Steganalysis of Overlapping Images

James M. Whitaker and Andrew D. Ker

Oxford University Department of Computer Science, Parks Road, Oxford OX1 3QD, England.

ABSTRACT

We examine whether steganographic images can be detected more reliably when there exist other images, taken with the same camera under the same conditions, of the same scene. We argue that such a circumstance is realistic and likely in practice. In 'laboratory conditions' mimicking circumstances favourable to the analyst, and with a custom set of digital images which capture the same scenes with controlled amounts of overlap, we use an overlapping reference image to calibrate steganographic features of the image under analysis. Experimental results show that the analysed image can be classified as cover or stego with much greater reliability than traditional steganalysis not exploiting overlapping content, and the improvement in reliability depends on the amount of overlap. These results are curious because two different photographs of exactly the same scene, taken only a few seconds apart with a fixed camera and settings, typically have steganographic features that differ by considerably more than a cover and stego image.

Keywords: Steganalysis, Calibration, Filtering, Digital Images, Machine Learning

1. INTRODUCTION

Perhaps the greatest challenge in the detection of steganographic data in digital media is the subtlety of the signal. In digital images, steganographic embedding methods from the earliest Least Significant Bit (LSB) methods to the most recent adaptive schemes[1–3] generally make changes of absolute value one: to pixel intensities in raw images, or transform coefficients in compressed images. Such changes are orders of magnitude smaller than natural variations between innocent images, especially when the payload is smaller than the maximum so that relatively few elements are changed, and moreso in uncompressed media than lossy compressed. Detection of hidden data is only possible if the media is processed to enhance the contrast between the stego signal and the cover content. In this paper we will concentrate on grayscale images, but the lessons apply equally to other media even if, at the moment, the literature for other domains is less developed.

Broadly speaking, steganalysis of digital images has employed two methods to enhance this 'signal-to-noise ratio'. The first is to *filter* the image using standard noise reduction methods, but to discard the content and keep only the noise: here the terminology can be confusing because the 'signal' is in the image noise, and the 'noise' is the cover content. The steganographic signal is attenuated by this filter to a smaller degree than the cover content. Filtering methods in steganalysis have developed steadily over the last 10 years, from simple Laplacian filters,[4] through multiple directional filters[5] to methods that also apply random projections.[6] Indeed, it is not unreasonable to say that steganalysis progress in the last 10 years has largely been driven by the development of better and more diverse filters in the feature calculations, as well as the capacity to manage large-dimensional 'rich features'.

The second method that is used to enhance the steganographic signal is known as *calibration*: by processing the suspected stego image, the analyst attempts to obtain some information about the cover that generated it, which in this work we call a *reference image*. Calibration has traditionally been limited to JPEG images, and performed using a decompress-crop-recompress cycle. Recently, steganalysis that exploits adaptive embedding, which can be thought of as a kind of calibration, has also been introduced. In this paper we propose a new kind of calibration based on an independent image, taken by the same camera with the same settings, of an overlapping scene. In uncompressed images, we show that a scene with overlapping content provides a good

Further author information:
A. D. Ker: E-mail: adk@cs.ox.ac.uk, Telephone: +44 1865 283530

reference, and gives improved detection accuracy when added to three existing steganalysis methods, on two different spatial-domain steganographic embedding algorithms, with two different payload sizes.

This paper is structured as follows. In Section 2 we discuss calibration in more detail, and how it might be used with images of overlapping content; we also argue that overlapping images may be a significant use-case for steganalysis in the real world.[7] In Section 3 we describe the experiments that test whether calibration by overlapping images is valuable for steganalysis, and report the results in Section 4: first, simply measuring the accuracy of calibrated against uncalibrated detectors, and then examining more closely the feature data in order to explain the results. Finally, we discuss ideas for making this idea more practical, and further directions for research, in Section 5.

2. CALIBRATION AND OVERLAPPING IMAGES

Many steganalysis methods make use of calibration, an idea first appearing in 2002.[8] A suspected stego object is manipulated with the ostensible aim of recovering information about the corresponding cover; we call the manipulated stego object a *reference* object.

Methods for creating the reference object have remained largely unchanged since their proposal. They are performed almost exclusively on JPEG images in the following manner: the suspected stego JPEG is decompressed, the first 4 rows and columns cropped away, then recompressed with the same quantization. By desynchronizing the 8×8 discrete cosine transform (DCT) grid, the stego noise is suppressed and some aspect of the cover content recovered: for example, the distribution of low frequency DCT coefficients in the reference should be similar to the cover. A detector can be created because steganalysis features from a stego object will depart from those of the reference object, whereas a cover and its reference should be close. An early detector was based on absolute differences between features of suspect and reference object.[9] With the advent of machine-learning steganalysis, calibration was applied by feeding the signed pointwise difference[10] or concatenation[11] of suspect and reference features into a classifier. Generally, it has been found that such calibrated features make more accurate and sensitive detectors of steganography, compared with uncalibrated features: for example, see the table of experimental results in Ref. 12.

In reality, calibration is itself a kind of filter,[11] in that it enhances stego noise and reduces cover content, and it is not necessarily the case that the reference object's features are particularly close to the original cover. We will return to this in Subsection 4.3

There has been almost no literature discussing calibration in raw uncompressed images. To our knowledge, the only successful endeavour is Ref. 13, where a crude form a calibration is achieved by adding adjacent pixels in an image. This is somewhat successful in the limited case of a particular detector based on histogram smoothness,[14] but has not been used subsequently.

Recently some steganalysis methods have been proposed[15,16] which make use of a different form of calibration (although the word is not used explicitly): reference features are computed from areas of an image less likely to contain hidden payload (through understanding the adaptivity of embedding or coding procedure), and compared with stego features computed from areas more likely to contain payload. Again, this enhances the stego signal, but only if the adaptivity criterion is somehow independent of the features, otherwise the part of the image used for a reference is not an unbiased sample.

If we wanted to use calibration in raw images, or perhaps improve it in JPEGs, how could we create a reference image? In this paper we consider the possibility that a steganalyst sees a large library of images from a target, some of which substantially overlap in scene content. If the overlapping images have equivalent camera settings then it is plausible they would have similar noise characteristics since they have approximately the same textures in frame. We emphasize that we would *not* expect them to have the same pixels, due to camera direction, changes in the scene over time, or even random fluctuations of photons hitting a camera CCD. But their aggregate characteristics should be similar. Modern steganalysis features for raw images are based on noise characteristics,[5,6] so this paper addresses the question: could an overlapping image be used as a reference?

Such a scenario is rather likely in practice, since the almost limitless storage capacity of digital media encourages photographers to make multiple captures of the same scene. Reasons for doing so include wanting a

photograph in which no person's eyes are closed, or everyone is smiling. Or consider situations where the photographer pans around a scene taking multiple photographs in order to capture the setting: these photographs will be substantially overlapping.

If a steganographer hides payload in an album of such images, but does not use every image (which would be a wise precaution to avoid pooled steganalysis[17]), this presents the steganalyst with an opportunity: a cover image substantially overlapping a stego image. Could a mismatch in their features lead to more accurate detection?

3. EXPERIMENTAL DESIGN

This is an initial study which mimics the most favourable circumstances, in classic 'laboratory conditions', simply to determine whether there is any value in calibrating using an overlapping image of the same scene. We begin with the following assumptions, all favourable to the detector:

(i) the steganalyst has access to the cover source, so that they can develop their own training data;

(ii) the embedding method and payload size are known;

(iii) the steganalyst is then given two images with overlapping content;

(iv) the first is known to be a cover, with the cover/stego status of the second to be determined;

(v) the camera characteristics (exposure, white balance, etc.) are the same for the two images.

These conditions are not outrageously favourable. (i) and (ii) are standard for 'laboratory conditions' steganalysis that dominates the literature. We have already argued that (iii) should often be true in practice, and it should be easy to automate detection of overlapping images in a set under examination. (iv) can be bypassed by testing each pair twice, with the images swapped: as long as the steganalyst sees *some* substantially overlapping images, and some such pairs truly consist of a cover and stego object, application should be possible. Even a prolific steganographer is unlikely to embed in every single image of their library. Finally, (v) can be verified if the images contain EXIF data.

Developing the calibration method to work in less favourable circumstances is for future work, although we touch briefly on (iv) in Subsection 4.2.

3.1 Overlapping image dataset

For our experiments we needed a library of natural cover images with controlled amounts of overlapping content, conforming to assumption (v) above. The first author generated such a set manually by the following procedure.

The chosen camera (Canon G16) was set on a tripod facing an outdoor scene, the exposure and lens characteristics were fixed, and a RAW image was taken. Shortly afterwards another RAW image was taken of the same scene (100% overlap), then the camera was panned right (horizontally) to overlap approximately 75% of the original image, and another image was taken. This was repeated to generate images overlapping the original by 50%, 25%, and finally 0% (a scene approximately adjacent to the original). The last image was captured twice. The seven images of each scene were labelled A to G, and in the experiments we will use images B to F as reference covers for image A. An example of the seven images is shown in Figure 1. This procedure was repeated 500 times with different scenes, taken across a few days in a variety of locations around Oxford, UK.

RAW images from this camera model are large – 4000×3000 pixels – and it was time consuming to take even as many as 500 sets of overlapping images. To increase the experimental base, we took all images in portrait orientation, and then cut every image into five horizontal slices of 3000×800. This leads to images that are not entirely independent, but does increase their diversity because some slices contain mostly sky, and some mostly textured ground. After slicing, we had 2500 sets of seven overlapping images: image sets A and B are independent captures of the same 2500 scenes, while images in sets A and C contain approximately 75% overlapping content, and so on.

The raw images were converted to bitmap using the Canon software with default settings, and then to grayscale. In this study we only examined uncompressed grayscale images; a small pilot study using JPEG versions of the same images will be mentioned in Section 5.

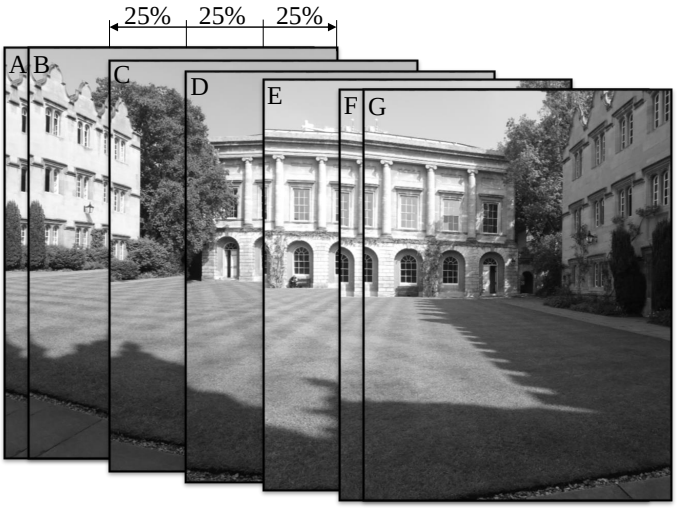

Figure 1. An example of seven overlapping images of the same scene, captured as part of the experimental library. They are shown here slightly staggered, so that all images can be seen.

3.2 Embedding

We tested two types of spatial-domain steganography: classical least significant bit matching (LSBM), and the adaptive embedding algorithm HUGO.[1] We selected payload rates of 0.01 and 0.02 bits per pixel (bpp) for LSBM, and 0.05 and 0.1 bpp for HUGO. These payloads are lower rates than typically found in steganalysis literature because of the image size: our images, even cut into slices, are substantially larger (2.4 Mpix) than the BOSSBase images (0.25 Mpix) on which steganalysis is often benchmarked, and the square root law of capacity[18] suggests that lower payload rates will be detectable in larger images. These payload rates lead to detector error rates of around 5-15% using uncalibrated steganalysis, which leaves scope for improvement.

While embedding, we found that a few image sets were unsuitable for HUGO. These were images where the top slice contained entirely sky, and which was particularly flat: HUGO embedding failed on such images. Where this happened we discarded the entire scene and replaced it with another for all experiments, so that the same scenes, with five slices, were usable in every case.

3.3 Features

We tested calibration of three different spatial-domain features from the steganalysis literature:

(i) **SPAM**[4] (Subtractive Pixel Adjacency Matrix) features were the state-of-art technique from around 2009–2011. The image is subject to Laplacian filters in different directions, and the features consist of co-occurrence probabilities for triplets of residuals in the filtered images. The standard version of SPAM, which we used, is 686-dimensional.

(ii) **SRM**[5] (Spatial Rich Model) features were proposed in 2012, although the ideas behind them emerged in 2011. They were the first features able to detect HUGO steganography reliably, and are still amongst the leading methods. They can be seen as a substantial generalization of SPAM, in that the image is subject to a number of different filtering operations before co-occurrences of triplets computed. There is some symmetrization, and quantization, for which various options exist. The version we use is called SRMQ1 and corresponds to the quantization $q = 1c$ in Ref. 5, and gives 12753-dimensional features.

(iii) **PSRM**[6] (Projected Spatial Rich Model) features were proposed in 2013,[6] a modification of SRM features. They use histograms of randomly filtered residuals, instead of co-occurrences. Computing so many convolutions is slow, but we used an optimized GPU implementation.[19] Various parameters can be adjusted,

and we opted to use the defaults from Ref. 6 except for slightly fewer random projections per residual (30 instead of 55) leading to more compact features that are 7020-dimensional.

We did not test the SPAM features against HUGO steganography, because they are known to be ineffective against it (one of HUGO's design criteria was to minimize distortion to SPAM features). We tested all other combinations of steganographic embedding, payload sizes, and steganalysis features.

3.4 Calibration

Suppose that a feature vector x has been extracted from a reference image, and y from a suspected stego image (both n-dimensional vectors). Recall that we are assuming that x is *known* to come from a cover image, but the status of y is unknown.

How should we use x to calibrate y? It will be a function that combines the features, $\kappa(x, y)$. Calibration in JPEG steganalysis literature either measures directly the difference between the features by constructing the n-dimensional vector $y - x^*$, or simply forms the $2n$-dimensional concatenation $x \cdot y$ and leaves the machine learning algorithm to find a good relationship between x and y for classification. Note in the second case we can expect the classifier to be slower to use and significantly slower to train, because its features are twice as large. On the other hand, if $\kappa(x, y) = y - x$ contains the best information then in theory a linear classifier on $\kappa(x, y) = x \cdot y$ can determine it, because the hypothesis space of linear classifiers depending on $y - x$ is contained in that depending on $x \cdot y$; this is not to say that the classifier will find the same solution in practice, and indeed it does not.

Subtraction is not the only way to measure difference between two feature elements, and one could use absolute difference, squared difference, or various normalized differences. Many options were explored by the first author in Ref. 20, which found that combinations of subtraction, concatenation, and an unusual quotient $(x^2 - y^2)/(x^2 + y^2)$[†] were the best performers, so in this paper we test the following six methods of calibration:

- Subtraction of the reference features from the suspected image features: $\kappa(x, y) = y - x$.
- Concatenation: $\kappa(x, y) = x \cdot y$ (known as *Cartesian calibration* in the literature).
- The strangely effective normalized difference: $\kappa(x, y) = (x^2 - y^2)/(x^2 + y^2)$.
- Three combinations of the above: $\kappa(x, y) = x \cdot (y - x)$, $\kappa(x, y) = x \cdot y \cdot (y - x)$, and $\kappa(x, y) = (x - y)^2/(x + y) \cdot (y - x)$.

We also tested $\kappa(x, y) = y$, which turns off calibration altogether, providing a baseline against which to test whether the overlapping images provided additional detection power.

It is entirely possible that better methods of calibration exist, including functions κ that do not necessarily operate pointwise on individual features. In steganalysis, calibration has not advanced to the same extent as filtering, and more research is called for.

3.5 Classification and benchmark

Finally, we want to use calibrated features to classify unknown images. We follow the current state-of-art in steganalysis by using the ensemble Fisher Linear Discriminant (FLD) classifier from Ref. 21. This uses a number of linear base learners, each a simple FLD trained on a random subspace selecting some of the features, and using a majority vote of base learners to classify unknown images.

The ensemble has two parameters: the number of dimensions sampled by each base learner d_{sub}, and the number of base learners L. The reference implementation of the ensemble[‡] can optimize both parameters

[*] All arithmetic operations on vectors are pointwise, and $\cdot$ represents concatenation (not multiplication) of one vector with another.

[†] We admit that we have no intuition as to why this normalized difference works better than others.

[‡] Available from `http://dde.binghamton.edu/download/ensemble/`.

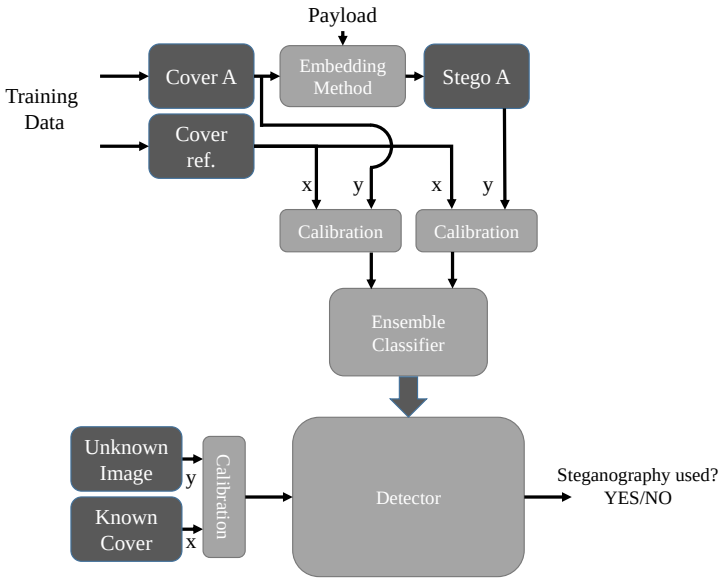

Figure 2. Use of overlapping image sets, calibration, and ensemble classifier, in our experiments.

automatically with respect to out-of-bag (OOB) error, and we allowed it to do so with respect to L, but chose to depart from this in the case of d_{sub}. We found that its search for optimal d_{sub} sometimes failed to find large enough values, leading to intermittent high error rates. In order to make comparisons between different experiments as fair as possible, we tried fixed values $d_{sub} \in \{50, 100, 150, 200, 250, 300, 400, 500, 600, 800, 1000, 1500, 2000, 3000\}$ and took the best-performing value in each case. We also fixed the random selection of subspaces to be the same in every experiment.

When classifying features formed by a concatenation, we fixed the random subspace selection to include corresponding components from each part: this allows the base learners to see the corresponding feature from x for every feature they see from y (or $y - x$, or both). In fact we observed little difference, when testing on preliminary data, between classifiers that enforce this selection and those that select purely at random. This seems surprising, but perhaps not when you consider that the SRM and PSRM features contain many strong colinearities (see Subsect. 4.3): thus some features can serve as a proxy for others.

A classifier is benchmarked by its accuracy in labelling unknown data as cover or stego, which involves setting a threshold for positive detection. We follow the literature in adopting the equal-prior error rate

$$P_{\mathrm{E}} = \tfrac{1}{2} \min(P_{\mathrm{FP}} + P_{\mathrm{FN}}),$$

where P_{FP} and P_{FN} represent the false positive and false negative rate on the testing data, and the minimum ranges over thresholds. Rather than measuring OOB error, which is asymptotically unbiased but not necessarily so in small samples, we used a cross-validation method: on each of five iterations, a different 20% of the data was used for testing and the remaining 80% for training and optimizing parameters, and we will report the average testing error over these five folds. There are certainly arguments for using other measures of accuracy, but what is important for this study is the relative, rather than absolute, performance of uncalibrated detection and calibration with different methods.

For each experiment, our suspect images will be from set A, and we will simulate training and testing stego images by embedding in them. The known reference images will be from one of the other sets; for example if we use set B (respectively C) as a reference then they overlap by 100% (75%), and so on down to adjacent images of the same scene that have no overlapping content. The process of training and testing images is depicted in Figure 2.

$\kappa(x,y)$		Overlap (uncropped)					Overlap (cropped)		
		100%	75%	50%	25%	none	75%	50%	25%
y (baseline)	0.089						0.101	0.112	0.139
$x \cdot y$		0.023	0.063	**0.078**	0.088	0.095	0.051	0.076	0.097
$x - y$		0.033	0.078	0.095	0.110	0.111	0.066	0.097	0.116
$\frac{x^2-y^2}{x^2+y^2}$		0.031	0.091	0.110	0.130	0.141	0.072	0.101	0.134
$x \cdot (y - x)$		0.022	**0.060**	0.081	**0.084**	0.093	0.049	0.077	0.096
$x \cdot y \cdot (y - x)$		0.023	0.067	0.080	0.090	0.095	0.049	0.077	0.095
$\frac{x^2-y^2}{x^2+y^2} \cdot (y - x)$		**0.018**	0.062	0.083	0.105	0.110	**0.048**	**0.075**	0.104

Table 1. Error rates of standard and calibrated steganalysis detectors (ensemble classifier, SRM features) for LSBM steganography (payload 0.01 bits per pixel). For each amount of overlap, and with the overlapped region cropped or not, we measure error rates under each different calibration method. Where it is superior to the baseline uncalibrated detector, we highlight the best performing calibration method in each column.

We also simulated the situation in which the steganalyst crops two partially-overlapping images to keep approximately the overlapping part, discarding the rest. We could do this by taking the right 75% of images from set A and calibrating them using the left 75% of images from set C, similarly 50% and 25% with sets D and E. Thus we can examine whether it is worthwhile to increase the proportion of overlap at the expense of the number of pixels of evidence.

4. EXPERIMENTAL RESULTS

We now describe experiments to evaluate whether calibration by overlapping images improves accuracy of detecting steganography (Subsection 4.1), which is the main aim of this paper, and some additional investigation of whether the calibration is robust when the status of the reference object is not known (Subsection 4.2). We also measure the extent to which overlapping scenes do have similar steganalysis features (Subsection 4.3).

4.1 Improvement in classification accuracy

We display the error rates for classifying cover objects against stego objects created using LSBM at 0.01 bits per pixel, using SRM features, with the six different methods of calibration (as well as no calibration), for images with 100% down to 0% overlap, and for images cropped to the overlapping region, in Table 1. We note that calibration improves detection accuracy, and the extent of the improvement depends on the amount that the reference image overlaps the suspect image: error rates decrease from 8.9% (uncalibrated) to around 2% (calibrated by various options including concatenation) when the reference overlaps by 100%, but when the overlap is only 50% the error rate can only be reduced to 7.8%, and if the overlap is less than 50% the improvement is insignificant or nil. We also see that cropping the image to the overlapping region makes a slight improvement over not doing so. Bear in mind that cropped images are smaller, and the square root law of capacity[18] implies that equal payload *rates* are less detectable in smaller images: we can see this in the increasing error rates of the baseline, uncalibrated, detector on cropped images in Table 1. The calibration has first to overcome the disadvantage of having a smaller image: it seems that it can do this, but not much more.

We should consider the possibility that the improved performance, with calibrated detectors, is purely due to additional training data: when we calibrate, we are supplying the machine learning algorithm with cover data from set B as well as set A. We should be particularly careful because the number of images per set, 2500 of which 2000 are used for training in each fold of the cross-validation, is not large compared with the dimensionality of the features, so our learners might be under-trained. To test this hypothesis, we trained an uncalibrated classifier using training cover data from both sets A and B combined, but stego data only from set A: this matches the

data available to the calibrated classifiers. We did this by adding cover data from set B into training data in every fold of the cross-validation.

Take the case corresponding to Table 1, LSBM steganography at 0.01 bpp, and using the SRM features. The ensemble FLD classifier improves its error rate to 8.5%, compared with 8.9% in the absence of set B. This is insignificant, particularly considering the improvement to 2.3% by cartesian calibration. When we tested HUGO embedding at 0.05 bpp, adding cover data in set B slightly worstened the error rate from 16.0% to 16.3%, compared with 9.8% obtained from the same data by cartesian calibration. It is evident that it is the connection between images in sets A and B – that they are different captures of the same scene – that provides detection power.

Error rates are easier to interpret in graphical form, and we include figures showing the experimental results for both embedding methods, both payloads, and each different steganalysis features, in Figures 3–5. From the figures we draw the following conclusions.

- If a substantially overlapping image is available, calibration is valuable. This is true for all combinations of embedding method, payload size, and features. The improvement depends on the amount of overlap, reducing errors by a factor of 2 or more at 100% overlap, decreasing as the overlap decreases, and retains some value when the reference image does not overlap against LSBM, but not HUGO.
- It seems that error rates are most reduced when the detector was quite good to start with. For example, when detecting HUGO at 0.05bpp via SRM features, error rates only drop from around 16% to 10%. But detecting LSBM at 0.02bpp with SRM features, the baseline error rate is around 4% and this drops to 1% on best calibration (a factor of over 4), and remains below the baseline even at 0% overlap; with PSRM features the baseline drops from around 6% to 1% (a factor of 6). There may be a theoretical model explaining the observation that calibration is most beneficial in strong detectors, but we have not yet developed one.
- Against LSBM embedding, it is valuable to crop the images to their overlapped region. This is particularly the case for PSRM features. On the other hand, against HUGO embedding it is counterproductive to crop; perhaps this is because HUGO payload is spread unevenly around an image.
- Of the calibration functions, the best option depends on the features used and the embedding method targeted, but the differences between them are mostly small. $\kappa(x, y) = x \cdot y$ is a good all-round performer, and $\kappa(x, y) = x \cdot (y - x)$ generally slightly better. The unusual normalized difference $\kappa(x, y) = (x^2 - y^2)/(x^2 + y^2)$ is effective in high levels of overlap, with PSRM and SRM features, against LSBM only.

4.2 Absence of known reference image

By using a reference image that is known to be a cover, our experiments have simulated a situation in which the analyst has a lot of information, perhaps unreasonably so. We briefly ask whether the calibration technique could be extended to cases where there are overlapping images but no known reference.

To examine this situation we tested classifiers, trained on pairs of known reference and stego object, in situations where the objects pairs were different. We use the notation $\mathbf{C}_A$, $\mathbf{C}_B$, etc, for cover images from sets A, B, etc, and $\mathbf{H}_A^{0.05}$ to indicate stego images from set A with HUGO embedding at 0.05 bpp, analogously $\mathbf{L}_A^{0.01}$ for LSBM. Training a calibrated detector on ($\mathbf{C}_B$, $\mathbf{H}_A^{0.05}$) or ($\mathbf{C}_B$, $\mathbf{H}_A^{0.1}$), we display the error rates when this detector is used to classify other pairs, in Table 2.

We make some preliminary observations from these results. From the first two columns of the table, we observe that we can use calibrated detectors in cases where the stego object has a different payload, with relatively little penalty. The exception is when training on large payload and testing on small payload, but this is most likely due to the P_{E} optimization choosing a detection threshold which does not catch the smaller payload, and can be fixed by setting a threshold for a desired false positive instead of maximizing P_{E}. From the third column of the table we see that we cannot detect anything if the reference object contains the same payload as the stego object; this is unsurprising. On the other hand, the fourth and fifth columns suggest that we *can* detect differences in

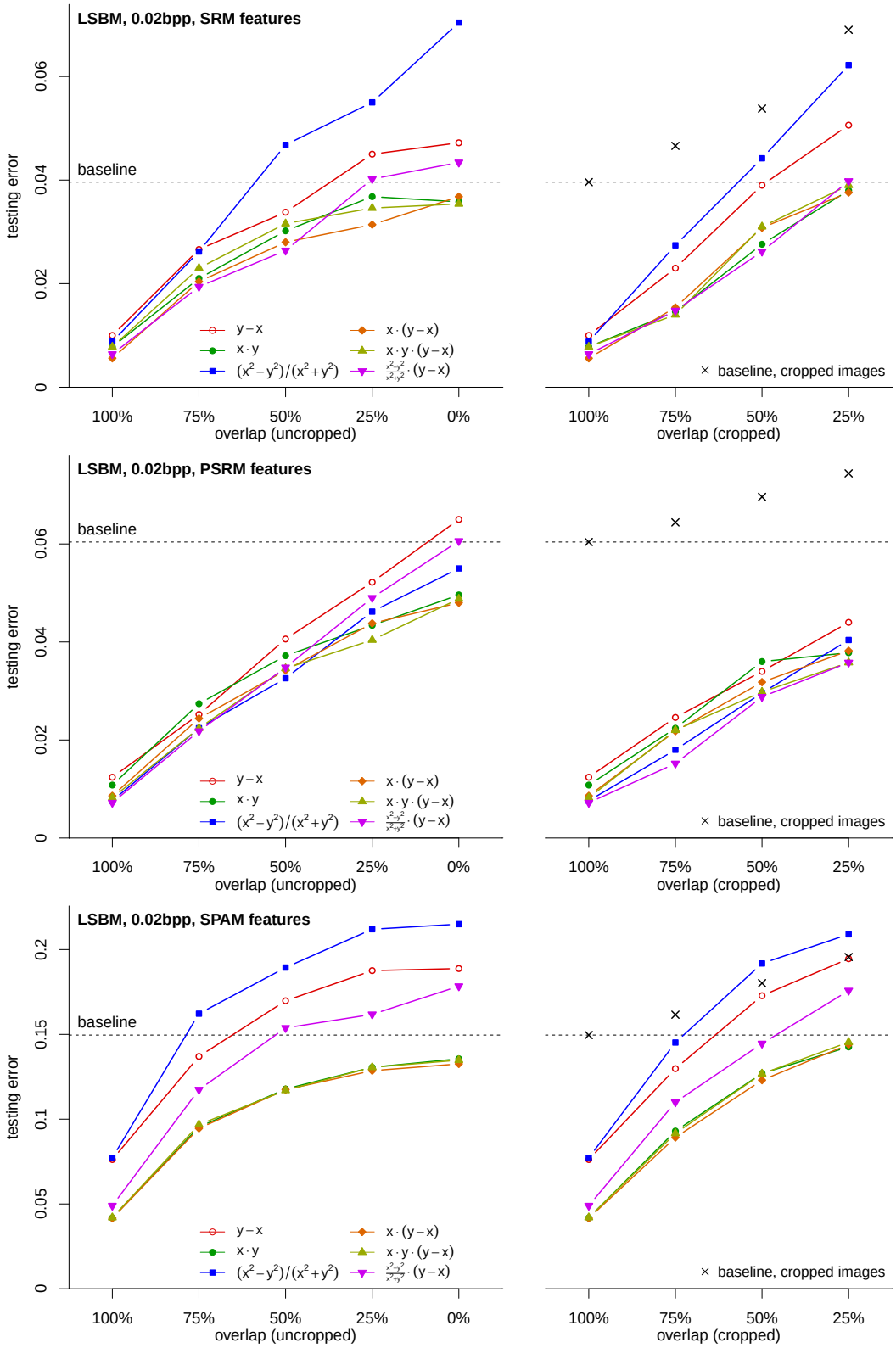

Figure 3. Error rates of detectors with six types of calibration, as amount of overlap varies. The uncalibrated detector is shown as a dashed line. For cropped images, crosses denote the performance of the uncalibrated detector on images of that size. LSBM steganography at 0.02 bits per pixel, with SRM (top), PSRM (middle), and SPAM (bottom) features.

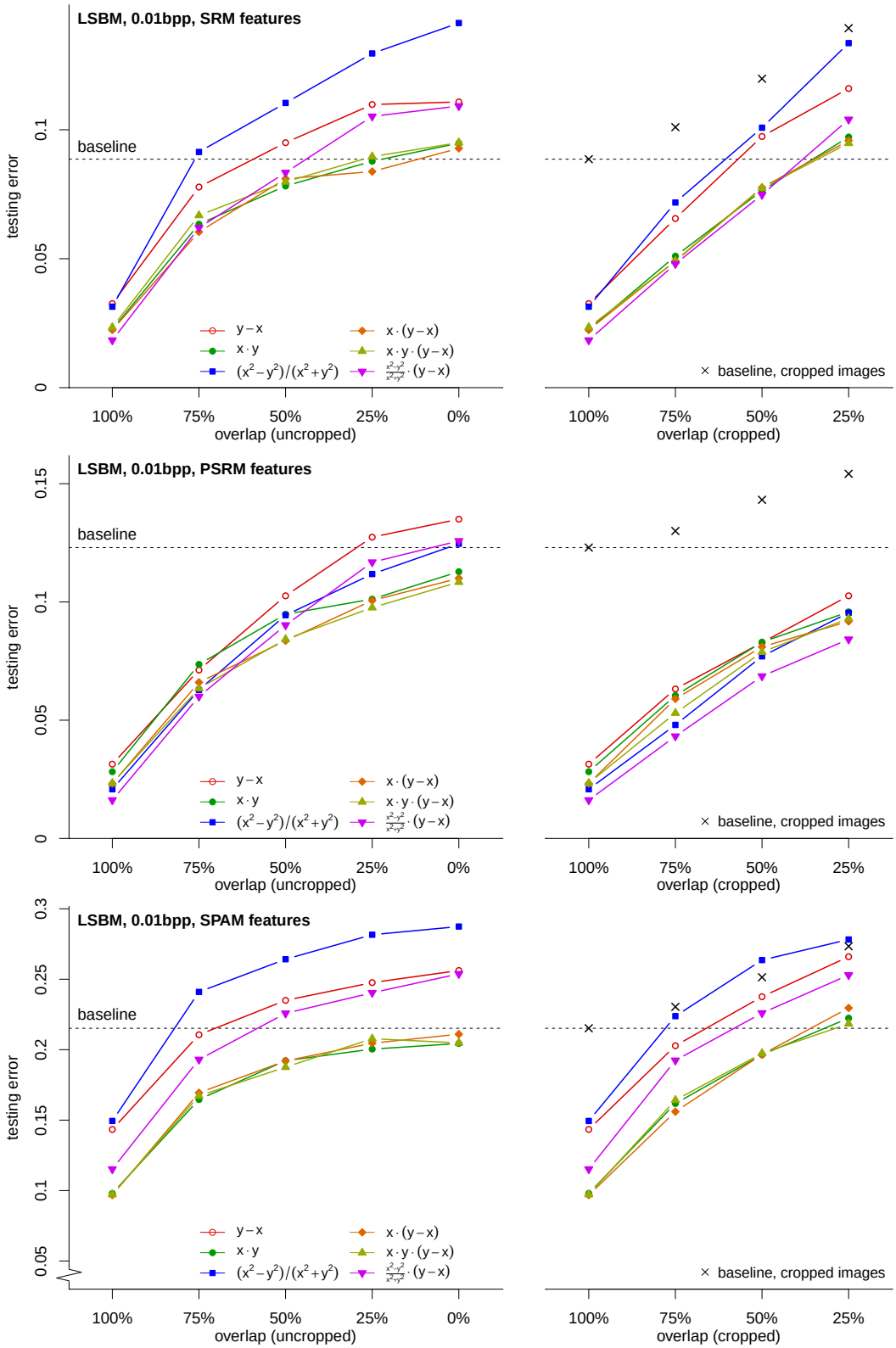

Figure 4. Error rates of calibrated and uncalibrated detectors for LSBM steganography at 0.01 bits per pixel, with SRM (top), PSRM (middle), and SPAM (bottom) features.

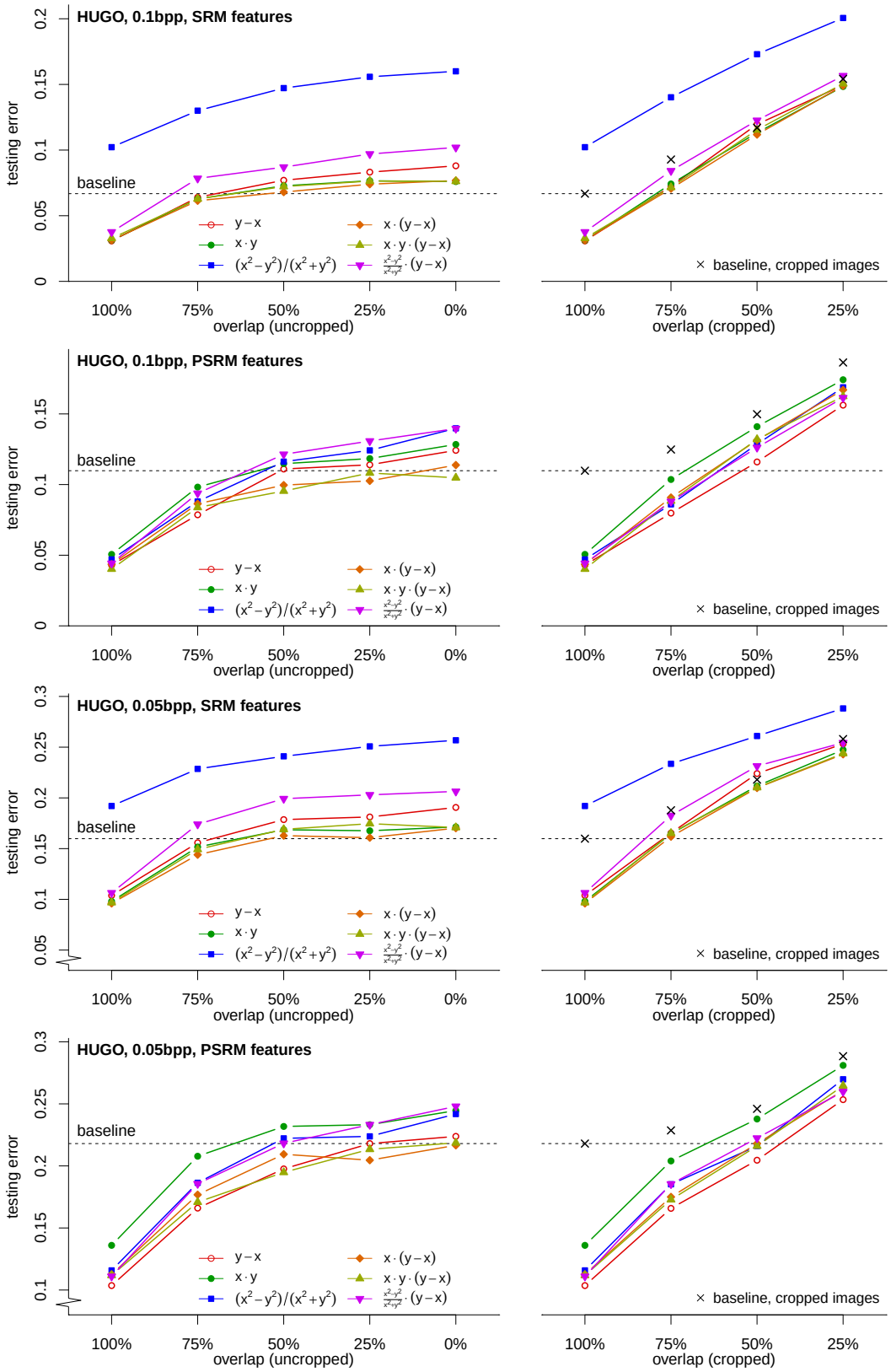

Figure 5. Error rates of calibrated and uncalibrated detectors for HUGO steganography at 0.05 and 0.1 bits per pixel, with SRM and PSRM features.

Training pairs	Test pairs					
	$(\mathbf{C}_B, \mathbf{H}_A^{0.05})$	$(\mathbf{C}_B, \mathbf{H}_A^{0.1})$	$(\mathbf{H}_B^{0.05}, \mathbf{H}_A^{0.05})$	$(\mathbf{H}_B^{0.05}, \mathbf{H}_A^{0.1})$	$(\mathbf{H}_B^{0.05}, \mathbf{C}_A)$	$(\mathbf{C}_C, \mathbf{H}_A^{0.05})$
$(\mathbf{C}_B, \mathbf{H}_A^{0.05})$	0.102	0.042	0.497	0.063	0.109*	0.175
$(\mathbf{C}_B, \mathbf{H}_A^{0.1})$	0.225	0.031	0.500	0.131	0.219*	0.247
						$\kappa(x, y) = y - x$
$(\mathbf{C}_B, \mathbf{H}_A^{0.05})$	0.105	0.061	0.463	0.068	0.405*	0.214
$(\mathbf{C}_B, \mathbf{H}_A^{0.1})$	0.340	0.034	0.489	0.112	0.474*	0.296
						$\kappa(x, y) = x \cdot y$

Table 2. Error rates of calibrated detectors trained and tested in mismatching pairs. $\mathbf{C}$ denotes covers and $\mathbf{H}^p$ stego objects embedded with HUGO at p bpp. *indicates that, after training, the sign of the projections and thresholds were swapped in each base learner.

payload between the reference and the suspect object, albeit with slightly lower accuracy and with the need to use the $\kappa(x, y) = x - y$ calibration, switching the signs inside the classifier, in the case when the 'reference' object has *more* payload than the image being analysed. This is a good direction for further research, suggesting that we should train a regressor, rather than a classifier, and that it will be estimating *difference* in payload. After setting a threshold symmetrical about zero, we should be able to construct a base learner for an ensemble which can detect payload in overlapping images whenever the payload rates are significantly unequal.

Finally, the last column of Table 2 suggests that calibration might still be valuable even if the training pairs contain a different amount of overlap from the testing pair. However, the accuracy is quite low, and further work would be needed to try to mitigate this. We could try including the amount of overlap as an additional feature, and train on overlapping images with diverse payloads and amounts of overlap, with the dependent variable being the difference in payload rate. This is for future research.

4.3 Closeness of reference to cover

Finally, we performed some simple experiments to investigate whether calibration really achieves its aims, asking: is the reference image actually a good proxy for the cover? Are they in some sense close?

Measuring distance between images is difficult. Distance in the spatial domain is irrelevant: what we care about is distance between the features used for steganalysis. But, apart from the weak SPAM features, the feature dimensionality is very high (7020 and 12753 for our versions of PSRM and SRM), much higher than the number of images in each set (2500). Furthermore, the features are strongly co-linear, which makes Euclidean distance a false metric.

We chose to measure distance in a feature space whitened by PCA. Our procedure was as follows. First, we took the set of SRM features pooled from $\mathbf{C}_A$ and $\mathbf{H}_A^{0.05}$, centered their mean on the origin, and applied PCA. This selection of features should ensure that directions that explain variation in both cover and stego images are preserved. We kept only the projections which explained at least 10^{-6} of the variance, which was only 18 dimensions, discarding the rest as insignificant noise. That only 18 dimensions explains more than 99.999% of the variation between features demonstrates how tightly-linked they are. Finally, after this PCA rotation, we normalized each dimension to unit variance.

After applying this same whitening projection to every image set, we calculated the mean Euclidean distance between different members of $\mathbf{C}_A$, and between corresponding members of $\mathbf{C}_A$ and $\mathbf{C}_B$, $\mathbf{C}_C$, etc, and $\mathbf{H}_A^{0.05}$. Finally, we normalized the measurements (whose scale is arbitrary) so that the mean distance between different members of $\mathbf{C}_A$ is unit. We display the resulting mean distances, in this whitened space, in the first row of Table 3. These results confirm our intuition that overlapping images should be closer to each other than independent images: the distance between features from different images in cover set A are approximately 16

SRM features: relative distance from $\mathbf{C}_A$	$\mathbf{C}_A{}^*$	$\mathbf{H}_A^{0.05}$	$\mathbf{C}_B$	$\mathbf{C}_C$	$\mathbf{C}_D$	$\mathbf{C}_E$	$\mathbf{C}_F$
... in whitened space	1.000	0.034	0.063	0.281	0.445	0.564	0.650
... projected onto regression vector	1.000	4.076	1.507	1.594	1.682	1.705	1.694
PSRM features: relative distance from $\mathbf{C}_A$	$\mathbf{C}_A{}^*$	$\mathbf{L}_A^{0.01}$	$\mathbf{C}_B$	$\mathbf{C}_C$	$\mathbf{C}_D$	$\mathbf{C}_E$	$\mathbf{C}_F$
... in whitened space	1.000	0.226	0.052	0.276	0.440	0.561	0.648
... projected onto regression vector	1.000	3.519	0.628	0.689	0.792	0.852	0.928

Table 3. Average distance between features, as a Euclidean distance after PCA and renormalization, and on the most effective detection direction. In each case the distance is measured between corresponding images of the same scene, except *, which measures the average distance between different cover images. All distances have been normalized so that the first column is 1.

times further apart than corresponding images in sets A and B (100% overlap), approximately 3.5 times further apart than corresponding images in sets A and C (75% overlap), and so on; adjacent but non-overlapping images in sets A and F are approximately 1.5 times closer than independent images.

Compare this, though, to the distance between a cover and the corresponding stego object, embedded by HUGO at 0.05 bpp: this is approximately *a factor of 30* smaller than the distance between different covers, and only about half as far as the distance between two consecutive captures of the same scene (A and B). This is a testament to how small the stego noise truly is! No wonder that HUGO at 0.05bpp is difficult to detect, when its effect on features is so tiny, at least in this whitened projection. We can say that the calibration is providing an approximation to the cover, but not one good enough to distinguish cover and stego by distance alone.

The lower part of the same table shows the corresponding results for PSRM features, with LSBM payload at 0.01bpp. Again, the distance between consecutive captures of the same scene is much smaller than between independent covers, and steganography still causes smaller distortion than the difference between independent images, but in this case it is larger than the distance between a cover A and a completely-overlapping reference B.

Why, then, does calibration help? We come to a similar conclusion to Kodovsky,[11] that our type of calibration is enhancing the distinction between cover and stego rather than necessarily returning a good approximation to the cover. We can see that this is so by choosing a different distance measure: instead of normalized PCA, we take a single dimension which is the FLD projection best classifying features in $\mathbf{C}_A$ from $\mathbf{H}_A^{0.05}$ (also $\mathbf{L}_A^{0.01}$), and measure distance projected onto this direction. With respect to this metric, mean distances between covers and corresponding images in other sets are displayed in the second lines of Table 3. In this direction, steganography causes 3–4 times more distortion than the difference between independent covers, and 2.5–6 times more distortion than between overlapping captures of the same scene. Thus we can say that steganographic distortion is overall *smaller in magnitude* than between different captures of the same scene, but *lies in a consistent direction* and is thereby detectable.

5. CONCLUSION

Calibration in steganalysis has not advanced much since it was first proposed, and has been largely limited to JPEG images. In this paper we have proposed a novel type of calibration, using an image which has identifiable overlapping content, and shown that it improves the accuracy of detection using standard steganalysis features and classifiers under laboratory conditions.

We note some limitations of our experiments. We used a carefully controlled data set, which excluded some potential sources of real-world error: images with substantial saturated regions, incorrectly-identified amounts of overlap, and camera movement (which was excluded by use of a tripod). Even if all camera settings are equal,

the noise characteristics of an image from cameras moving at different speeds will have significant mismatch, which would probably throw off the calibration technique. A similar problem might arise for scenes where the lighting changes between overlapping images (if the sun goes behind a cloud, for example). We also note that our data set, with 2500 images obtained by slicing 500 scenes, is not large, although we did exclude under-training as the reason for calibration improvement.

We explored a range of calibration functions, but there is scope to look more widely, including at functions that do not simply apply pointwise or linear operations. Given a *set* of overlapping images, one could imagine calibration that measures some kind of whitened distance, although a lot of regularisation would be needed. Indeed, there might even be scope for learning the best calibration automatically. In our experiments, the exact calibration formula did not make a huge difference, and only $\kappa(x, y) = x - y$ showed promise for the case of reference images of unknown status.

We have restricted our experiments to uncompressed images and spatial-domain features, and in a sense our method complements the traditional calibrated steganalysis only available to JPEG images. There is scope for extending this work to JPEG images also, for which we can use the same image library. This is the subject of a research project by a current student supervised by the second author, and its preliminary results are a little different: it seems that calibration by overlapping JPEG images is approximately as good as traditional calibration of JPEGs (decompression, cropping, and recompression), but only in limited cases is it any better. For example, we tested JPEG versions of the images in sets A and B, compressed at quality factor 80, steganography using nsF5 at payload of 0.02 bits per nonzero coefficient, and JRM features.[21] Completely uncalibrated features give an error rate of 5.6%; cartesian calibrated features produced in the traditional manner give rise to an error rate of 4.9%, and cartesian calibration using 100% overlapping images instead gives an error rate of 4.7%. We might expect that quality factor 100 images, which are as close as JPEG can come to raw uncompressed images, would behave more like the raw images tested here, but they do not: perhaps this is due to the difference in the steganographic embedding method.

Calibration of overlapping JPEGs could be more sophisticated, because we can use *both* the overlapping reference image *and* the reference image obtained in the traditional way. It is unknown, at the moment, how best to combine this information with the suspected stego object.

Finally, in order to be practically useful, this work would need to be extended. Ideally we would like to use images from real-world sources such as social networks, automatically identifying overlapping images with equal camera settings (this should be straightforward) but modifying calibration so that it works with varying amounts of overlap and the absence of known reference images (we discussed possibilities for this in Subsect. 4.2). Whether the technique is valuable most likely depends on how often overlapping images are seen in the real world, which is another study in itself.

ACKNOWLEDGMENTS

Some of this work formed the first author's MSc in Computer Science dissertation at Oxford University in 2014.[20]

The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. They also thank Jan Kodovský for making available reference implementations of the SRM feature extractor and ensemble classifier.

REFERENCES

[1] Pevný, T., Filler, T., and Bas, P., "Using high-dimensional image models to perform highly undetectable steganography," in [*Proc. Information Hiding, 12th International Conference*], *LNCS* **6387**, 161–177, Springer (2010).

[2] Holub, V. and Fridrich, J., "Designing steganographic distortion using directional filters," (2012). Poster at IEEE Workshop on Information Forensics and Security, Tenerife.

[3] Holub, V., Fridrich, J., and Denemark, T., "Universal distortion function for steganography in an arbitrary domain," *EURASIP Journal on Information Security* **2014**(1) (2014).

[4] Pevný, T., Bas, P., and Fridrich, J., "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on Information Forensics and Security* **5**(2), 215–224 (2010).

[5] Fridrich, J. and Kodovský, J., "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security* **7**(3), 868–882 (2012).

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

[7] Ker, A. D., Bas, P., Böhme, R., Cogranne, R., Craver, S., Filler, S., Fridrich, J., and Pevný, T., "Moving steganography and steganalysis from the laboratory into the real world," in [*Proc. 1st ACM Workshop on Information Hiding and Multimedia Security*], 45–58, ACM (2013).

[8] Fridrich, J., Goljan, M., and Hogea, D., "Attacking the OutGuess," in [*Proc. 5th ACM Workshop on Multimedia & Security*], ACM (2002).

[9] Fridrich, J., "Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes," in [*Proc. 6th Information Hiding Workshop*], *LNCS* **3200**, 67–81, Springer (2004).

[10] Pevný, T. and Fridrich, J., "Merging Markov and DCT features for multi-class JPEG steganalysis," in [*Security, Steganography, and Watermarking of of Multimedia Contents IX*], *Proc. SPIE* **6505**, 0301–0314 (2007).

[11] Kodovský, J. and Fridrich, J., "Calibration revisited," in [*Proc. 11th ACM Workshop on Multimedia & Security*], 63–74, ACM (2009).

[12] Kodovský, J. and Fridrich, J., "Steganalysis of JPEG images using rich models," in [*Media Watermarking, Security, and Forensics XIV*], *Proc. SPIE* **8303**, 0A01–0A13 (2012).

[13] Ker, A. D., "Resampling and the detection of LSB matching in colour bitmaps," in [*Security, Steganography, and Watermarking of of Multimedia Contents VII*], *Proc. SPIE* **5681**, 1–15, SPIE (2005).

[14] Ker, A. D., "Steganalysis of LSB matching in grayscale images," *IEEE Signal Processing Letters* **12**(6), 441–444 (2005).

[15] Denemark, T., Fridrich, J., and Holub, V., "Further study on the security of S-UNIWARD," in [*Media Watermarking, Security, and Forensics 2014*], *Proc. SPIE* **9028**, 1601–1615, SPIE (2014).

[16] Carnein, M., Schöttle, P., and Böhme, R., "Predictable rain?: Steganalysis of public-key steganography using wet paper codes," in [*Proc. 2nd ACM Workshop on Information Hiding & Multimedia Security*], 97–108, ACM (2014).

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

[18] Ker, A. D., Pevný, T., Kodovský, J., and Fridrich, J., "The square root law of steganographic capacity," in [*Proc. 10th ACM Workshop on Multimedia & Security*], 107–116, ACM (2008).

[19] Ker, A. D., "Implementing the projected spatial rich features on a GPU," in [*Media Watermarking, Security, and Forensics 2014*], *Proc. SPIE* **9028**, 1801–1810, SPIE (2014).

[20] Whitaker, J. M., "Steganalysis in overlapping images," (2014). MSc dissertation, Oxford University.

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