EFFECT OF SATURATED PIXELS ON SECURITY OF STEGANOGRAPHIC SCHEMES FOR DIGITAL IMAGES

Vahid Sedighi and Jessica Fridrich

Binghamton University
Department of ECE
Binghamton, NY

ABSTRACT

When hiding messages in digital images, care needs to be exercised how the embedding changes are executed in or near saturated pixels. In this paper, we consider three different rules that are currently being used that adjust the embedding in saturated pixels and assess their impact on empirical steganographic security of four modern embedding algorithms. Surprisingly, the rules can have a major effect, especially in image sources with stronger noise. We show that the preferred way to treat saturated patches during message hiding is to adjust the pixel costs to entirely avoid making embedding changes in saturated pixels despite the ensuing loss of embedding capacity. This paper hopes to raise the awareness of the importance of treatment of saturated pixels in steganography to avoid introducing easily correctable flaws that may negatively affect security.

***Index Terms*—** Steganography, saturation, overexposure, steganalysis, content adaptive

1. INTRODUCTION

Steganography is the art of hiding information in objects so that the very presence of hidden data is not obvious and cannot be proved using statistical hypothesis testing. Steganalysis, on the other hand, strives to discover the presence of embedded secrets. Currently, the vast majority of work on steganography in digital media has focused on imagery [1, 2, 3].

Modern embedding algorithms first assign costs of changing individual pixels based on their local neighborhood and then hide the payload using coding techniques with (near) minimal total embedding cost [4]. Pixels in regions that are easily modelable, such as blue sky, are typically assigned larger costs while pixels in textured regions have smaller costs. Most embedding algorithms use ternary coding and modify pixels by at most ± 1. For pixel values at the boundary of the dynamic range, there are essentially three options: 1) either map out-of-range values back to the original range, 2) restrict the polarity of changes, or 3) avoid making changes altogether. Because the number of saturated pixels in standard image sets is typically very low (i.e., on average less than 1% in BOSSbase 1.01 [5]), not much attention has been paid to the treatment of saturated pixels within the embedding algorithm. For example, the embedding simulators for WOW [6] and S-UNIWARD [7] adopt rule 2), which is restricting the polarity of embedding changes. As shown in this paper, this is not a very good option as it can increase the detection accuracy of current detectors by $1\% - 20\%$ depending on the cover source and embedding algorithm.

Effect of saturated pixels on accuracy of steganalytic detectors is a topic that has been studied before. In [8], the authors describe a correction to the so-called Weighted-Stego image quantitative detector to return more accurate estimates of embedded payload size in images containing saturated pixels. The effect of the relative number of saturated pixels on the error distribution of quantitative detectors has been analyzed in [9, 2]. These contributions focus on the effect of saturation on steganalysis and do not investigate how the embedding algorithm itself should be adjusted for better empirical security.

In this paper, we analyze all three above-mentioned strategies for treating pixels at the boundary of the dynamic range for four modern content-adaptive embedding algorithms and four different image sources commonly used for benchmarking in steganography and digital forensics. The drop in security associated with rule 1) and 2) is especially pronounced in images with stronger noise and not necessarily an elevated number of saturated pixels. We show that the security is undermined due to the fact that the embedding algorithms introduce changes near the boundary of saturated patches. The most conservative embedding rule 3), that avoids making changes in saturated pixels, is the most secure option for the steganographer and should always be adopted.

In Section 2, we introduce the basic concepts and necessary

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

background. Three different strategies for assigning costs to pixels at the boundary of the dynamic range are listed in Section 3. In Section 4, we summarize the common core of all experiments. Section 5 contains the results of all experiments and their interpretation. The paper is concluded in Section 6.

2. BACKGROUND

In this paper, we work with 8-bit grayscale images whose pixel values are denoted with $\mathbf{X}$ and $\mathbf{Y}$ for cover and stego images, respectively, both $n_1 \times n_2$ matrices with elements $x_{ij}, y_{ij} \in \{0, \dots, 255\}$.

Currently, all content-adaptive algorithms use a rule that specifies the cost of making an embedding change by $+1$ and -1 at each pixel: $\rho_{ij}^{(+)} \geq 0$ and $\rho_{ij}^{(-)} \geq 0$. The embedding algorithm hides a given secret payload while minimizing the expected distortion between cover and stego images computed as a sum of costs of all changed pixels:

$$D(\mathbf{X}, \mathbf{Y}) = \sum_{x_{ij} \neq y_{ij}} \rho_{ij}^{(y_{ij} - x_{ij})}. \tag{1}$$

It can be easily shown [4] that such optimal embedding will execute changes with probabilities:

$$\Pr\{y_{ij} = x_{ij} \pm 1\} = \frac{e^{-\lambda \rho_{ij}^{(\pm)}}}{1 + e^{-\lambda \rho_{ij}^{(\pm)}} + e^{-\lambda \rho_{ij}^{(\mp)}}} \triangleq \beta_{ij}^{(\pm)} \tag{2}$$

$$\Pr\{y_{ij} = x_{ij}\} = 1 - \beta_{ij}^{(+)} - \beta_{ij}^{(-)}. \tag{3}$$

and thus embed a payload of $H_3(\beta_{ij}^{(+)}, \beta_{ij}^{(-)})$ bits, where $H_3(u, v) \triangleq -u \log_2 u - v \log_2 v - (1-u-v) \log_2 (1-u-v)$ is the ternary entropy function.

In most academic work, researchers study embedding simulators that merely execute the changes with probabilities (2)–(3) but do not perform actual embedding. A practical embedding scheme that operates near the corresponding payload-distortion bound can be built using syndrome-trellis codes [4].

3. BOUNDARY RULES

All state-of-the-art embedding schemes for the spatial domain that minimize additive distortion (1) use symmetric costs, $\rho_{ij}^{(+)} = \rho_{ij}^{(-)}$. However, when a cover pixel has a borderline value, $x_{ij} = 0$ or 255, the embedding obviously needs to be modified so that the stego image stays within the dynamic range. This can be assured in at least three different ways.

Embed and correct. Here, the sender embeds while making out-of-range changes and then adjusts the values in the stego image $\mathbf{Y}$ to comply with the dynamic range. Since most embedding algorithms use ternary coding, $y_{ij} = -1$ is changed to $y_{ij} = 2$ and $y_{ij} = 256$ to changed to $y_{ij} = 253$. This measure does not decrease the embedding capacity of an image at the expense of increased distortion.

$$\begin{aligned} &\rho_{ij}^{(-)} = \rho_{ij}^{(+)} \implies \text{Embedding} \\ \implies &\begin{cases} y_{ij} = -1 & \implies y_{ij} = 2 \\ y_{ij} = 256 & \implies y_{ij} = 253. \end{cases} \end{aligned} \tag{4}$$

Forbid changes outside range. For pixel values at the boundary of the dynamic range, the cost of changing the pixel value outside of the dynamic range is set to a very large value C. This rule decreases the embedding capacity of an image.

$$\begin{aligned} &\begin{cases} x_{ij} = 0 & \implies \rho_{ij}^{(-)} = C \\ x_{ij} = 255 & \implies \rho_{ij}^{(+)} = C \end{cases} \\ &\qquad \implies \text{Embedding}. \end{aligned} \tag{5}$$

Avoid saturated pixels altogether. This is the most conservative rule because it prohibits the embedding from changing the borderline values altogether.

$$\begin{aligned} &(x_{ij} = 0 \text{ or } x_{ij} = 255) \\ \implies &\rho_{ij}^{(+)} = \rho_{ij}^{(-)} = C \\ \implies &\text{Embedding}. \end{aligned} \tag{6}$$

In Section 5, these three rules are subjected to tests on four modern embedding algorithms and four image sources.

4. SETUP OF EXPERIMENTS

In this section, we list the common core of all experiments in this paper: the image sources, steganographic methods, and steganalysis methodology.

4.1. Image sources

BOSSbase 1.01 [5] is by far the most frequently used database for designing steganography and benchmarking. It contains 10,000 images taken in the RAW format by seven different cameras, converted to grayscale, downsampled and cropped to the final size of 512×512 pixels. The script used for the conversion and processing is also available from the same web site as the database itself.[10]

BOSSbaseNRC (Non-interpolated Red Channel) was formed from the same RAW images as BOSSbase 1.01 but with the color-interpolation step as well as the resizing skipped. Instead, the images were subsampled by a factor of 2 to form two-times smaller images only from pixels with a red color filter array sampled at 8 bits. The processing did involve gain and gamma adjustment to obtain a naturally looking content. Even though this is an unusual source, it is not completely artificial. Hiding in images from BOSSbaseNRC is rather close to hiding in the RAW format, which is increasingly being used even by casual photographers. This source was included intentionally because the effect of Rules 1–3 is most pronounced on this source.

Source	N	S	$S_{3\times3}$	$\overline{S}_{3\times3}$	B	$B_{3\times3}$	$\overline{B}_{3\times3}$	$F_{3\times3}$	$\overline{F}_{3\times3}$
BOSSbase 1.01	10,000	0.0097	0.0084	0.0093	0.0026	0.0016	0.0020	0.0616	0.1007
BOSSbaseNRC	10,000	0.0221	0.0171	0.0208	0.0133	0.0016	0.0021	0.0188	0.0230
NRCS-C	6,644	0.0037	0.0024	0.0030	0.0009	0.0000	0.0001	0.0074	0.0123
NikonD90	2,276	0.0110	0.0101	0.0166	0.0058	0.0023	0.0031	0.0570	0.0802

Table 1. The total number of images in each source, N, and the relative number of saturated pixels, S, black pixels, B, pixels from a union of 3×3 blocks of saturated, black, and flat pixels, $S_{3\times3}$, $B_{3\times3}$, $F_{3\times3}$, and their versions dilated by a 3×3 neighborhood, $\overline{S}_{3\times3}$, $\overline{B}_{3\times3}$, $\overline{F}_{3\times3}$.

NRCS-C was derived from the NRCS database of 3,322 raw scans of negatives coming from the USDA Natural Resources Conservation Service [11]. Two 512×512 images were obtained by cropping the central 512×1024 part of each NRCS image, splitting it in two, and converting each image to grayscale. Thus, the NRCS-C image set contains a total of $2 \times 3,322$ = 6,644 images.

NikonD90 is a subset of RAISE dataset [12] taken with Nikon D90 camera. RAISE dataset is commonly used for benchmarking digital forensic algorithms. We downloaded the version that is part of LIRMM [13] in which the images were converted to grayscale and cropped to 512×512. It contains 2,276 images.

Table 1 shows the average relative number of saturated, black, and flat pixels across the four sources. Also shown are the statistics for pixels in the form of a union of 3×3 squares of saturated, black, and flat pixels and their dilated versions (e.g. '$\overline{S}_{3\times3} = \mathrm{imdilate}(S_{3\times3}, \mathrm{ones}(3,3))$' in Matlab) to obtain a better picture about the spatial distribution of such pixels (i.e., whether they are scattered or form connected segments). BOSSbaseNRC contains the largest number of saturated pixels (2.2% on average). By comparing the counts S, $S_{3\times3}$, and $\overline{S}_{3\times3}$, one can see that saturated pixels also form connected regions in all four sources. On the other hand, the number of black pixels is comparatively much smaller and the pixels are more scattered across the images. The number of flat pixels $F_{3\times3}$ and $\overline{F}_{3\times3}$ show that BOSSbaseNRC and NRCS-C are generally much noisier than BOSSbase 1.01 and NikonD90.

Four spatial-domain content-adaptive embedding algorithms are investigated in this paper: WOW [6], S-UNIWARD [7], HILL [14], and MiPOD [15]. These four algorithms represent current state of the art in steganography built around additive distortion functions (1).

Security is evaluated experimentally by training the FLD ensemble [16] for the classes of cover and stego images embedded with a fixed relative payload in bits per pixel (bpp). The security is reported with $\overline{P}_{\mathrm{E}}$, the minimal total error probability under equal priors, $P_{\mathrm{E}} = \frac{1}{2}(P_{\mathrm{FA}} + P_{\mathrm{MD}})$,[1] on the testing set averaged over ten 50/50 database splits into training and testing sets. The selection-channel-aware spatial rich model, the maxSRMd2 [17], was used in all experiments.

[1] P_{FA} and P_{MD} are the false-alarm and missed-detection rates.

Saturation Type	WOW	S-UNIWARD	HILL	MiPOD
BSR	20.73	14.95	10.06	11.46
MSR	2.54	2.42	1.28	1.48

Table 2. Percentage of changed saturated pixels on the boundary of saturated regions (BSR) and in the middle of the saturated regions (MSR) as the result of embedding with Rule 1 in NikonD90 at payload 0.4 bpp. A pixel in BSR has at least one non-saturated pixel in its 3×3 neighborhood, while all 8 pixels in the 3×3 neighborhood of a pixel from the MSR are saturated.

5. EXPERIMENTS

Figure 1 shows the average detection error $\overline{P}_{\mathrm{E}}$ for four embedding algorithms, four image sources, and three boundary rules (Section 3) for one small and one large payload. First, we wish to point out the results on BOSSbaseNRC. While all four embedding algorithms are practically undetectable with Rule 3 ($\overline{P}_{\mathrm{E}} \approx 0.5$), with Rule 1 and 2 the detectability increases by up to 20% in terms of the detection error. The high noise level in BOSSbaseNRC images essentially prevents detection of steganography everywhere except for the boundaries of saturated regions where Rules 1 and 2 allow changes (c.f. Table 2). Avoiding the saturated regions altogether (Rule 3) removes this flaw. The effect of Rules 1–3 on security is also apparent in the other three sources but is less pronounced, especially for the smaller payload as all four algorithms strive to avoid making embedding changes in saturated pixels. The smallest impact is observed for BOSSbase 1.01.

As Table 2 shows, under Rule 1 most changes in saturated pixels are near the boundary and not in the middle of saturated patches as all embedding schemes tend to avoid saturated pixels because of their higher embedding costs ρ_{ij}. However, because the costs are determined from a neighborhood of each pixel, borders of saturated patches are still changed during embedding. Note that HILL and MiPOD generally experience a smaller drop in security with Rules 1 and 2 (Figure 1), which is because they make fewer changes in the borders of saturated patches (Table 2) since their costs are postprocessed

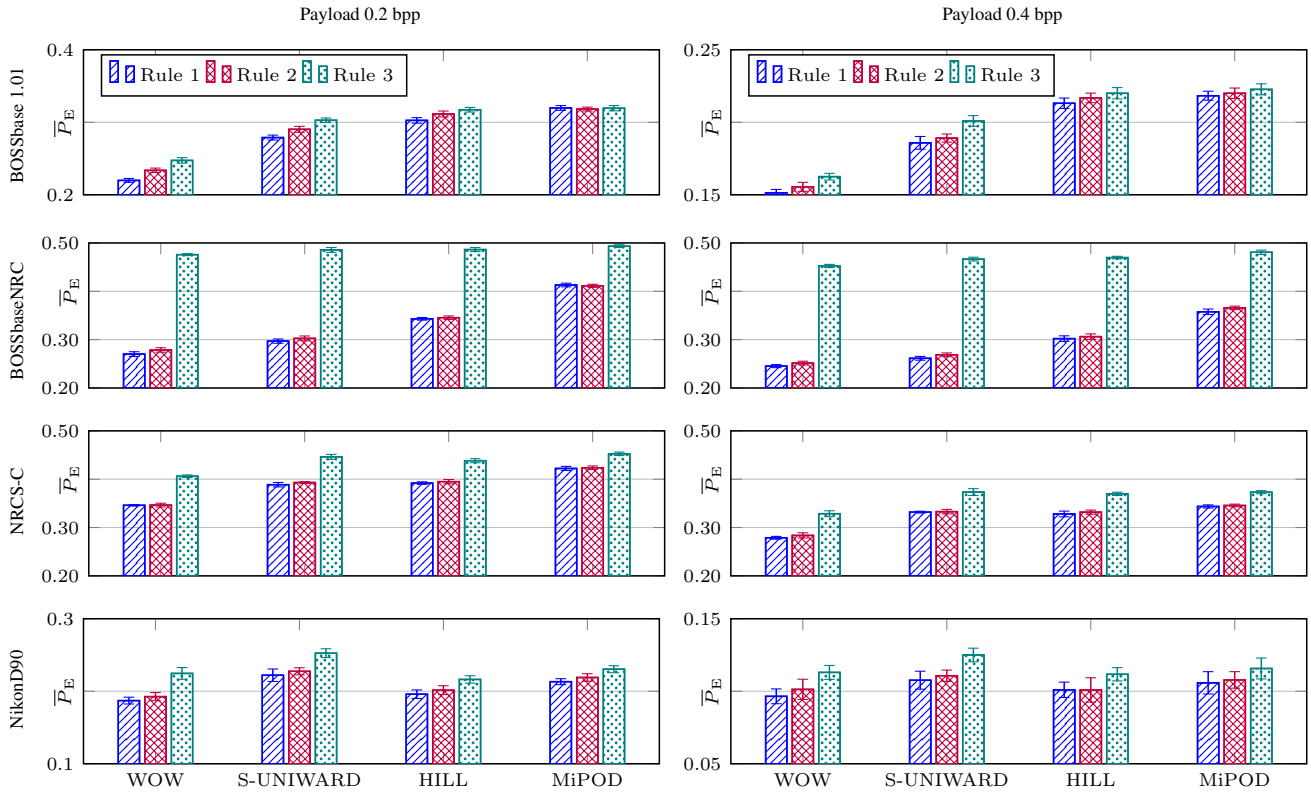

Fig. 1. Average detection error $\overline{P}_E$ for four embedding algorithms and boundary Rules 1–3 for four image sources, BOSSbase 1.01, BOSSbaseNRC, NRCS-C, and NikonD90 (by rows), and two payloads, 0.2 and 0.4 bpp (by columns). Note the different range of y-axis to save space.

using an averaging filter that "spills" high embedding costs into their neighborhood.

Furthermore, we would like to stress that the increased detectability of Rule 1 and 2 is almost entirely due to saturated pixels (value 255) rather than black pixels with value 0. This is because saturation occurs even in very noisy images where "underflow" is unlikely due to the noise. Black pixels are also much more scattered and do not form connected regions as can be seen from Table 1 and the discussion in Section 3. To confirm this claim, we provide a limited scale experiment with S-UNIWARD at 0.4 bpp. In Table 3, we show the detection error for Rules 1–3 and a modification of Rule 2 that treats black pixels as Rule 3 (avoids changing them) but applies Rule 2 for saturated pixels (allows changes by -1). The fact that the results obtained with this Rule 2 NE0 do not statistically deviate from Rule 2 confirms the more important role of saturated pixels.

	Rule 1	Rule 2	Rule 3	Rule 2 NE0
NikonD90	.1065±.0038	.1131±.0056	.1241±.0033	.1134±.0037
NRCS-C	.3195±.0028	.3337±.0030	.3740±.0035	.3387±.0042

Table 3. Detection error for Rule 1–3 and Rule 2 NE0 that avoids embedding in black pixels but applies Rule 2 in saturated pixels.

6. CONCLUSIONS

Content-adaptive steganography is nowadays a mature area of research. Little attention, however, has been paid to the proper adjustment of the embedding algorithm at pixels with values at the boundary of the dynamic range. We hypothesize that this lack of interest may be due to the incorrect belief that saturated pixels are rare and their effect on detectability in current steganalysis is negligible.

In this paper, we investigate three different rules for treatment of such pixel values that sound plausible: 1) allow changes by ± 1 everywhere and then correct for the finite dynamic range, 2) allow only one-sided changes at boundary values, 3) prohibit changes of boundary values entirely. On experiments with four modern steganographic schemes and four image sources, we show that the impact of the above rules on detectability can be substantial, increasing the detection accuracy of classical steganalysis with rich image models by 1%–20%, depending on the embedding algorithm and image source. The most conservative Rule 3 leads to the best empirical security and should be used in practice.

7. REFERENCES

[1] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*, Cambridge University Press, 2009.

[2] R. Böhme, *Advanced Statistical Steganalysis*, Springer-Verlag, Berlin Heidelberg, 2010.

[3] S. Katzenbeisser and F. Petitcolas, *Information Hiding:*, Artech House Publishers, 2016.

[4] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 920–935, September 2011.

[5] P. Bas, T. Filler, and T. Pevný, "Break our steganographic system – the ins and outs of organizing BOSS," in *Information Hiding, 13th International Conference*, T. Filler, T. Pevný, A. Ker, and S. Craver, Eds., Prague, Czech Republic, May 18–20, 2011, vol. 6958 of Lecture Notes in Computer Science, pp. 59–70.

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

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

[8] A. D. Ker and R. Böhme, "Revisiting weighted stego-image steganalysis," in *Proceedings SPIE, Electronic Imaging, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X*, E. J. Delp, P. W. Wong, J. Dittmann, and N. D. Memon, Eds., San Jose, CA, January 27–31, 2008, vol. 6819, pp. 5 1–17.

[9] R. Böhme, "Assessment of steganalytic methods using multiple regression models," in *Information Hiding, 7th International Workshop*, M. Barni, J. Herrera, S. Katzenbeisser, and F. Pérez-González, Eds., Barcelona, Spain, June 6–8, 2005, pp. 278–295.

[10] "BOSSbase 1.01," `http://agents.fel.cvut.cz/stegodata/`.

[11] "NRCS database," `http://photogallery.nrcs.usda.gov`.

[12] "RAISE database," `http://mmlab.science.unitn.it/RAISE/`.

[13] "LIRMM database," `http://www.lirmm.fr/~chaumont/LIRMMBase.html`.

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

[15] 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.

[16] 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.

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