STEGANOGRAPHY AND ITS DETECTION IN JPEG IMAGES OBTAINED WITH THE “TRUNC” QUANTIZER

Jan Butora and Jessica Fridrich, Fellow, IEEE

Binghamton University
Department of ECE
Binghamton, NY

ABSTRACT

Many portable imaging devices use the operation of “trunc” (rounding towards zero) instead of rounding as the final quantizer for computing DCT coefficients during JPEG compression. We show that this has rather profound consequences for steganography and its detection. In particular, side-informed steganography needs to be redesigned due to the different nature of the rounding error. The steganographic algorithm J-UNIWARD becomes vulnerable to steganalysis with the JPEG rich model and needs to be adjusted for this source. Steganalysis detectors need to be retrained since a steganalyst unaware of the existence of the trunc quantizer will experience 100% false alarm.

Index Terms— Steganography, side-information, trunc quantizer, steganalysis, JPEG

1. INTRODUCTION

Steganography in JPEG images is usually executed by partially decompressing the JPEG file and modifying the quantized DCT coefficients by at most ±1. To the best of the authors’ knowledge, the entire bulk of previous art on JPEG steganography assumes that the last step of JPEG compression involves rounding the DCT coefficients to the nearest integer [16]. Such JPEG images will be referred to as coming from the *round source.* As recently pointed out in [1], however, many modern portable imaging devices, such as iPhone 5c, Canon EOS 10D, Samsung Galaxy Tab 3 8.0, replace the rounding with “rounding towards zero” due to its easier (more efficient) hardware implementation. We will refer to JPEG images processed this way as coming from the *trunc source.*

This paper studies both steganography and steganalysis in trunc JPEGs. In the next section, we introduce notation, datasets, and the setup of experiments. In Section 3, we show that a steganalyst unaware of the existence of the trunc quantizer will experience 100% false alarm rate independently of the steganography and the detector. We also show that steganography in trunc JPEGs is more secure. In Sections 4 and 5, J-UNIWARD and SI-UNIWARD stego algorithms are redesigned to reflect the specifics of the new source. The paper is concluded in Section 6.

2. PRELIMINARIES

For simplicity, we only work with 8-bit grayscale images. Pixel values and unquantized DCT coefficients in an 8×8 block will be denoted x_{ij} and d_{kl}, $0 \leq i, j, k, l \leq 7$. The classical rounding operation will be denoted $\mathcal{Q}_{round}(x) = [x]$ while the trunc quantizer is $\mathcal{Q}_{trunc}(x) = \lfloor x \rfloor$ for $x \geq 0$ and $\mathcal{Q}_{trunc}(x) = \lceil x \rceil$ for $x < 0$, where $\lfloor x \rfloor$ and $\lceil x \rceil$ represent flooring and ceiling. Quantized DCT coefficients are $\mathcal{Q}_{(\cdot)}(d_{kl}/q_{kl})$, where q_{kl} are the luminance quantization steps. The rounding error during compression is defined as $e_{kl} = c_{kl} - \mathcal{Q}_{(\cdot)}(c_{kl})$, where we denoted $c_{kl} = d_{kl}/q_{kl}$.

All experiments are carried out on the union of BOSSbase 1.01 and BOWS2 datasets, each with 10,000 grayscale images, resized from their original size 512×512 to 256×256 using `imresize` with default setting in Matlab. Cover JPEG images coming from the round source were obtained in Matlab using the command `imwrite`. Cover JPEG images from the trunc source were obtained in Matlab by applying Matlab’s `dct2` on blocks of pixels, dividing the coefficients by the quantization matrix, applying the trunc quantizer $\mathcal{Q}_{trunc}(x)$, and saving them to a JPEG file using Phil Sallee’s `jpeg_write`. Decompression to the spatial domain for experiments with empirical detectors was obtained by multiplying the DCT coefficients by quantization steps and applying a block inverse DCT without rounding or clipping in Python by applying ’fftpack.idct’ with the parameter norm = ’ortho’, from

The work on this paper was partially supported by NSF grant No. 1561446 and by DARPA under agreement number FA8750-16-2-0173. 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.

Python's SciPy library, horizontally and vertically.

For training empirical detectors, we randomly selected 4,000 images from BOSSbase and the entire BOWS2 dataset with 1,000 BOSSbase images set aside for validation. The remaining 5,000 BOSSbase images were used for testing. In summary, $2 \times 14,000$ cover and stego images were used for training, $2 \times 1,000$ for validation, and $2 \times 5,000$ for testing. This dataset and the split into training and testing has also been used in [18, 19, 2].

For steganalysis, we selected the SRNet [2], the cartesian-calibrated JPEG Rich Model (ccJRM) [14], and Gabor Filter Residual features (GFR) [17] with the FLD ensemble [15]. The ensemble was trained on the union of the training and validation sets. For training SRNet from scratch, we set the initial learning rate (LR) to 10^{-3} for 400k iterations and continued for 100k more iterations with LR 10^{-4} and batch size 32. When seeding, we use LR 10^{-3} for 100k iterations and a lower the LR to 10^{-4} for additional 50k iterations.

3. COMPARING THE SOURCES

For experiments in this section, we selected the steganographic algorithm nsF5 [7] with relative payload 0.2 bpnzac, J-UNIWARD [10] with payload 0.4 bpnzac, and UED [8, 9] with payload 0.3 bpnzac.

3.1. Quantizer mismatch

First, we study what happens when the detector is unaware of the existence of the trunc source and uses a detector trained on the round source for steganalysis of trunc JPEGs. Experiments were executed with three different detectors for quality factors 85 and 100 and various steganographic algorithms and payloads. To be more specific, we trained a classifier for a given stego algorithm (and fixed payload) on cover and stego images from the round source. This detector was then tested on cover and stego JPEGs embedded with the same stego algorithm and payload but starting with trunc JPEG covers instead. The end result was always the same – both cover and stego images from the trunc source were detected as stego irrespectively of the embedding algorithm, payload and detector, with the false alarm rate ranging between 99.1% and 100%.

Fortunately, it is easy to reliably identify the type of the DCT quantizer and build separate detectors for each source. Table 1 shows the accuracy of a classifier trained on two classes: cover JPEG images coming from the trunc and round source for quality 85 and 100. The training was stopped after 70k iterations, since the validation accuracy already saturated at 100%. Note that this detector correctly reveals the DCT quantizer even when presented with stego images embedded with various payloads and different stego algorithms. Having this classifier, from now on we will assume that the steganalyst knows whether an image under investigation comes from the round or the trunc source.

Algorithm	Payload	QF85	QF100
Covers	0	0.9999	0.9989
nsF5	0.2	0.9998	0.9987
JUNI	0.4	0.9999	0.9987
UED	0.3	0.9997	0.9987

Table 1. Accuracy of detecting the DCT quantizer. The detector is the SRNet trained between cover classes from the round and trunc sources and tested on 5,000 pairs of images from each of the four sources.

3.2. Effect of truncation on security

Since the histogram bin for zero coefficients in the trunc source is twice as wide as all other bins, cover images in trunc source have more zeros than covers in the round source. For a fixed image, its "effective" size, the number of non-zero DCT coefficients [13], is smaller in the trunc source than in the round source. For a fair comparison of the security of a given stego algorithm in both sources, we thus adjust the size of the embedded payload according to the square root law [13, 6, 12, 11]. The relative payload in the trunc source, α_{trunc}, was scaled as

$$\alpha_{trunc} = \alpha_{round} \cdot \sqrt{\frac{N_{round}}{N_{trunc}}} \cdot \frac{\log(N_{trunc})}{\log(N_{round})}, \qquad (1)$$

where N_{trunc} and N_{round} stand for the number of non-zero AC DCT coefficients from a given image in trunc and round sources, respectively. The accuracy[1] shown in Figure 1 were obtained with three different detectors: SRNet and the ensemble classifier with JRM and GFR features on the same embedding algorithms and payloads as above. SRNets on quality 75 were trained from scratch, while 95 was trained via curriculum training from 75. For nsF5, the network was first trained on quality 95 from scratch and then retrained for the smaller quality because the higher quality is more detectable [3]. Note that even with the scaled payload, the detection accuracy is larger in the round source across all algorithms and detectors, indicating that it is harder to detect steganography in the trunc source. A *surprising exception* is J-UNIWARD, which is best detected in trunc source with JRM. As shown in the next section, this is because J-UNIWARD embeds "too much" into zero DCT coefficients, which are much more populated

[1] Accuracy for the ensemble with rich models is computed as $1 - P_{\mathrm{E}}$, where P_{E} is the minimum average total probability of error.

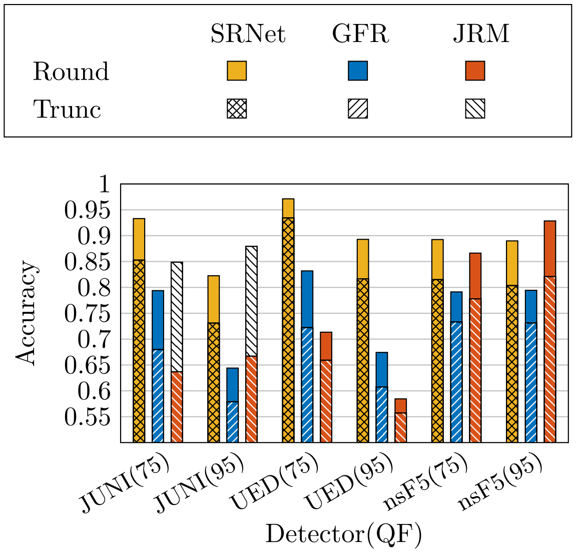

Fig. 1. Detection accuracy in the trunc source and the round source when adjusting for the square root law for J-UNIWARD, UED, and nsF5 with relative payloads 0.4, 0.3, and 0.2 bpnzac.

in the trunc source, and consequently introduces artifacts detectable by JRM.

4. J-UNIWARD FOR TRUNC SOURCE

As mentioned in the previous section, J-UNIWARD in the trunc source is best detected with the JRM because it embeds too much into zero coefficients. Figure 2 top left shows that stego images have significantly fewer zero coefficients than cover images. This lead us to the following adjustment of the embedding algorithm.

For a fixed DCT mode (k, l), let β_i be the average J-UNIWARD change rate on such coefficients that are equal to i in the cover image. If there are no coefficients equal to i, we set $\beta_i = 0$. Let ρ_i be the corresponding "average cost"

$$\rho_i = 1/\lambda \log(1/\beta_i - 2), \tag{2}$$

where $\lambda > 0$, is a Lagrange multiplier. We wish to adjust $\rho_0 \to \tilde{\rho}_0 = \eta\rho_0$, $\eta > 0$, so that the new change rate of zeros

$$\tilde{\beta}_0 = \frac{e^{-\lambda\rho_0\eta}}{1 + 2e^{-\lambda\rho_0\eta}} \tag{3}$$

preserves on average the number of zero coefficients:

$$(1 - 2\tilde{\beta}_0)h[0] + \beta_{-1}h[-1] + \beta_1 h[1] = h[0], \tag{4}$$

where $h[i]$ is number of cover coefficients equal to i. Assuming $\beta_{-1} = \beta_1$ and using $\log(1 + z) \approx z$ for small $z > 0$, (3) and (4) give

$$\eta = \frac{\rho_1}{\rho_0} + \frac{1}{\lambda\rho_0}\log\left(\frac{2h[0]}{h[1] + h[-1]}\right). \tag{5}$$

Computing the average change rate on coefficients equal to 1 or -1, $\beta_{|1|} = (\beta_{-1} + \beta_1)/2$, from (2) and (5)

$$\eta = \frac{\log(1/\beta_{|1|} - 2)}{\log(1/\beta_0 - 2)} + \frac{\log\left(\frac{2h[0]}{h[1]+h[-1]}\right)}{\log(1/\beta_0 - 2)}. \tag{6}$$

Technically, the change rates in (4) have a different Lagrange multiplier because, first, J-UNIWARD simulator is used to compute the average change rates β_i, the costs of zero coefficients are then updated, and a new Lagrange multiplier needs to be found to satisfy the payload constraint. As indicated by our experiments, however, the new Lagrange multiplier produced by such modulation of costs on zero coefficients is almost identical to the original one, justifying thus our simplified approach. Note that, if there are no coefficients equal to 1 or -1, the update rule (5) naturally sets the costs to wet costs. We call this scheme J-UNIWARD with histogram correction (hcJ-UNIWARD). Figure 2 top right shows that the embedding indeed roughly preserves the number of zero coefficients.

To show that hcJ-UNIWARD is more secure than J-UNIWARD across quality factors, we trained the SRNet, the ensemble classifier with JRM features, as well as the concatenation of JRM and the features extracted by the SRNet (the 512-dimensional input to the IP layer) with the low-complexity linear classifier [4]. The improvement in security ranges from $7 - 15\%$ in terms of accuracy of the best detector among the three detectors mentioned above (see Figure 3).

5. SIDE INFORMATION

In side-informed JPEG steganography, the rounding errors during the quantization of DCT coefficients are used to modulate the embedding costs by $1 - 2|e_{kl}|$. In trunc source, however, the rounding errors have a different range, and the modulation has to be adjusted. Note that a modulation by $1 - 2|e_{kl}|$ would lead to negative costs. Moreover, it does not correspond to what one would intuitively expect because zero cost should be associated with $e_{kl} \approx 0$ and $|e_{kl}| \approx 1$. In this section, we focus on the ternary version of SI-UNIWARD [10, 5].

We propose to modulate by the minimum perturbation of the precover that makes it quantize to the desired stego value. Denoting $\rho_{kl}(-1)$, $\rho_{kl}(+1)$ J-UNIWARD's costs of changing the kl-th DCT coefficient by -1 and $+1$, respectively, the side-information modulated costs ρ'_{kl} for cover DCT coefficients c_{kl} that quantize to a non-zero integer ($|c_{kl}| \geq 1$)

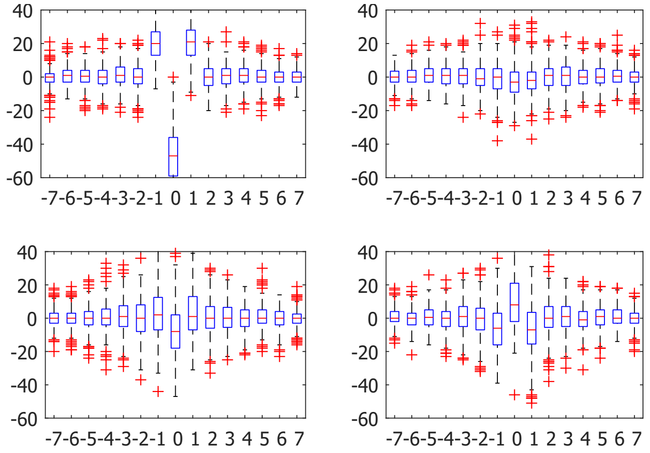

Fig. 2. Boxplots showing the differences between stego (0.4 bpnzac) and cover histograms of DCTs across 300 randomly selected images. From left to right by rows: J-UNIWARD, hcJ-UNIWARD, SI-UNIWARD, hcSI-UNIWARD.

$$\rho'_{kl}(\mathrm{sign}(e_{kl})) = (1-|e_{kl}|)\rho_{kl} \tag{7}$$

$$\rho'_{kl}(-\mathrm{sign}(e_{kl})) = |e_{kl}|\rho_{kl} \tag{8}$$

and for those that quantize to 0 ($|c_{kl}| < 1$)

$$\rho'_{kl}(+1) = (1-e_{kl})\rho_{kl} \tag{9}$$

$$\rho'_{kl}(-1) = (1+e_{kl})\rho_{kl}. \tag{10}$$

This makes good intuitive sense because when either $e_{kl} \approx 0$ or $|e_{kl}| \approx 1$ the non-quantized coefficient c_{kl} is the most sensitive to noise and should be given a small cost (modulation close to zero). The separate treatment for coefficients that quantize to zero is necessary because the quantization bin for zero is twice as large in the trunc source. Indeed, when $0 < c_{kl} < 1$, $e_{kl} = c_{kl}$, and it takes a perturbation of $1 - e_{kl}$ to quantize to 1 and $1 + e_{kl}$ to quantize to -1.

SI-UNIWARD was implemented and tested in the trunc source for quality factors 75 and 95 at $1, 0.8, 0.6$, and 0.4 bpnzac. Starting with the largest payload, curriculum learning was used to train on the next smaller payload. Detection accuracy of the SRNet is shown in Table 2. For the smallest tested payload, the algorithm is practically undetectable, which validates the proposed modulation of costs. Only SRNet's accuracy is shown because the detection power of the ensemble classifier with JRM features was substantially worse.

We also implemented SI-UNIWARD with histogram correction (hcSI-UNIWARD) in the same way we implemented hcJ-UNIWARD, only this time, the modulation factor (6) was computed with SI-UNIWARD's average change rates. This, however, decreased the security by about 4%. We hypothesize that the modulation of costs of zeros in SI-UNIWARD (9)–(10) already addresses the problem with embedding into zeros too much because the total cost of changing a zero is $\rho'_{kl}(+1)+\rho'_{kl}(-1) = 2\rho_{kl}$, while for coefficients that quantize to a non-zero value, this sum is ρ_{kl}. This is supported by the box plots in Figure 2 bottom.

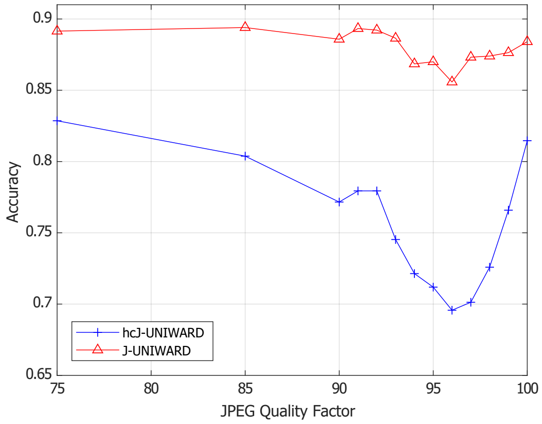

Fig. 3. Accuracy of the best detector in trunc source for hcJ-UNIWARD and J-UNIWARD at 0.4 bpnzac.

bpnzac	1	0.8	0.6	0.4
QF75	0.8164	0.7436	0.6485	0.5653
QF95	0.7984	0.6972	0.6050	0.5420

Table 2. Detection accuracy of SRNet for various payloads of ternary SI-UNIWARD in the trunc source.

6. CONCLUSIONS

JPEG compressors that use rounding towards zero (trunc) instead of rounding are common in portable electronic devices. This quantizer has profound implications for steganography. Steganalyst unaware of the existence of such a source will experience 100% false alarms. The “trunc JPEGs” are more friendly to steganography than “round JPEGs” even when adjusting the payload according to the square root law. Moreover, and surprisingly, J-UNIWARD's embedding is faulty in trunc JPEGs as it embeds too much into zeros. We describe an effective fix for this problem. Finally, we also propose a novel modulation of costs for side-informed steganography in trunc JPEGs.

7. REFERENCES

[1] S. Agarwal and H. Farid. Photo forensics from rounding artifacts. In C. Riess, editor, *The 8th ACM Workshop on Information Hiding and Multimedia Security*, Denver, CO, 2020. ACM Press.

[2] M. Boroumand, M. Chen, and J. Fridrich. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5):1181–1193, May 2019.

[3] J. Butora and J. Fridrich. Effect of JPEG quality on steganographic security. In R. Cogranne and L. Verdoliva, editors, *The 7th ACM Workshop on Information Hiding and Multimedia Security*, Paris, France, July 3–5, 2019. ACM Press.

[4] R. Cogranne, V. Sedighi, T. Pevný, and J. Fridrich. Is ensemble classifier needed for steganalysis in high-dimensional feature spaces? In *IEEE International Workshop on Information Forensics and Security*, Rome, Italy, November 16–19, 2015.

[5] T. Denemark and J. Fridrich. Side-informed steganography with additive distortion. In *IEEE International Workshop on Information Forensics and Security*, Rome, Italy, November 16–19 2015.

[6] T. Filler, A. D. Ker, and J. Fridrich. The Square Root Law of steganographic capacity for Markov covers. In N. D. Memon, E. J. Delp, P. W. Wong, and J. Dittmann, editors, *Proceedings SPIE, Electronic Imaging, Media Forensics and Security*, volume 7254, pages 08 1–11, San Jose, CA, January 18–21, 2009.

[7] J. Fridrich, T. Pevný, and J. Kodovský. Statistically undetectable JPEG steganography: Dead ends, challenges, and opportunities. In J. Dittmann and J. Fridrich, editors, *Proceedings of the 9th ACM Multimedia & Security Workshop*, pages 3–14, Dallas, TX, September 20–21, 2007.

[8] L. Guo, J. Ni, and Y.-Q. Shi. An efficient JPEG steganographic scheme using uniform embedding. In *Fourth IEEE International Workshop on Information Forensics and Security*, Tenerife, Spain, December 2–5, 2012.

[9] L. Guo, J. Ni, and Y. Q. Shi. Uniform embedding for efficient JPEG steganography. *IEEE Transactions on Information Forensics and Security*, 9(5):814–825, May 2014.

[10] 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*, 2014:1, 2014.

[11] A. D. Ker. On the relationship between embedding costs and steganographic capacity. In M. Stamm, M. Kirchner, and S. Voloshynovskiy, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, June 20–22, 2017. ACM Press.

[12] A. D. Ker. The square root law of steganography. In M. Stamm, M. Kirchner, and S. Voloshynovskiy, editors, *The 5th ACM Workshop on Information Hiding and Multimedia Security*, Philadelphia, PA, June 20–22, 2017. ACM Press.

[13] A. D. Ker, T. Pevný, J. Kodovský, and J. Fridrich. The Square Root Law of steganographic capacity. In A. D. Ker, J. Dittmann, and J. Fridrich, editors, *Proceedings of the 10th ACM Multimedia & Security Workshop*, pages 107–116, Oxford, UK, September 22–23, 2008.

[14] J. Kodovský and J. Fridrich. Steganalysis of JPEG images using rich models. In A. Alattar, N. D. Memon, and E. J. Delp, editors, *Proceedings SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics 2012*, volume 8303, pages 0A 1–13, San Francisco, CA, January 23–26, 2012.

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

[16] W. Pennebaker and J. Mitchell. *JPEG: Still Image Data Compression Standard*. Van Nostrand Reinhold, New York, 1993.

[17] X. Song, F. Liu, C. Yang, X. Luo, and Y. Zhang. Steganalysis of adaptive JPEG steganography using 2D Gabor filters. In P. Comesana, J. Fridrich, and A. Alattar, editors, *3rd ACM IH&MMSec. Workshop*, Portland, Oregon, June 17–19, 2015.

[18] J. Ye, J. Ni, and Y. Yi. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11):2545–2557, November 2017.

[19] M. Yedroudj, F. Comby, and M. Chaumont. Yedroudj-net: An efficient CNN for spatial steganalysis. In *IEEE ICASSP*, pages 2092–2096, Alberta, Canada, April 15–20, 2018.