We are able to achieve decent detection results without changing the source domain or target domain distributions.

Methods

In this section we explain the datasets and methodology used in our experiments.

Image Datasets

We choose to use two datasets: the BOSSbase database [17] and iPhone image data that was collected as part of a forensic database project [18]. We choose the former because it is a well-known and benchmarked dataset of images from digital still cameras. The latter we choose because it consists of images from mobile devices. As increasingly more images "in-the-wild" originate from cell phone cameras, it is it is important to collect data from these sources [19, 20, 21, 22, 23]. While the iPhone dataset used in this work is not in the query part of the StegoAppDB database, the iphone dataset is available for download after March 1, 2019, by visiting the StegoAppDB homepage [18] and clicking on the link for "Algorithm Mismatch Dataset."

The BOSSbase dataset contains 10,000 RAW images from seven digital still cameras. We convert the RAW images to TIFF images in Photoshop. Then we center-crop 512x512 subimages, convert them to 256-bit grayscale and save them in the PNG format, all in Matlab. These 512x512 grayscale images serve as cover images. The BOSSbase images are a mixture of auto-exposure and manual exposure images.

We use 1,927 TIFF auto-exposure images collected on two iPhone 6s, two iPhone 6s Plus, and two iPhone 7 devices using a camera app. We convert each TIFF image to 256-bit grayscale, crop it into five 512x512 disjoint subimages, and save in the PNG file format, totaling 9,635 cover images.

We create stego images from both datasets in the same manner. From each cover image we create stego images using four embedding algorithms, LSB matching, MiPOD, S-UNIWARD and WOW, and three embedding rates, 10%, 20%, and 40%, for the BOSSbase dataset. Due to time constraints we use one embedding rate, 10%, for the iPhone dataset.

Methodology of Algorithm Mismatch Experiments

In an *algorithm mismatch experiment* we train an Ensemble Classifier [10] with Spatial Rich Model features [11] on covers and a single embedding algorithm - LSB matching, MiPOD, S-UNIWARD, or WOW - and embedding rate. We choose this classifier and feature set because they are both well-known and widely used in the steganalysis community. We test the trained classifier on covers and all four embedding algorithms with the same embedding rate. The detection error from the *algorithm mismatch case*, where the training and testing algorithms are different, is compared to the the detection error from the *best-case* classifier, where the training and testing algorithms are the same.

We perform algorithm mismatch experiments on the full set of 10,000 images from all seven BOSSbase devices and the full

set of 9,635 auto-exposure images from six iPhone devices. We randomly select a training set of 5,000 cover images and corresponding stego images from a single embedding algorithm and embedding rate. The test set is comprised of the remaining cover images and corresponding stegos from the same embedding rate used in training and all four embedding algorithms. The results are averaged over five repetitions.

Previous work has shown that detection accuracy can be improved by training and testing on a single device [22], so we also perform algorithm mismatch experiments on individual devices to determine if algorithm mismatch experiments see similar improvement. Because we have a smaller number of available images from any single device, for individual device experiments we perform five repetitions of ten-fold cross-validation on randomly selected samples of 700 cover images and corresponding stego images from a single embedding algorithm and embedding rate. For each fold during the cross-validation process, 70 cover images and the corresponding stegos from all four algorithms are set aside for testing. The training set consists of 630 covers and corresponding stegos from a single embedding algorithm. The 630 stegos from each of the other three algorithms are neither used for training nor testing.

We calculate the detection error rate in Equation 1 as the average of the false alarm rate P_{FA} and the missed detection rate P_{MD} for a single algorithm. For example, for the classifier trained on covers and MiPOD, the LSB detection error rate is the average of the false alarm rate and the rate of LSB stegos classified as cover.

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

In our initial experiments we trained Ensemble Classifiers on LSB matching, MiPOD, S-UNIWARD, and WOW image data, training one classifier for each algorithm. Each trained classifier was used to test each of the four algorithms. The detection errors of these four classifiers for detecting MiPOD using BOSSbase 40% embedding rate image data are shown in Figure 1. We see that the S-UNIWARD and WOW classifiers achieve detection errors close to the best-case classifier, the MiPOD classifier. However, the LSB trained classifier results in error rates close to random guessing when testing MiPOD images. The same experiment on different embedding rates, as well as on the iPhone dataset produced similar results.

LSB matching is the simplest and least complex of the four embedding algorithms. This motivates us to try to improve the LSB trained classifier to achieve better detection error on MiPOD, S-UNIWARD, and WOW. With such an improvement, a steganalyst could quickly produce and use LSB data for training classifiers, and not need to produce any other stego images for training. We discovered that adjusting the value of the decision threshold within the Ensemble Classifier achieves this goal. We call an Ensemble Classifier with an adjusted decision threshold trained on LSB matching data an *LSB Adjusted* classifier. Figure 2 shows that the LSB Adjusted classifier achieves much lower detection error than the LSB classifier. We discuss the LSB Adjusted classifier in greater detail in the next subsection.

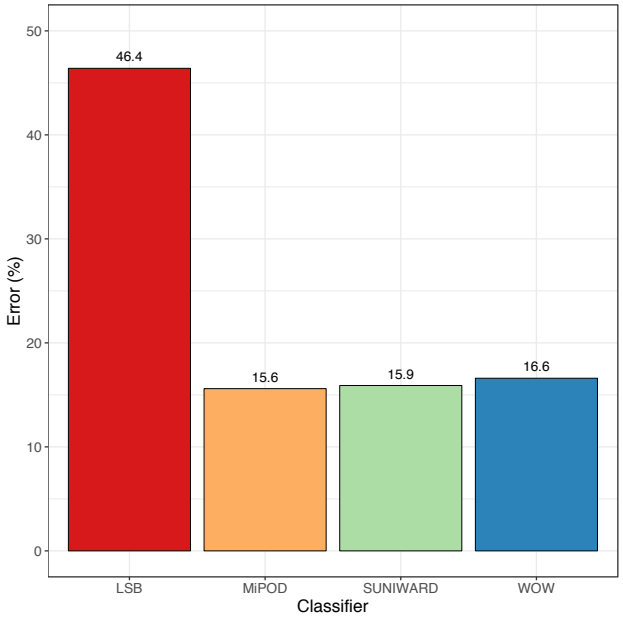


Figure 1. *Average error rate by classifier on MiPOD test set at 40% embedding on images from all BOSSbase devices (training size=5,000)*

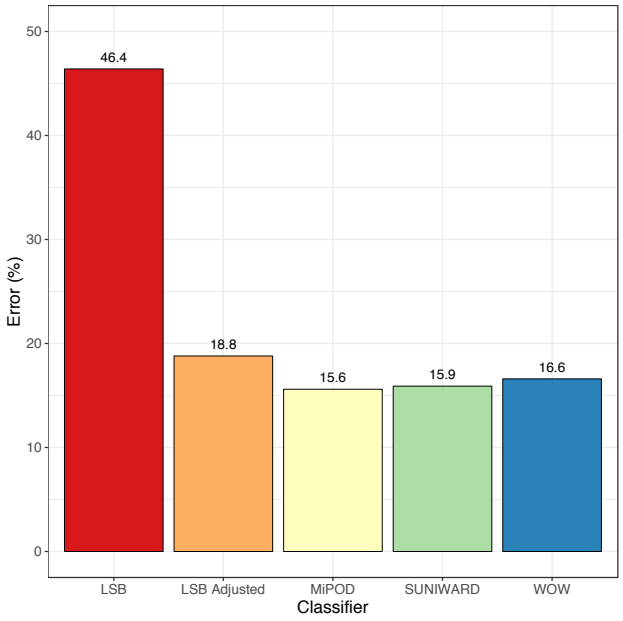


Figure 2. *Average error rate by classifier on MiPOD test set at 40% embedding on images from all BOSSbase devices (training size=5,000)*

Adjusting the Decision Threshold to Improve Classification Results

We found that with slight modifications an Ensemble Classifier can be trained solely on covers and LSB matching data and achieve a decent detection error rate when testing MiPOD, S-UNIWARD, and WOW data.

In order to explain our modifications to the Ensemble Classifier, we first give a brief overview of the classifier's pertinent parts. For a more in depth description see [10]. The standard Ensemble Classifier is implemented as a collection of Fisher Linear Discriminant (FLD) base learners.

To understand the FLD, suppose we have a training set of cover-stego pairs and suppose we can accurately classify the training set using only two features. This setup is highly improbable, but working in 2-dimensions will allow us to graph our features. The features (x,y) of each training image are plotted in Figure 3. The vector w is calculated to point in the direction that maximizes the between-class variance and minimizes the within-class variance. The decision hyperplane (dashed line) is orthogonal to w. The classifier predicts whether a test image x is cover or stego by projecting its features onto w:

$$g(x) = w^T x. \tag{2}$$

The projection value $g(x)$ is then compared to the decision threshold b:

$$\begin{cases} g(x) > b, & x \texttt{ is cover} \\ g(x) < b, & x \texttt{ is stego} \\ g(x) = b, & \texttt{class randomly assigned to } x. \end{cases} \tag{3}$$

The standard decision threshold is chosen to minimize the detection error in equation 1 on the training data, but it can be changed. In fact, we will change the location of the decision threshold to improve the detection error of LSB trained classifiers on MiPOD, S-UNIWARD and WOW.

Each FLD base learner in the Ensemble Classifier is constructed as described above except on a larger training set and feature space. Cover-stego pairs are randomly selected, with replacement, to be used for training. A subset of features from these cover-stego pairs is randomly selected, without replacement. A test image is voted on by each base learner in the classifier and the class with the majority of votes wins.

In order to improve the detection error of LSB trained classifiers, we adjust the standard decision threshold b for each individual base learner as follows:

$$b_{adj} = b - \lambda c \tag{4}$$

where λ is a tuning parameter and c is the standard deviation of the FLD projections of the training images.

We found the tuning parameter $\lambda = 0.75$ to produce decent classification results for experiments on iPhone data. In future research we plan to develop a definition of the optimal λ for a given dataset, as well as a systematic method for determining it. We believe it is likely that the optimal λ would be dependent on the dataset.

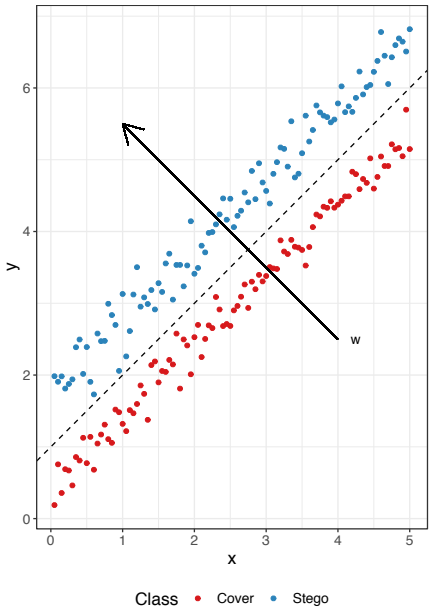


Figure 3. *Example of the FLD normal vector w and decision hyperplane (dashed line) on a feature set with 2 dimensions, x and y*

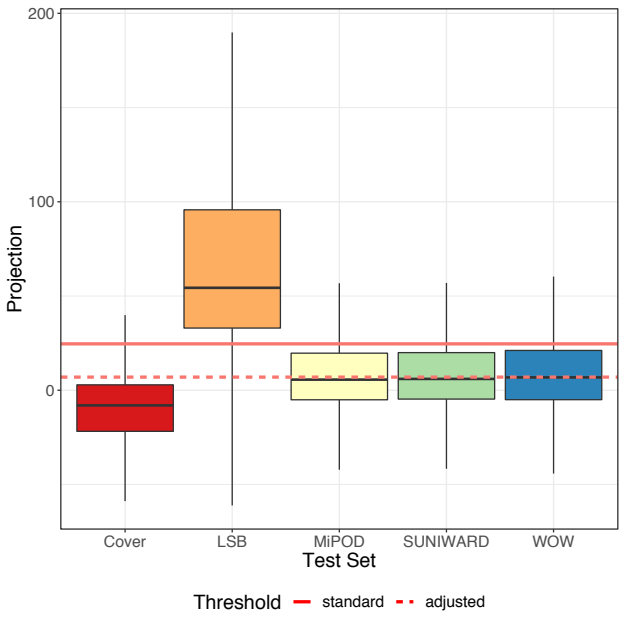


Figure 4. *Boxplots of the projections $g(x) = w^T x$ of test image features x onto the normal vector w of each FLD base-learner in an LSB trained Ensemble Classifier. Test images are classified based on which side of the decision threshold their projections fall.*

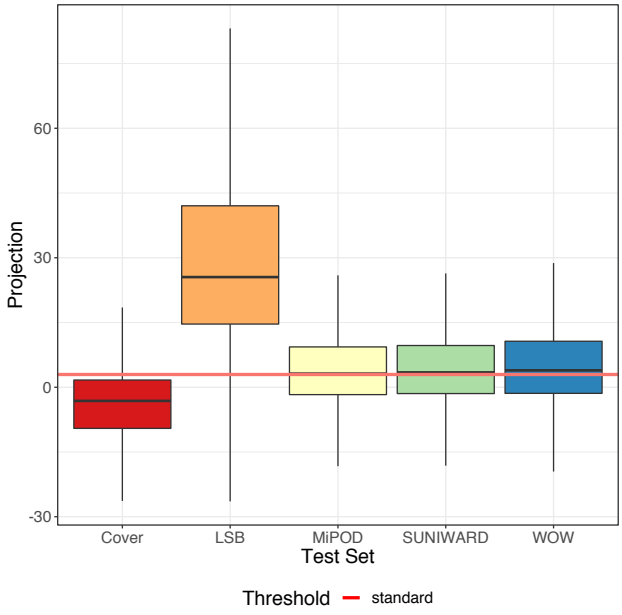


Figure 5. *Boxplots of the projections $g(x) = w^T x$ of test image features x onto the normal vector w of each FLD base-learner in an MiPOD trained Ensemble Classifier. Test images are classified based on which side of the decision threshold their projections fall.*

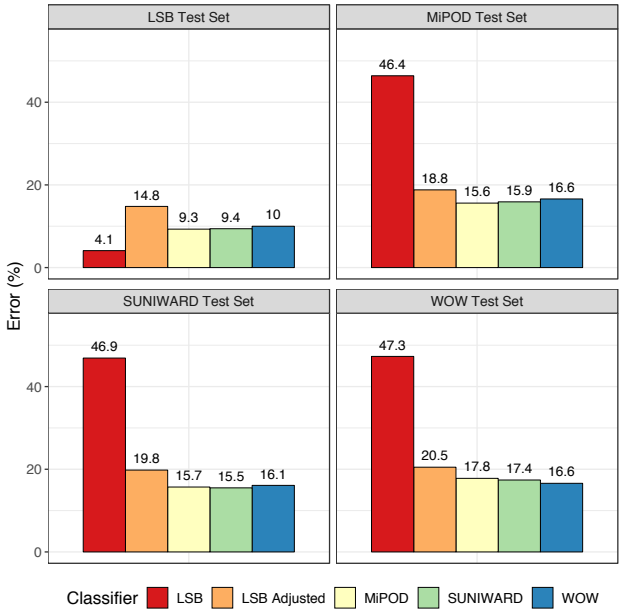


Figure 6. *Average error rate by classifier and test algorithm at 40% embedding on images from all BOSSbase devices (training size=5,000)*

Figure 4 shows the projections of test images when tested by LSB matching trained Ensemble Classifiers over five repetitions of ten-fold cross-validation with sample size 700. The figure does not show outliers, which account for roughly 12% of the projections for each image type. The median standard decision threshold is a solid red line and the median adjusted decision threshold is a red dashed line. In general, the projections of the cover images are the smallest, the projections of the LSB matching images are the largest and the, projections of MiPOD, S-UNIWARD and WOW fall in the middle. The median standard threshold successfully separates the cover images and LSB matching images, but mistakenly classifies the MiPOD, S-UNIWARD and WOW images as cover. However, the median adjusted threshold is able to recognize many of the MiPOD, S-UNIWARD, and WOW images as stego. Moreover, the adjusted decision threshold shown in figure 4 has the same relationship to the projections of all five image types as the median standard decision threshold of MiPOD trained classifiers shown in 5. This gives credence to the notion that an LSB matching trained classifier can achieve detection error rates when testing on MiPOD comparable to a MiPOD trained classifier testing on MiPOD.

Results

BOSSbase Dataset

We perform algorithm mismatch experiments on image data from all seven BOSSbase devices for three embedding rates: 10%, 20% and 40%. The results are shown in Figures 6-8. Each figure shows the detection error rates of five classifiers. The name of the classifier refers to the embedding algorithm upon which it was trained. The LSB Adjusted classifier is trained on cover and LSB matching data and the decision threshold is adjusted within the Ensemble Classifier during training as described in subsection Adjusting the Decision Threshold to Improve Classification Results of the Methods section. The other four classifiers, LSB, MiPOD, S-UNIWARD, and WOW, use the standard decision threshold. The grey titles denote the embedding algorithm being tested. The error is the detection error as calculated in equation 1 when testing covers and a single embedding algorithm.

Figures 6-8 show that adjusting the decision threshold when training on LSB matching data drastically improves the detection error when testing MiPOD, S-UNIWARD, and WOW, and produces error rates comparable to the best-case classifiers, typically within 3% or 4%. The MiPOD, S-UNIWARD and WOW classifiers achieve decent detection error rates on all four algorithms. Not unexpectedly, the overall error rates increase as the embedding rate decreases, but for each embedding rate the LSB Adjusted classifiers are fairly close to the best-case classifiers.

Pentex K20D Experiments

Previous work has shown that training and testing on a single device [22] can reduce detection error. We conduct algorithm mismatch experiments on a single BOSSbase device, the Pentex K20D, and show that algorithm mismatch classifiers see improved results comparable to the improvements for the best-case classifiers.

We use image data from the Pentex K20D digital still camera and run algorithm mismatch experiments using sample sizes of 700 and 1,300 and embedding rate 10%. The sample size 700 results are shown in Figure 9. The results on the larger sample size are similar, so we omit them here. As expected, we see that restricting the dataset to a single device shows slight reduction in error rates in the best-case classifiers compared to experiments on the entire dataset displayed in Figure 8. The algorithm mis-

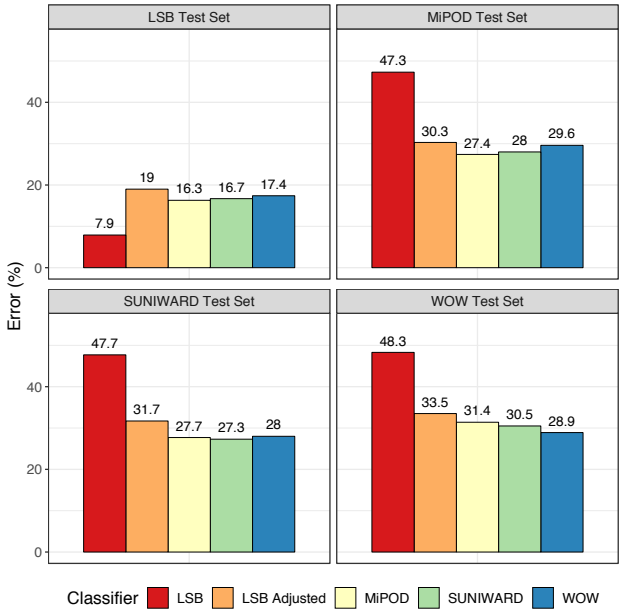


Figure 7. *Average error rate by classifier and test algorithm at 20% embedding on images from all BOSSbase devices (training size=5,000)*

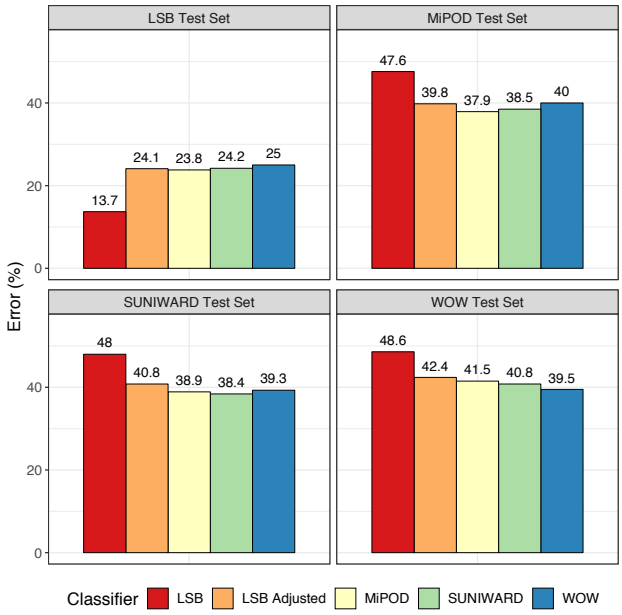


Figure 8. *Average error rate by classifier and test algorithm at 10% embedding on images from all BOSSbase devices (training size=5,000)*

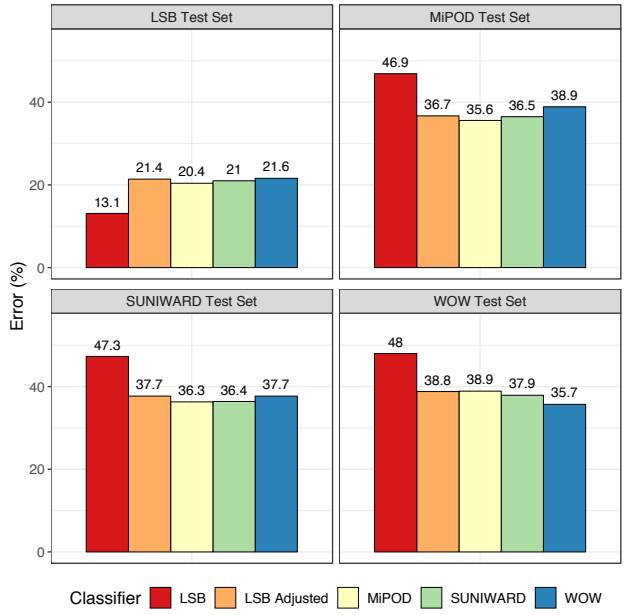


Figure 9. *Average error rate by classifier and test algorithm at 10% embedding on images from BOSSbase Pentex K20D device (cross-validation with sample size=700)*

match classifiers see similar reductions in error rates on the single device. The MiPOD and S-UNIWARD trained classifiers have testing errors within 1% of each other for all four testing algorithms. The LSB Adjusted classifier achieves testing errors within 2% of the MiPOD and S-UNIWARD trained classifiers for all four testing algorithms. The WOW trained classifier obtained testing errors within 3% of the MiPOD and S-UNIWARD trained classifiers.

iPhones Dataset

We perform algorithm mismatch experiments on the iPhone dataset. The results are shown in Figure 10. As we saw with the BOSSbase dataset, the LSB Adjusted classifier achieves decent detection error rates on MiPOD, S-UNIWARD and WOW, and MiPOD and S-UNIWARD do remarkably well at detecting each other.

Individual Device Experiments

We perform algorithm mismatch experiments on individual iPhone devices. This reduces the detection error rates of the best-case classifiers for four of the devices, while two devices see an increase in error rates. However, on all devices the algorithm mismatch classifiers achieve decent error rates in comparison to the best-case classifiers.

Figure 11 shows the detection error rates for the iPhone 6s (1) device with 10% embedding rate. The results for the iPhone 6s (2) and iPhone 6s Plus (1) devices are similar, within 1% of those shown in Figure 11 in most cases, so we omit them here. Figures 12-14 show the detection error rates for the iPhone 6s Plus (2), iPhone 7 (1) and iPhone 7 (2) devices respectively. Restriction to a specific device decreases the average testing errors in general for the iPhone 6s and 6s Plus devices, while the average testing errors

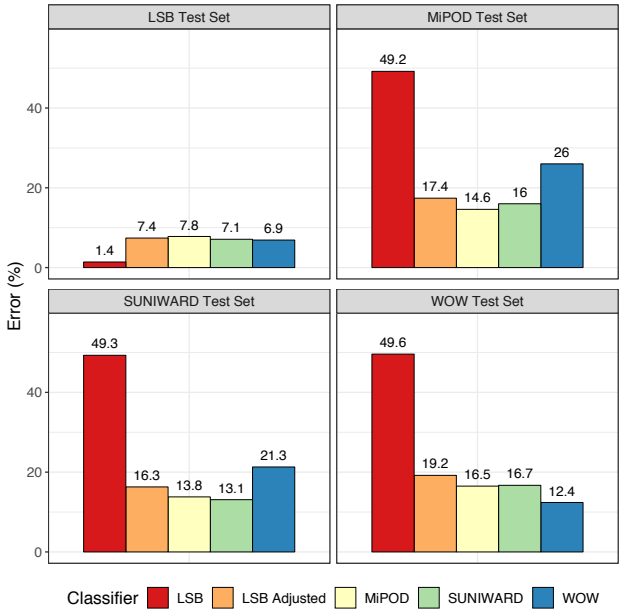


Figure 10. *Average error rate by classifier and test algorithm at 10% embedding on auto-exposure images from six iPhone devices (training size=5,000)*

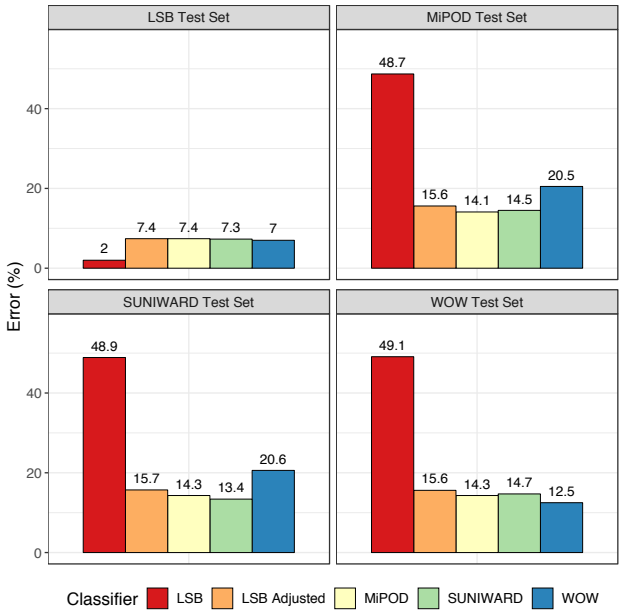


Figure 11. *Average error rate by classifier and test algorithm at 10% embedding on auto-exposure images from iPhone 6s (1) device (cross-validation with sample size=700)*

generally increase for the iPhone 7 devices. The LSB Adjusted classifier obtains average testing errors within 4% of the best-case classifiers for MiPOD and S-UNIWARD on all six devices. The LSB Adjusted classifier obtains average testing errors for WOW within 5% of the best-case classifier on the iPhone 6s and 6s Plus devices, within 7% for the iPhone 7 (1), but only within 12% for the iPhone 7 (2).

Cross-Phone Experiments

We consider the scenario where the steganalyst does not have the same device but has a device of the same model as well as devices from different models of the same make. We explored this scenario in previous work and showed that in some cases detection error on cross-device tests can be reduced by fixing the iso and exposure settings of the training and testing data [22]. Here we perform algorithm mismatch experiments across devices.

We train on auto-exposure image data from one iPhone device and test on each of the other iPhone devices. The results when training on the iPhone 6s (2) are displayed in Figure 15. We see that irrespective of the testing devices the MiPOD and S-UNIWARD trained classifiers obtain similar testing errors to each other on all four embedding algorithms. We also see that generally the LSB Adjusted classifiers achieve decent detection errors in comparison to the best-case classifiers. The error rates for training on the iPhone 6s (1) and both iPhone 6s Plus devices and testing on the other devices are similar to those shown in Figure 15 so we omit them here. The detection error rates when training on the iPhone 7 devices and testing on the other devices are almost all above 40%. This means that we can't necessarily expect to get adequate results when testing on an unseen device. If the best-case classifiers perform badly, so do the LSB Adjusted classifiers. However, if the best-case classifiers do well, the LSB

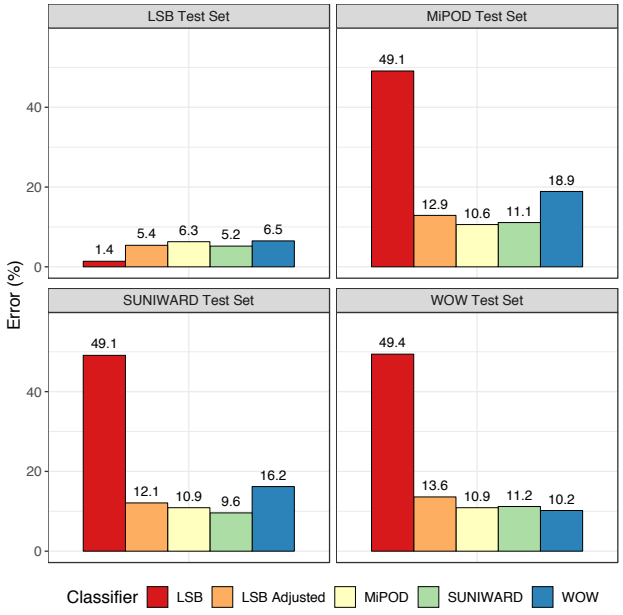


Figure 12. *Average error rate by classifier and test algorithm at 10% embedding on auto-exposure images from iPhone 6s Plus (2) device (cross-validation with sample size=700)*

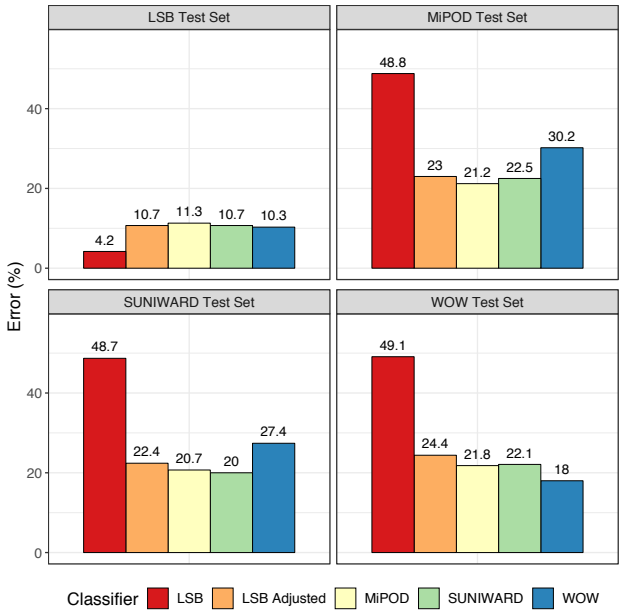


Figure 13. *Average error rate by classifier and test algorithm at 10% embedding on auto-exposure images from iPhone 7 (1) device (cross-validation with sample size=700)*

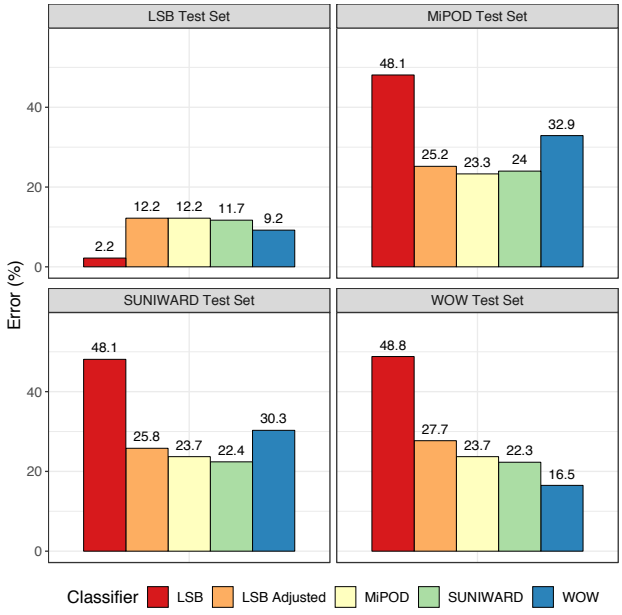


Figure 14. *Average error rate by classifier and test algorithm at 10% embedding on auto-exposure images from iPhone 7 (2) device (cross-validation with sample size=700)*

Adjusted classifiers achieve decent results as well.

Conclusions and Future Work

New stego algorithms will undoubtably continue to be created, increasing the likelihood that steganalysis classifiers will encounter unseen algorithms. We present a straight-forward and non-data-intensive steganalysis framework to address algorithm mismatch, the case where a classifier is trained on one algorithm and tested on another.

We train Ensemble Classifiers with Spatial Rich Model features on one of four embedding algorithms - LSB matching, MiPOD, S-UNIWARD, or WOW - and test the classifier on all four algorithms. We adjust the decision threshold of the Ensemble Classifier when training on LSB matching data and use the standard decision threshold when training on the other three algorithms. We use two datasets for training and testing: BOSSbase with 10%, 20% and 40% embedding rates, and iPhone data with 10% embedding rate. The average detection errors for the best-case classifiers, classifiers trained and tested on the same algorithm, for BOSSbase data were much larger than for the iPhone data. However, the LSB trained classifier with adjusted threshold, and the MiPOD and S-UNIWARD trained classifiers, achieved decent detection errors in comparison to the best-case classifiers on both datasets.

We plan to conduct algorithm mismatch experiments with more stego embedding algorithms, including at least one iPhone stego app. We also plan to further investigate and improve the selection of the tuning parameter λ used to adjust the decision threshold when training on LSB matching data.

Acknowledgements

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

References

[1] T. Pevný and J. Fridrich, "Novelty detection in blind steganalysis," in *Proceedings of the 10th ACM workshop on Multimedia and security*. ACM, 2008, pp. 167–176.

[2] X. Kong, C. Feng, M. Li, and Y. Guo, "Iterative multi-order feature alignment for jpeg mismatched steganalysis," *Neurocomputing*, vol. 214, pp. 458–470, 2016.

[3] T. Pevný and J. Fridrich, "Towards multi-class blind steganalyzer for jpeg images," in *International Workshop on Digital Watermarking*. Berlin, Heidelberg: Springer, 2005, pp. 39–53.

[4] ——, "Multi-class blind steganalysis for jpeg images," in *Security, Steganography, and Watermarking of Multimedia Contents VIII*, vol. 6072. International Society for Optics and Photonics, 2006, p. 60720O.

[5] ——, "Merging markov and dct features for multi-class jpeg steganalysis," in *Security, Steganography, and Watermarking of Multimedia Contents IX*, vol. 6505. International Society for Optics and Photonics, 2007, p. 650503.

[6] ——, "Multiclass detector of current steganographic meth-

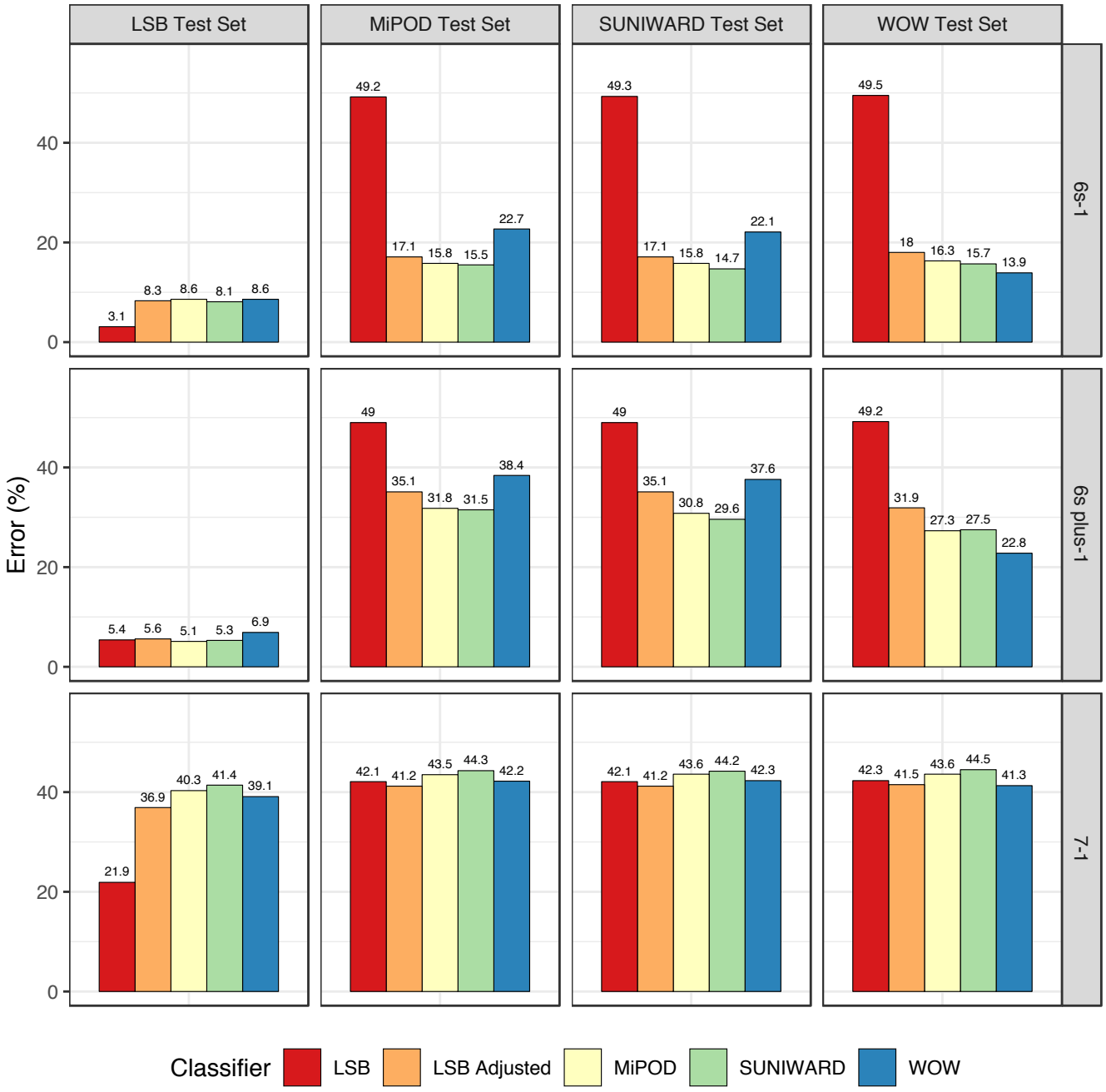


Figure 15. *Classifiers trained on auto-exposure images from iPhone 6s-2 device and tested on auto-exposure images from iPhone 6s-1, iPhone 6s Plus-1 and iPhone 7-1 devices. Average error rate by classifier, test algorithm and test device at 10% embedding (cross-validation with sample size=700)*

ods for jpeg format," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 635–650, 2008.
[7] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 221–234, 2016.
[8] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security*, vol. 2014, no. 1, p. 1, 2014.
[9] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *IEEE Workshop on Information Forensics and Security (WIFS)*. Tenerife, Canary Islands: IEEE, December 2012, pp. 234–239.
[10] J. Kodovský, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2012.
[11] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.
[12] T. Pevný, J. Fridrich, and A. D. Ker, "From blind to quantitative steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 445–454, 2012.
[13] T. Holotyak, J. Fridrich, and S. Voloshynovskyy, "Blind statistical steganalysis of additive steganography using wavelet higher order statistics," in *Proc. of the 9th IFIP TC-6 TC-11 Conference on Communications and Multimedia Security*, Salzburg, Austria, Sept 2005.
[14] M. Goljan, J. Fridrich, and T. Holotyak, "New blind steganalysis and its implications," in *Security, Steganography, and Watermarking of Multimedia Contents VIII*, vol. 6072. International Society for Optics and Photonics, 2006, p. 607201.
[15] A. D. Ker and T. Pevný, "A new paradigm for steganalysis via clustering," in *In Media Watermarking, Security, and Forensics III*, vol. 7880. International Society for Optics and Photonics, 2011, p. 78800U.
[16] Y. Wu, T. Zhang, X. Hou, and C. Xu, "New blind steganalysis framework combining image retrieval and outlier detection," *KSII Transactions on Internet & Information Systems*, vol. 10, no. 12, 2016.
[17] P. Bas, T. Filler, and T. Pevný, "Bossbase database." [Online]. Available: http://agents.fel.cvut.cz/stegodata/
[18] Center for Statistics and Applications in Forensic Evidence, "Stegoappdb homepage," https://forensicstats.org/stegoappdb/, 2018.
[19] W. Chen, Y. Wang, Y. Guan, J. Newman, L. Lin, and S. Reinders, "Forensic analysis of android steganography apps," in *IFIP International Conference on Digital Forensics*, G. Peterson and S. Shenoi, Eds. Springer, Cham, 2018, pp. 293–312.
[20] W. Chen, L. Lin, M. Wu, and J. Newman, "Tackling android stego apps in the wild," in *APSIPA ASC 2018*. (in press).
[21] L. Lin, W. Chen, Y. Wang, S. Reinders, Y. Guan, J. Newman, and M. Wu, "The impact of exposure settings in digital image forensics," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, October 2018, pp. 540–544.
[22] L. Lin, J. Newman, S. Reinders, Y. Guan, and M. Wu, "Domain adaptation in steganalysis for the spatial domain," in *IS&T Electronic Imaging: Media Watermarking, Security, and Forensics (IS&T, Burlingame, CA, 2018)*, no. 7, pp. 1–9.
[23] J. Newman, L. Lin, W. Chen, S. Reinders, Y. Wang, Y. Guan, and M. Wu, "Stegoappdb: a steganography apps forensics image database," in *IS&T Electronic Imaging: Media Watermarking, Security, and Forensics (IS&T, Burlingame, CA, 2019)*. (in press).

Author Biography

Stephanie Reinders received her BA in Journalism and Asian Languages and Literatures from the University of Minnesota (2005). After working for several non-profit organizations as an administrative assistant, she returned to school to earn a graduate degree in mathematics. She received a post-baccalaureate certificate in Mathematics from Smith College (2013) and currently is pursuing a PhD in Applied Mathematics and Computer Engineering at Iowa State University.

Li Lin received his B.S. degree in Mathematics from Capital Normal University, Beijing, China. He is currently pursuing the Ph.D degree in Applied Mathematics at Iowa State University, Ames, Iowa. His research interests include statistical image forensics, steganalysis, and statistical learning.

Dr. Yong Guan is a Professor of Electrical and Computer Engineering, the Associate Director for Research of Information Assurance Center, and the cyber forensics coordinator for NIST-CSAFE at Iowa State University. He received his Ph.D. degree in Computer Science from Texas A&M University. Supported by NSF, NIST, IARPA, ARO and Boeing, his research focuses on security and privacy issues, including digital forensics, network security, and privacy-enhancing technologies for the Internet.

Dr. Min Wu is a Professor of ECE and Distinguished Scholar-Teacher at the University of Maryland, College Park. She received her Ph.D. degree in electrical engineering from Princeton University. She leads the Media and Security Team (MAST), with main research interests on information security and forensics, and multimedia/multimodal signal and data science. She was elected as an IEEE Fellow and an AAAS Fellow for contributions to signal processing, multimedia security and forensics.

Dr. Jennifer Newman received her BA in Physics from Mount Holyoke College and her PhD in Mathematics from the University of Gainesville, FL. She is an Associate Professor of Mathematics at Iowa State University in the Department of Mathematics, her research focusing on image processing, stochastic modeling, steganalysis and image forensics. She is a member of IEEE, IS&T, SIAM, and IAI.

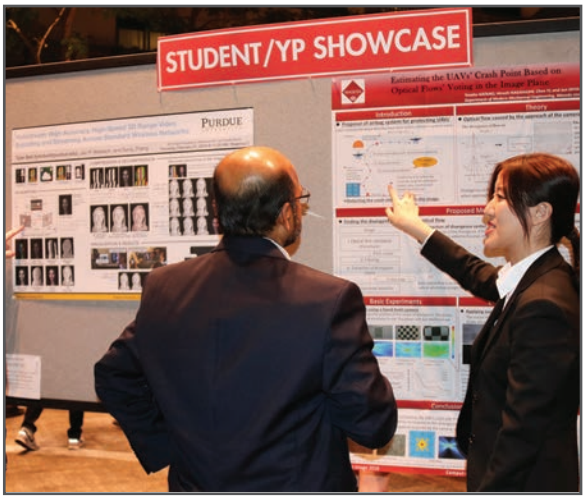
JOIN US AT THE NEXT EI!

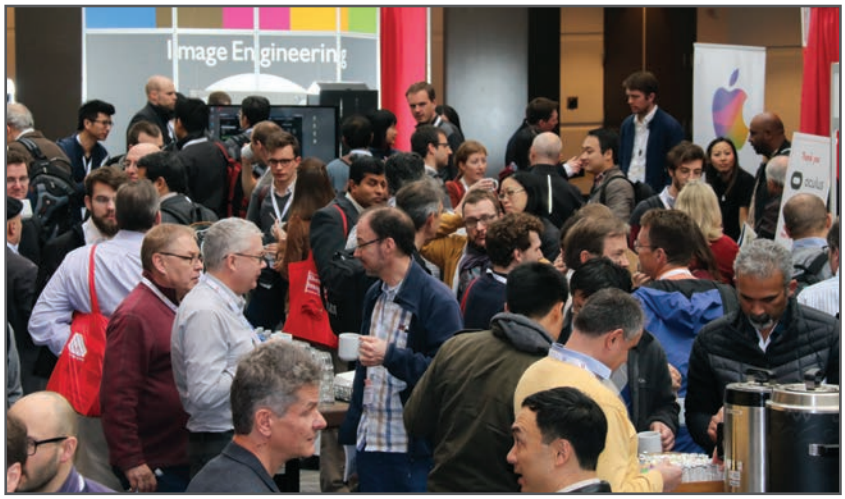
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org