A VARIATIONAL APPROACH TO JPEG ANTI-FORENSICS

Wei Fan *,†, Kai Wang *, François Cayre *, and Zhang Xiong †

*GIPSA-Lab, CNRS UMR5216, 11 rue des Mathématiques, F-38402 St-Martin d’Hères Cedex, France
†School of Computer Science and Engineering, Beihang University, Beijing 100191, P. R. China

ABSTRACT

The objective of JPEG anti-forensics is to remove all the possible footprints left by JPEG compression. By contrary, there exist detectors that attempt to identify any telltale of the image tampering operation of JPEG compression and JPEG anti-forensic processing. This paper makes contribution on improving the undetectability of JPEG anti-forensics, with a higher visual quality of processed images. The employment of constrained total variation based minimization for deblending successfully fools the forensic methods detecting JPEG blocking, and another advanced JPEG forensic detector. Calibration-based detector is also defeated by conducting a further feature value optimization. Experimental results show that the proposed method outperforms the state-of-the-art methods in a better trade-off between forensic undetectability and visual quality of processed images.

Index Terms— Digital image forensics, anti-forensics, JPEG compression, total variation, subgradient method

1. INTRODUCTION

Increasing development of high-quality cameras and powerful photo-editing tools significantly reduces the difficulty to make visually plausible fake images. Doctored images are appearing with growing frequency, for instance, in advertising and in political and personal attacking. Doubts of the authenticity have been thrown upon digital images. Image forensics has enjoyed its popularity to restore some trust, as it serves as a passive and blind authentication measure as:

\[ H_I(n) = \sum_n |H_I(n) - H_{I1}(n)|, \]

where \( H_I(n) \) and \( H_{I1}(n) \) are normalized histograms of pixel differences across block boundaries and within the block, respectively.

2. RELATED WORK

During JPEG compression (here we refrain from repeating the standard JPEG compression process), two known artifacts appear, indicating the JPEG compression history of one image. The first one is the quantization artifacts in DCT domain. The DCT coefficients are clustered around the integer multiples of the quantization step length, leaving a comb-like distribution of DCT coefficients in each subband. The second one is the blocking artifacts in spatial domain. There are consistent discontinuities across block borders. Both of them are traces left from an image’s JPEG compression history.

2.1. Detecting JPEG compression

Fan and De Queiroz [1] proposed an algorithm for maximum-likelihood estimation (MLE) of the JPEG quantization table, from a spatial-domain bitmap representation of the image. The method can also serve as a detector to classify an image as not JPEG compressed, if all the entries of the estimated quantization table are either 1 or “undetermined” [1, 2].

Fan and De Queiroz [1] also proposed a JPEG blocking signature measure as:

\[ K_F = \sum_n |H_I(n) - H_{I1}(n)|, \]

where \( H_I(n) \) and \( H_{I1}(n) \) are normalized histograms of pixel differences across block boundaries and within the block, respectively.

2.2. JPEG anti-forensics

In order to disguise the quantization artifacts, Stamm et al. [2] proposed to add a dithering signal \( d \) to the DCT coefficient \( z \) in each subband:

\[ z_d = z + d, \]

so that the dithered signal \( z_d \) approximates the distribution of the unquantized coefficients. For AC components, \( d \) is distributed in
such a way that \( z_d \) approximately follows a Laplacian distribution; for the DC component, \( d \) follows a uniform distribution. The adding of \( d \) succeeds in fooling the MLE of the quantization table in [1].

After the dithering operation, Stamm et al. [3] later proposed an anti-forensic deblocking operation against the blocking artifact detector of Fan and De Queiroz [1], that is the blocking signature measure \( K_f \) of Eq. (1). For each pixel at location \((i, j)\), the anti-forensically deblocked image pixel value is obtained according to:

\[
g_{i,j} = \text{med}_s(x_{i,j}) + w_{i,j},
\]

where \( x_{i,j} \) is the original pixel value, \( \text{med}_s(\cdot) \) is the median filtering operation with window size \( s \), and \( w_{i,j} \) is a low-power white Gaussian noise of variance \( \sigma^2 \).

2.3. Countering JPEG anti-forensics

Valenzise et al. [6, 4] claimed that the dithering signal [2] degraded the image quality. A forgery detector is therefore designed by measuring the noisiness of the re-compressed image, employing the TV of the image (the \( \ell_1 \) norm of the spatial first-order derivatives) [7]. For a given image, the detector re-compresses it using different quality factors \( q \), as a function of which, TV \( (q) \) is computed. The backward finite difference quotient TV \( (h) \) with lag \( h \) is calculated as:

\[
TV^h(q) = \frac{1}{h} \left( TV(q) - TV(q - h) \right).
\]

The forensic measure is:

\[
K_f = \max \{TV^h(q)\}.
\]

Lai and Böhme [5] proposed another calibration-based detector to counter Stamm et al.’s JPEG anti-forensic method [2]. They borrowed the idea of calibration from steganalysis [8] for cropping image matrix \( X \) by 4 pixels both horizontally and vertically to obtain \( X_{cal} \). The calibrated feature \( K_L \) is established as:

\[
K_L = \frac{1}{28} \sum_{k=1}^{28} \frac{(v_{X,k} - v_{X_{cal},k})}{v_{X,k}},
\]

where \( v_{X,k} \) and \( v_{X_{cal},k} \) are the variances of the \( k \)-th high-frequency subband (defined in [5]) of \( X \) and \( X_{cal} \), respectively.

3. JPEG DEBLOCKING USING CONSTRAINED TV-BASED MINIMIZATION

Some researchers have investigated the problem of removing JPEG blocking artifacts. However, their efforts mainly focused on improving image visual quality [9, 10, 11], especially for highly compressed images. Nevertheless, the objective of anti-forensics takes higher priority of statistical undetectability than perceptual quality. For JPEG anti-forensic deblocking purpose, we hereby propose to remove JPEG blocking artifacts using a variational approach minimizing a TV-based energy, which is composed of two terms: a TV term and a TV-based blocking measurement term.

Inspired by [10], which aims to improve the visual quality of images compressed at low bit-rates by solving a constrained and weighted TV-based minimization problem, for an image \( X \) of size \( H \times W \), we first define the TV term as:

\[
TV_b(X) = \sum_{1 \leq i \leq H, 1 \leq j \leq W} e_{i,j}
\]
4. JPEG ARTIFACTS BEYOND BLOCKING

In practice, we found out that after our deblocking process described in Sec. 3, besides Fan and De Queiroz’s blocking artifact detector [1], we also succeeded in fooling the TV-based detector of Valenzise et al. [4]. Meanwhile the calibrated feature value, i.e., $K_t$ of Eq. (6), has also been significantly decreased. However, for genuine, uncompressed images, this feature value is highly condensed in an interval of very small values. It is hard to further decrease this value by performing deblocking, while keeping good visual quality.

In this section, as a first trial to defeat this detector, we will directly optimize an energy function which is very close to Eq. (6) for de-calibration purpose.

Denote $D = [d_1^T, \ldots, d_N^T]^T$ as the block DCT matrix in JPEG compression, where $d_i^T$, $i = 1, \ldots, 64$ is a $64 \times 1$ vector. We define a stacking operator $\text{stc}()$, so that $S = \text{stc}(X)$ is a $64 \times \frac{H \times W}{64}$ matrix. Each column vector $s_i$ of $S$ is a column-wise stacking of the $j$-th $8 \times 8$ block in $X$. Therefore, $d_i s_i$ is the $i$-th DCT coefficient of the $j$-th block in $X$ after block DCT. And the row vector $d_i S$ contains all the DCT coefficients in the $i$-th subband.

Hence the minimization problem is formulated as:

$$X^* = \arg \min_X \sum_{k=1}^{28} \left| \text{var}(d_{i_k} \text{stc}(X)) - \text{var}(d_{i_k} \text{stc}(X_{\text{cal}})) \right|, \quad (13)$$

where $d_{i_k}$ corresponds to the $k$-th high-frequency subband defined in [5], and $\text{var}(\cdot)$ returns the variance of the input vector. This problem can also be solved using subgradient method.

5. EXPERIMENTAL RESULTS

Our large-scale test was carried out on 1338 images of size $512 \times 384$ from the UCID corpus [14]. Without loss of generality, only the luminance component of the image is considered. Denote $\{J_k | k = 1, 2, \ldots, 1338\}$ as the set of all the original uncompressed images in UCID. We compressed each UCID image at a quality factor $1338$ randomly selected in the interval of $[30, 90]$. Thereafter a set of JPEG compressed images $\{J_k\}$ is generated. We also created five other image sets from $\{J_k\}$ as follows:

- $\{F_k^A\}$, with the application of Alter et al.’s deblocking method [10] on $\{J_k\}$;
- $\{F_k^{A,S}\}$, with the application of Stamm et al.’s deblocking method [3] on $\{J_k\}$, with parameters $\rho = 3$ and $\sigma^2 = 2$, as suggested in [3];
- $\{F_k^{A,S,\text{grd}}\}$, with Stamm et al.’s dithering signals [2] added on $\{J_k\}$ first, and then Stamm et al.’s deblocking operation [3] was applied, with parameters $s = 3$ and $\sigma^2 = 2$;
- $\{F_k^{A,F}\}$, with the application of our proposed deblocking operation described in Sec. 3 on $\{J_k\}$;
- $\{F_k^{A,F,\text{grd}}\}$, with the application of our proposed de-calibration operation described in Sec. 4 on $\{F_k^{A,F}\}$.

In order to verify the undetectability of the proposed deblocking method, it is necessary to test against detectors using different blocking criteria [15]. Therefore, besides the blocking artifact detector in [1], we build another JPEG blocking signature measure:

$$K_{\nu}^{\text{B}} = \| B_{\nu}^p (X) - B_{\nu}^p (X_{\text{cal}}) \|,$$  \hspace{1cm} (14)

where $B_{\nu}^p$ is the gradient aware blockiness [16], which is the normalized $\ell_2$ norm of the weighted gradient computed from each group of four adjacent pixel values across block borders.

In this paper, the detectors described in Sec. 2, together with the above blocking artifact detector, are used for the testing of forensic undetectability of the image sets. For the sake of simplicity, we name the detectors directly using the feature value name, that is $K_{\nu}^{\text{B}}$, $K_{L}$, $K_{F}$, and $K_{\nu}^{\text{F}}$ in Eqs. (5), (6), (1), and (14) respectively. Here, we consider the parameters $h = 5$, $p = 1$, and $p = 2$. Figure 2 shows the ROC curves of different image sets against the detectors $K_{\nu}^{\text{F}}$, $K_{L}$, $K_{F}$, $K_{\nu}^{\text{F}}$, and $K_{\nu}^{\text{B}}$ respectively. The detection reliability $\rho = 2a - 1$, where $a$ is the area under the ROC curve [16], is reported in Table 1. The average PSNR and SSIM [17] values are provided in Table 2 for the comparison of the quality of processed images.

Apparently fooling detectors is not the goal of Alter et al.’s work [10], as their main focus is to improve the image perceptual quality.

\footnote{1All the JPEG compression and decompression operations in this paper were performed using libjpeg version 6b provided by Independent JPEG Group. More information can be found at http://www.ijg.org/.

2Note that in Table 2, the average PSNR and SSIM values of $\{F_k^{A}\}$ are both lower than those of $\{J_k\}$. The reason may be that the parameter setting in [10] is optimized for low bit-rate compression, but not for the whole quality factor range of [30, 90].}
The regularization parameter $\lambda$ (please refer to the electronic version for a better visibility). Eq. (12) are two parameters we can adjust. During the simulation, we set $\lambda = 1.5$, and $t = 1/k$ at the $k$-th iteration. This parameter setting works well for our JPEG anti-forensic purposes.

In order to converge to $X^*$ in Eq. (12), one would like to run as many iterations as possible. However, we found that the iteration giving the best $K_U$, $K_V$, or $K_U^*$ values are not always in the convergence. This can be explained by the fact that we deblocked the image not in a way directly minimizing the blocking signature measures, as they have different blocking criteria. In practice, we run 50 iterations, and choose the one giving the smallest $K_U$ value as the final result. This gives satisfying results against all the three blocking artifact detectors. Moreover, it is also able to fool another advanced detector $K_V$. The reason may be that the TV term of Eq. (7) suppresses the unnatural noises that can be detected by $K_V$.

For the de-calibration operation, as we directly minimize the feature value, we are capable of obtaining very small $K_L$ values when converging to $X^*$ in Eq. (13). In order to fool the detector, a random threshold for each image is drawn from the distribution of the calibrated feature values for genuine, uncompressed images, and the iteration stops once $K_L$ value drops below it.

During our deblocking and de-calibration processes, we did not explicitly smooth the DCT coefficient histogram. However, in practice, we observe that the gaps in the DCT coefficient histogram were plausibly smoothed (example DCT histograms of $F_{k_0}$ subband are shown in Fig. 3). We also used the MLE method [1], the targeted detector of Stamm et al.’s DCT histogram smoothing method in [2], to estimate the quantization table of the processed images $\{F_{k_0}^{k_0}P_c\}$. None of estimates was correct, and a very high portion of 93.20% of the images had the estimated quantization table full of entries being either 1 or “undetermined”. This is an interesting observation that deserves future investigation. In all, our JPEG anti-forensic images can be passed off as never compressed by testing under existing JPEG forensic detectors. However we are aware that there might still exist some artifacts which can be detected by more reliable detectors to be designed in the future.

Further research shall be devoted to advanced optimization methods for solving the minimization problems, to further investigation of the calibrated feature in order to avoid optimizing the feature value directly, and to explicitly smoothing the DCT coefficient histogram to better approximate the original distribution.

Another interesting point might be to apply our JPEG anti-forensics to counter the most recent forgery localization method of JPEG images [18, 19], and to compare the performance of our method with that of other advanced histogram-based anti-forensic methods, e.g., [20].

6. DISCUSSION AND CONCLUSION

The regularization parameter $\lambda$ in Eq. (11) and the step size $t$ in Eq. (12) are two parameters we can adjust. During the simulation, the calibration-based detector

Fig. 3. Example results (close-up images) of the proposed method compared with Stamm et. al.’s methods [2, 3]. The original $k_0$-th image $I_{k_0}$ in UCID corpus has been JPEG compressed with quality factor 50 to obtain $J_{k_0}$. The small figures in the bottom left are the histograms of (2, 2) DCT coefficient of the corresponding images.

<table>
<thead>
<tr>
<th>${J_{k_0}}$</th>
<th>$K^0_V$</th>
<th>$K_L$</th>
<th>$K_P$</th>
<th>$K_U^0$</th>
<th>$K_U^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>${F_{k_0}^V}$</td>
<td>0.9853</td>
<td>0.9925</td>
<td>0.9998</td>
<td>0.9964</td>
<td>0.8812</td>
</tr>
<tr>
<td>${F_{k_0}^L}$</td>
<td>0.9435</td>
<td>0.9764</td>
<td>0.6352</td>
<td>0.9901</td>
<td>0.7229</td>
</tr>
<tr>
<td>${F_{k_0}^{P}}$</td>
<td>0.2837</td>
<td>0.3057</td>
<td>−0.0106</td>
<td>0.5298</td>
<td>−0.0090</td>
</tr>
<tr>
<td>${F_{k_0}^{S}}$</td>
<td>0.7163</td>
<td>0.1666</td>
<td>−0.2293</td>
<td>0.4430</td>
<td>0.0994</td>
</tr>
<tr>
<td>${F_{k_0}^{V\cdot P}}$</td>
<td>−0.2264</td>
<td>0.9896</td>
<td>−0.0111</td>
<td>0.2579</td>
<td>0.1048</td>
</tr>
<tr>
<td>${F_{k_0}^{V\cdot S}}$</td>
<td>−0.2217</td>
<td>−0.1867</td>
<td>0.0031</td>
<td>0.2688</td>
<td>0.1609</td>
</tr>
</tbody>
</table>

Table 2. Comparison of image quality.

<table>
<thead>
<tr>
<th>PSNR</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>35.1355</td>
<td>34.1031</td>
</tr>
<tr>
<td>30.2676</td>
<td>29.8271</td>
</tr>
<tr>
<td>34.1186</td>
<td>34.0484</td>
</tr>
<tr>
<td>0.9870</td>
<td>0.9820</td>
</tr>
<tr>
<td>0.9528</td>
<td>0.9407</td>
</tr>
<tr>
<td>0.9479</td>
<td>0.9789</td>
</tr>
<tr>
<td>0.9785</td>
<td></td>
</tr>
</tbody>
</table>

Stamm et al.’s deblocking method [3] successfully fools blocking artifact detectors $K_P$ and $K_V$, however $K_U$ still keeps a detection reliability value of around 0.5 for both $\{J_{k_0}\}$ and $\{F_{k_0}^{V\cdot P}\}$. And the noise introduced by the dithering signal [2] can still be detected by $K_V^*$ after the deblocking operation [3]. Moreover, as pointed out in [15], their deblocking attack has not been tested against median filtering forensic detectors.

As shown in Fig. 2-(e), the proposed deblocking method successfully fools all the three blocking artifact detectors, as well as the TV-based detector $K_V$, at the cost of slightly lower visual quality than the JPEG compressed images (Table 2). After the de-calibration process, the calibration-based detector $K_L$ is also defeated, as shown in Fig. 2-(f); meanwhile we keep high undetectability against other detectors and a high level of image visual quality (Table 2). Compared to Stamm et al.’s methods [2, 3], our method achieves a better trade-off between undetectability and the visual quality of processed images: the average PSNR value has been improved by 4.2 dB.

Figure 3 shows the processed anti-forensic images from an example JPEG compressed image at quality factor 50. As expected, Fig. 3-(d) processed using our TV-based method better preserves the image details such as textures and edges than (c). It can also be observed that even after the deblocking operation [3], the spatial-domain noise introduced by the dithering signal [2] can still be noticed at the smooth areas of the image, e.g., the sky area in Fig. 3-(c) (please refer to the electronic version for a better visibility).
7. REFERENCES


