Residual Information of Redacted Images Hidden in the Compression Artifacts

Nicholas Zhong-Yang Ho   Ee-Chien Chang

School of Computing
National University of Singapore
iam@nicholasho.net   changec@comp.nus.edu.sg

Abstract. Many digital images need to be redacted before they can be disseminated. A common way to remove the sensitive information replaces the pixels in the sensitive region with black or white values. Our goal is to study the effectiveness of this simple method in purging information. Since digital images are usually lossily compressed via quantization in the frequency domain, each pixel in the spatial domain will be "spread" to its surroundings, similar to the Gibbs-effect, before it is redacted. Hence, information of the original pixels might not be completely purged by replacing pixels in the compressed image. Although such residual information is insufficient to reconstruct the original, it can be exploited when the content has low entropy. We consider a scenario where the goal of the adversary is to identify the original among a few templates. We give two approaches and investigate their effectiveness when the image is compressed using JPEG or wavelet-based compression scheme. We found that, if a redacted image is compressed in higher bit rate compared to the compression of the original image, then the correct template can be identified with noticeable certainty. Although the requirements are stringent, it will not be surprising that redacted images matching the requirements can be found in the public domain. Hence, our findings highlight a subtle attack that must be considered when declassifying images.

1 Introduction

Many digital images need to be redacted before they are disseminated. Consider a scenario where an archive of scanned documents is to be released to the public and some sub-regions in the scanned documents contain sensitive information, such as name, age or address of an individual. These information have to be removed before the whole image is released. In many cases, it is infeasible to redact the hard copies and re-digitize them, since the original hard copies may not be available, not be allowed to be damaged, or be simply too complicated to do so. A typical approach is to digitally redact the image by replacing each sensitive region with other values, for example white, black or some images indicating that this region has been redacted (Fig. 1(b)). Other examples are images of driver's license with sensitive information such as birth date, or images of road accidents with vehicle's plate numbers that needed to be redacted.

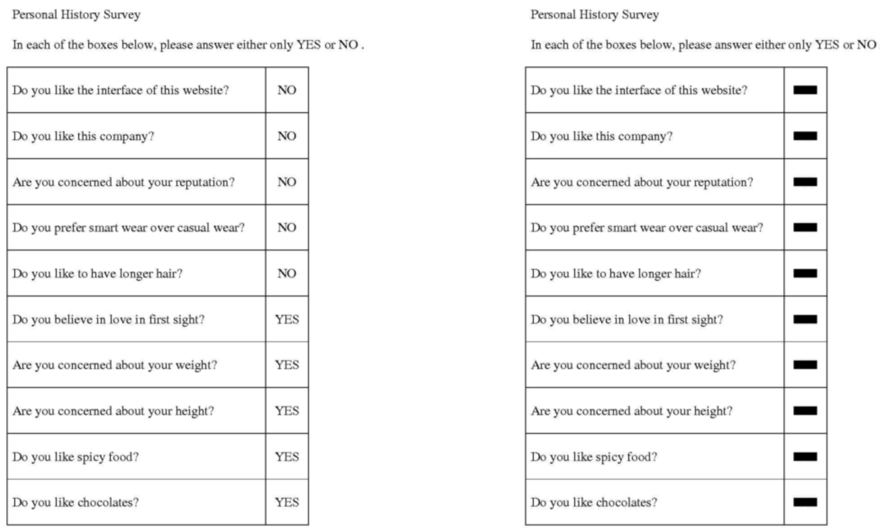
An interesting question now is, although the sensitive region has been replaced, can we deduce its content from the redacted image? Most images in the public domain are lossily compressed. Popular compression like JPEG and JPEG2000 quantized the coefficients in the transformed domain. Hence, information of a pixel will “spread” to other pixels in the spatial domain, creating compression artifacts like the Gibbs-effect. In other words, before an image is redacted, information in the sensitive region has already “spread” to surrounding. Hence, even if the sensitive region is replaced, some residual information might still remain in the surrounding regions. Thus, it is clear that there is information leakage through the compression artifacts. The next interesting question is whether such information is sufficient in deriving the content in the redacted region. Although it is unlikely that the original image can be reconstructed, the residual information might be useful when the content of the removed region has low entropy. The image in Fig. 1(a) & 6(a) are such examples. Each redacted region contains either the word “YES” or “NO”. Furthermore, the fonts can be derived from other parts of the image. In such a situation, we can assume that an adversary has a few possible templates of the region removed, and his goal is to identify the template that is closest to the original.

Note that we do not consider information leakage that can be inferred from the semantic of the image. An example of such information is the size of the region, which revealed useful information[3, 10]. Another example is words that are not completely covered[10]. We are also not considering physical redaction like markings on the hard copies[10] or using a lower resolution optical device [2]. There are some techniques [14, 9, 1] on document redaction that works on the documents directly, like the tools available to redact PDF documents[8]. We also do not consider these techniques and tools, since they handle the documents directly before the documents are converted to images. Instead, in this paper, we are looking for artifacts that are generated as side-products of image processing. A digital image typically has to undergo a series of image processing operations before they are redacted and published. Many of these artifacts are not purged during redaction and residual information might be hidden in the artifacts. These are the types of information that we wish to exploit in recovering the secrets.

We propose two methods in recovering the secrets. The first method assumes that the adversary has a good estimate of the original raw image in the non-sensitive region, whereas the second method does not make such an assumption.

Note that the redacted image actually has been compressed at least twice: before redaction, and right after redaction. The quantization level applied in these two steps will affect the amount of information retained. Furthermore, it is unlikely that the exact original is one of the templates, due to noise like geometric distortion. In addition, the adversary may not know the exact compression parameters. In our experimental studies, we investigate the proposed methods under various types of noise and uncertainties.

Our experimental results show that the success rate of the adversary is noticeable when the second compression rate is of a much higher quality compared to the first compression, and the noise in the template is low. Such requirements are stringent. Nevertheless, it will not be surprising to find some redacted images meeting the requirements. Hence, this subtle attack must still be taken into consideration when redacting sensitive images.



Personal History Survey

In each of the boxes below, please answer either only YES or NO .

Do you like the interface of this website?	NO
Do you like this company?	NO
Are you concerned about your reputation?	NO
Do you prefer smart wear over casual wear?	NO
Do you like to have longer hair?	NO
Do you believe in love in first sight?	YES
Are you concerned about your weight?	YES
Are you concerned about your height?	YES
Do you like spicy food?	YES
Do you like chocolates?	YES

Personal History Survey

In each of the boxes below, please answer either only YES or NO .

Do you like the interface of this website?	[redacted]
Do you like this company?	[redacted]
Are you concerned about your reputation?	[redacted]
Do you prefer smart wear over casual wear?	[redacted]
Do you like to have longer hair?	[redacted]
Do you believe in love in first sight?	[redacted]
Are you concerned about your weight?	[redacted]
Are you concerned about your height?	[redacted]
Do you like spicy food?	[redacted]
Do you like chocolates?	[redacted]

(*a*) (*b*)

Fig. 1. (a) A document image. (b) A redacted version.

(*a*) (*b*)

Fig. 2. Two templates derived from the redacted image in Fig. 1 (b).

2 Problem formulation

Let $C_\delta(I)$ be the lossily compressed I with quantization parameter δ in the transformed domain[1]. Given an image I, let $R(I, r, M)$ be the modified image where the pixels of I in the region r is replaced by the *mask* M. The region r is typically rectangular and can be represented by its corners. The mask can be all white, black, or image of symbols indicating that the region has been redacted. Let $T_{I,r}$ to be the sub-image of I in the region r. When it is clear in the context, for simplicity, we write $R(I, r, M)$ as $R(I)$, and the $T_{I,r}$ as T.

2.1 Process of redaction

Let I_0 to be the *raw image*, which can be document image captured by camera, scanner or image generated by document editing tools, before compression. The raw image is lossily compressed, giving $I_1 = C_{\delta_1}(I_0)$ and passed to the redactor. The redactor wants to remove information in a region r. Note that the actual intention of the redactor is to remove information from the raw image I_0, which the redactor does not have. Instead, the redactor replaces pixels in I_1 by some mask M, giving the modified image $I_2 = R(I_1, r, M)$. Let us call I_2 the *raw redacted image*. Next, I_2 is lossily compressed with parameter δ_2, giving the final *redacted image* I_3 that is to be disseminated.

Here are the detailed steps in obtaining the redacted I_3 from the raw I_0.

1. The raw image I_0 is compressed giving I_1.

$$I_1 = C_{\delta_1}(I_0). \tag{1}$$

2. The redactor replaces pixels in region r by the mask M, giving the raw redacted image I_2.

$$I_2 = R(I_1, r, M). \tag{2}$$

3. The raw redacted image is compressed with parameter δ_2, giving the final redacted image I_3.

$$I_3 = C_{\delta_2}(I_2) = C_{\delta_2}(R(C_{\delta_1}(I_0))). \tag{3}$$

2.2 Goal of the adversary

The adversary has the redacted image I_3 in equation (3). We assume that the adversary knows the mask M and the region redacted r. In addition, he has two templates $\widetilde{T}_0$ and $\widetilde{T}_1$, where one of them is a noisy version of $T_{I_0,r}$. The adversary derives the templates from I_3 together with other background knowledge, for example, a font file. Hence, we write $\widetilde{T}_i = \texttt{Template}(I_3, i)$ for $i = 0, 1$ where

[1] The type of the parameter δ depends on the compression scheme, for example, it is the quantization table for JPEG.

`Template` is the method the adversary employs in guessing the templates. Since it is unreasonable to assume that the algorithm `Template` is able to output a template that is exactly same as $T_{I_0,r}$, we assume that there is noise like additive white noise and geometric deformation.

Let the *secret* $s = 0$ if $\widetilde{T}_0$ is the noisy version of $T_{I_0,r}$, and $s = 1$ otherwise. The secret can be viewed as a one-bit content that is removed from the image. Let us assume that the raw image is from a source such that the secret is equally likely to be 0 or 1. Thus, without seeing the redacted image I_3, the adversary can correctly guess the secret with probability 0.5.

The goal of the adversary, given I_3, is to correctly guess the secret s. If he succeeds with probability $0.5 + \epsilon$, we say that he achieves an advantage of ϵ in identifying the original. In other words, with the redacted image I_3, he can improve his chances by ϵ. If the adversary has non-zero advantage, the redacted image I_3 must still contain some information of the secret s. Note that this security notion is loosely inspired by the formulation of semantic security [7].

We assume that the adversary knows the redaction process, in particular, the compression scheme in used. He knows the parameter δ_2, which can be easily obtained from the header information in I_3. The adversary does not know δ_1. However, he can obtain an estimation of δ_1 by analyzing the distribution of the coefficients of I_3. There are a number of techniques that estimate the quantization in the studies of image forensic[5, 16, 13] and image steganography[6]. Let the estimated parameter be $\tilde{\delta}_1$.

Below is the summary of what the adversary knows.

- I_3, the redacted image.
- r, M, the region and the mask.
- $\widetilde{T}_0, \widetilde{T}_1$, two templates obtained using some background information and I_3.
- δ_2, the quantization parameter for the second compression.
- $\tilde{\delta}_1$, an estimation of the parameter δ_1 for the first compression.

In addition, the adversary may be able to reduce compression artifacts from I_3. That is, getting an approximation of $R(I_0, r, M)$. This is possible in some cases. For example, if the image is a document and the adversary is aware of the fonts library, he may attempt to reconstruct the document. If an accurate approximation of $R(I_0, r, M)$ is obtained, then the adversary can easily obtain the compression artifacts $R(I_0, r, M) - I_3$. On the other hand, the size of I_3 is generally much larger than the redacted region r. Thus, total error in estimating $R(I_0, r, M)$ could be significant. Nevertheless, such assumption is still reasonable when the compression scheme is JPEG, which divides the images into small 8×8 blocks. One of our proposed methods exploits this assumption.

- The adversary knows $\widetilde{R}$, an approximation of $R(I_0, r, M)$.

The performance of an adversary will be affected by the noise in estimating the templates, the relationship between δ_1 and δ_2, and the noise in estimating δ_1. In addition, the accuracy of the approximation of $R(I_0, r, M)$ if the adversary chooses to exploit this information.

3 Proposed methods

We will present two general methods. The first method requires and exploits the assumption that the adversary has an approximation of $R(I_0, r, M)$, whereas the second method does not require that. The first method is suitable for JPEG because each 8×8 block is relatively small, and it is feasible to estimate the raw image for the small block accurately. On the other hand, it is not easy to be applied on wavelet-based compression because each coefficient contains information from a large region.

Intuitively, the first method, starting from an estimate of the raw image, simulates the redaction process and then compares the differences between the actual redacted image I_3 and the simulated image in the spatial domain. The second method, starting from an estimate of the raw sub-image in the redacted region, obtains an estimate of the compressed (under the first compression) sub-image. Next the redaction process is simulated, and finally the actual image I_3 is compared with the simulated redacted image in the transformed domain.

3.1 First method - Comparison in the spatial domain

Recall that, given the redacted image I_3 and background knowledge, the adversary can derive $\widetilde{R}$, an approximation of $R(I_0, r, M)$, and two templates T_0 and T_1. Let T_0^β and T_1^β be the geometrically distorted copy of the respective T_0 and T_1 under some parameter β. Let $\mathcal{T}$ be a collection of T_0^β and T_1^β for all β's. For example, $\mathcal{T}$ can be the collection of 18 templates that are translated horizontally, vertically by 1 pixel, and combinations of both.

The main idea is to find the $\widetilde{T} \in \mathcal{T}$ such that a composed image of $\widetilde{T}$ and $\widetilde{R}$ is most similar to I_3. The corresponding undistorted template of $\widetilde{T}$ (that is, either T_0 or T_1), is then declared as the revealed secret.

Here are the detailed steps: For a $\widetilde{T} \in \mathcal{T}$, the following are carried out.

1. A composed image $\widetilde{I}$ is obtained by replacing the redacted region in $\widetilde{R}$ by $\widetilde{T}$.
2. The redaction process described in Section 2.1 is performed on $\widetilde{I}$ using the parameters $\widetilde{\delta}_1$ and δ_2. Let the redacted image be $\widetilde{I}_3$.
3. Compute the difference of $\widetilde{I}_3$ and I_3. Let the difference be $d_1(\widetilde{T})$.

Finally, determine the $\widetilde{T}$ that minimizes $d_1(\widetilde{T})$. If $\widetilde{T}$ is derived from T_0, then declare the secret is 0, otherwise, declare the secret as 1.

3.2 Second method - Comparison in the transformed domain

Unlike the previous section, $\widetilde{R}$ is not available. So, a straightforward comparison of the composed image and I_3 cannot be carried out. Instead, in this method, they are compared in the transformed domain. The main idea is as follows: Consider $T_{I_1,r}$, which is the sub image in the redacted region of I_1(see Section 2 for the notations). The coefficients of the combined image of $T_{I_1,r}$ and I_3 should follow closely the distribution of coefficients quantized with parameter δ_1. Hence, given a $\widetilde{T} \in \mathcal{T}$, the adversary can try to obtain an estimate of $T_{I_1,r}$, which can then be filled into I_3. The distribution of the coefficients of the composed image is then examined. Note that the effect of the second compression is not taken into consideration and is treated as noise.

Here are the detailed steps: For a $\widetilde{T} \in \mathcal{T}$, the following are carried out.

1. An image $\widetilde{I}$ is obtained by replacing the redacted region in I_3 by $\widetilde{T}$.
2. The image $\widetilde{I}$ is compressed with quantization $\widetilde{\delta}_1$. Let the compressed image be $I_{\texttt{temp}}$. The sub-image of $I_{\texttt{temp}}$ in the redacted region is treated as an approximation of $T_{I_1,r}$. Let us write this sub-image as $\widetilde{T}_{I_1,r}$
3. Compose an image by replacing the redacted region in I_3 by $\widetilde{T}_{I_1,r}$. This can be viewed as an approximation of I_1 and let this image be $\widetilde{I}_1$.
4. Next, $\widetilde{I}_1$ is transformed and quantized one more time with parameter $\widetilde{\delta}_1$. Let $d_2(\widetilde{T})$ be the quantization error.

Finally, determine the $\widetilde{T}$ that minimizes $d_2(\widetilde{T})$. If $\widetilde{T}$ is derived from T_0, then declare that the secret is 0, otherwise, declare the secret as 1.

There are a few ways to measure quantization error in step 4. In our experiments, we employ a weighted Euclidean distance, where the weight is the inverse of the step size. That is, suppose $C = \{c_1, c_2, \ldots, c_k\}$ a set of k coefficients, and s_i is the quantization step size for the coefficient c_i, then the quantization error is:

$$\sqrt{\sum_i^k \frac{1}{s_i} \left| c_i - s_i \cdot round\left(\frac{c_i}{s_i}\right) \right|^2}$$

4 Experiment

Test Images. We conduct experiments on two sets of images. The first set of images are uniformly randomly generated images, where each pixel is uniformly distributed in the range 0 to 255. The main purpose of using random images is to obtain a large number of images, so as to facilitate analysis of the attack effectiveness against different types and levels of noise.

The second set of images consists of a document image and a mobile phone image. The document image I_1 is shown in Fig. 1(a), and the redacted image shown in Fig. 1(b), where the sensitive information is covered by the black boxes. The size of I_1 is 1034×1494 pixels, and the size of each redacted region is 70×28 pixels. The two templates of “Yes” and “No” shown in Fig. 2 are derived from Fig. 1(b). The mobile phone image is (Fig. 6(a)) captured by a mobile phone with manufacturer recommended parameters.

Compression. We focus on two image compression schemes - JPEG compression and Wavelet-based compression (used in JPEG2000)[12]. The JPEG quantization matrices used in our experiments are obtained from a Matlab JPEG

Toolbox by Sallee [15]. Each quality value (ranging from 0 to 100) is assigned a quantization matrix. Appendix A shows some matrices and their corresponding quality values.

For the wavelet-based compression, we use the Cohen-Daubechies-Feauveau (CDF) 9/7 wavelet transform [4]. The lossy compression is done by applying scalar quantization on the coefficients. The subsequent lossless compression does not play a role in our problem.

4.1 Random JPEG Images

General setting. Since JPEG divides an image into 8×8 blocks and lossily compresses the blocks independently, it is suffice to work on random images of size 8×8. The experiments are conducted with varying levels of noise parameters, and are designed to aid in the analysis of how the following affect the adversary's success rate:

1. The area redacted. Specifically, the number of columns redacted in a block.
2. The parameters of the two JPEG compression, δ_1 and δ_2.
3. Noise in the templates.
4. The uncertainty in obtaining the first compression parameter $\widetilde{\delta}_1$.
5. The noise in $\widetilde{R}$.

Generating the random images and templates. Without loss of generality, let the secret s be 0. Here are the steps in preparing the following information for the adversary: a redacted random image I_3, the templates T_0 and T_1, and the estimated redacted image $\widetilde{R}$.

1. Let I_0 be a uniformly and randomly generated 8×8 pixels block.
2. Extract template T_0 from image I_0. Extract template T_1 from another randomly generated 8×8 pixels block.
3. Compress image I_0 at JPEG compression quality δ_1 to get I_1.
4. Image I_1 is redacted and compressed at JPEG compression quality δ_2 to produce the redacted image I_3. (Equations (2) and (3))
5. Gaussian white noise is added to I_0, which in turn gives $\widetilde{R}$. Noise is also added to T_0 and T_1 to give $\widetilde{T}_0$ and $\widetilde{T}_1$.

Success rate. We call the variance of the white noise as the noise level. Given the randomly generated I_3, $\widetilde{T}_0$ and $\widetilde{T}_1$, the proposed method is carried out to produce a guess of the secret. For each set of parameters, the experiment is repeated for 1000 samples of randomly generated I_3, $\widetilde{T}_0$ and $\widetilde{T}_1$. The ratio of the correct guess is the estimated success rate. Note that the success rate is for a single block. If the image in question contains multiple blocks along the boundary of the redacted region, the adversary can make a decision using majority vote, which significantly improves the overall success rate.

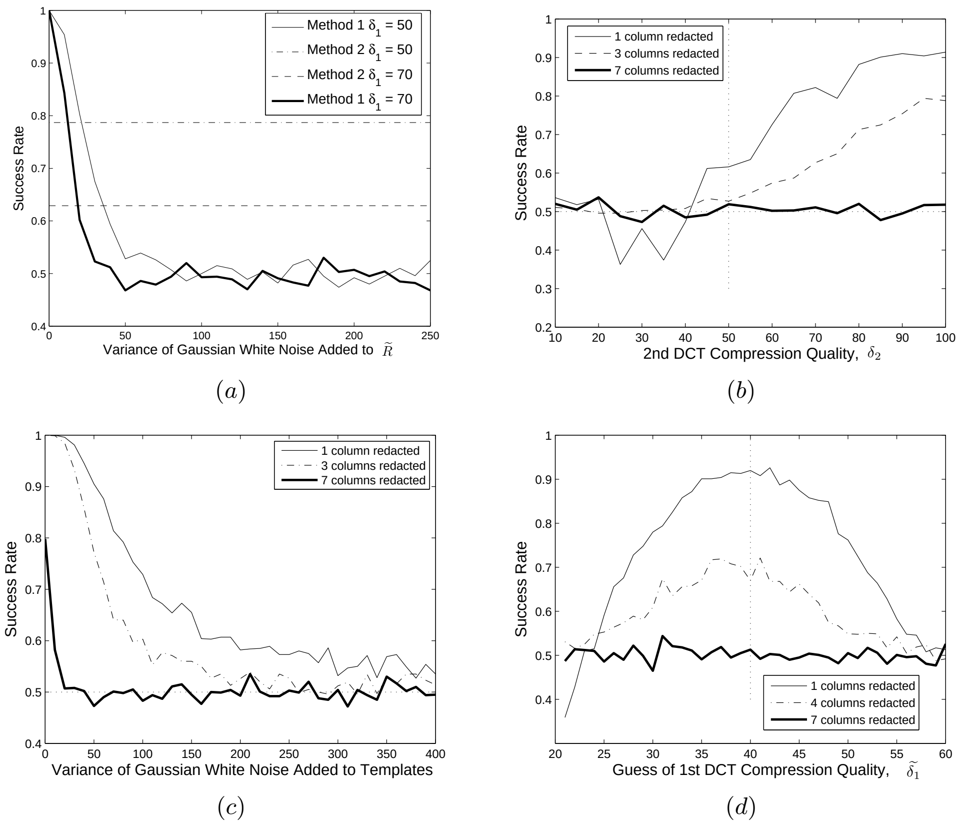


Fig. 3. (a) Success rate for image block, $\delta_1 = 50, 70$, $\delta_2 = 95$. All curves are results of redacting 3 columns. Both lines generated by the second method have Gaussian white noise (variance is 50 per pixel) added to templates. (b) Success rate of second method for image block, $\delta_1 = 50$ (indicated by the vertical line), Gaussian white noise (variance = 50) added to templates. (c) Success rate of second method for image block, $\delta_1 = 50$, $\delta_2 = 95$. (d) Success rate of second method for image against adversary's guess of first compression $\widetilde{\delta_1}$, The actual $\delta_1 = 40$ is indicated by the vertical line. The parameter $\delta_2 = 90$, and variance of Gaussian white noise added to templates is 50 (per pixel).

Effect of the area redacted. Fig. 3(b) shows the success rate for various values of δ_2, with δ_1 fixed at 50. Gaussian white noise with variance = 50 has been added into the templates. We have repeated the experiment with $1, 2, \ldots, 7$ columns redacted. The results show that the larger the area of redaction, the lower the success rate near the larger δ_2 values.

Effect of the two JPEG compression parameters δ_1 and δ_2. Fig. 3(b) also shows that at higher δ_2 values, the success rate of the adversary improves almost linearly. However, at the smaller values of δ_2, success rate falls to 0.5.

Effect of noise in the templates. Fig. 3(c) shows the success rate of curves for various noise levels, where $\delta_1 = 50$ and $\delta_2 = 95$. Under the noise, each pixel in the template is corrupted by additive Gaussian white noise. The results show that as the amount of Gaussian white noise is added into the templates, the success rate decreases.

Effect of adversary guessing δ_1 wrongly. Fig. 3(d) shows the success rates for guessing δ_1, where actual $\delta_1 = 40$, $\delta_2 = 90$. The results in the figure shows that the closer the adversary's guess of δ_1 is to the actual δ_1, the better the success rates of the adversary to reveal the data hidden by redaction.

Effect of accuracy of approximating R on adversary success rate. Fig. 3(a) shows the success rate for both methods at two different values of $\delta_1 = 50, 70$ as accuracy of approximating $\widetilde{R}$ varies. In the figure, we can see that the first method's success rate is very sensitive to the accuracy of approximating $\widetilde{R}$. With noise level above 25 to 30, the first method fares worse than the second method.

4.2 Random JPEG2000 Images

General Setting. Due to the use of wavelet transform in JPEG2000, the visual artifacts are "spread" over a much wider area, as compared to the DCT compression artifacts. As a result, the first method is unsuitable to be used for JPEG2000 images. Thus, in this paper, only the second method will be discussed for all experiments involving wavelet transformed images. Lossy compression is achieved by scalar quantization. We call the reciprocals of the quantization step the compression quality.

Generation of Random Images and Template. The method of generation of the random images and template is similar to that described in Section 4.1 except that the size of images is 256×256 pixels. The redacted portion consists of vertical columns of the pixel block starting from the left side.

Parameters of the compression quality δ_1 and δ_2. Fig. 4 shows a similar trend as those seen in JPEG experiments so far. That is, at higher δ_2 values, the success rate of the adversary improves. However, at the smaller values of δ_2, success rate falls to around 0.5.

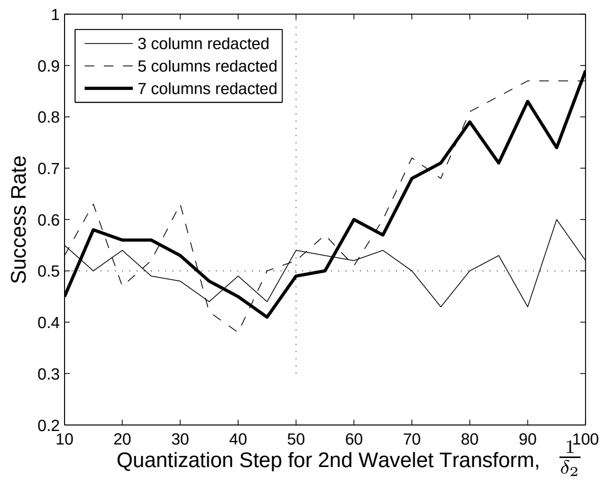


Fig. 4. Success rate for image JPEG2000 block using second method, $\delta_1 = 50$ (indicated by the vertical line), Gaussian white noise (variance = 10) added to templates.

4.3 Document Image

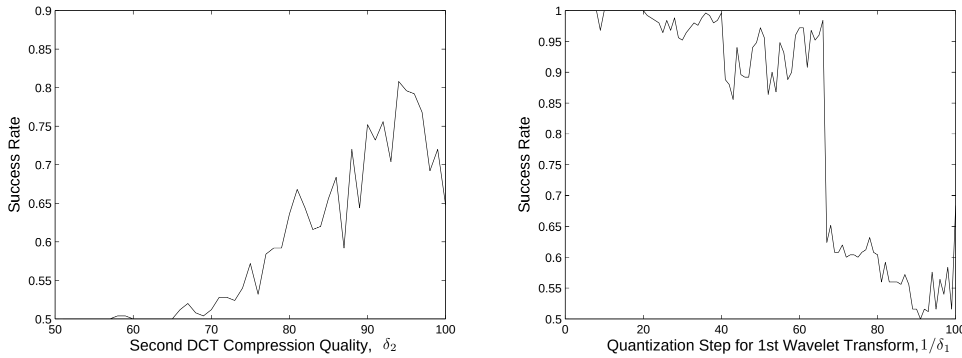
General setting. Instead of applying the method on a 8×8 pixels block in Section 4.1, this section will deal with applying the 2nd method on a redacted binary document image (shown in Fig. 1(a). Both JPEG and wavelet transform will be tested on the document image using the method described in Section 3.2.

Let I_0 be the raw image shown in Fig. 1(a). The redacted image I_2 and templates T_0 and T_1 are prepared in the following way:

1. Compress image I_0 with quality δ_1 to give I_1.
2. Five "YES" and "NO" subimages are extracted from I_0, from which the two templates "YES" and "NO" are derived manually.
3. Image I_1 is redacted and compressed with quality δ_2 to produce image I_3 shown in Fig. 1(b). (Equations (2) and (3))
4. In addition, during guessing, in order to correct the geometric distortion, each template is translated horizontally and vertically by at most a pixel. Thus, there are a total of translated 9 copies for each template.

Since JPEG involves block-wise compression, the success rate in Fig. 5(a) is calculated by collectively comparing all the blocks intersecting the border of the redacted zone. As for JPEG2000, since it does not involve block-wise transformations, the whole document is compared to determine the success rate in Fig. 5(b).

Relationship of δ_1 and δ_2 for JPEG Compression In Fig. 5(a), observe that when the second compression $\delta_2 < 65$, the chances of the adversary are only as good as guessing.



(a) (b)

Fig. 5. (a) Success rate for binary image using second method with JPEG compression quality $\delta_1 = 50$
(b) Success rate for binary image using second method with wavelet transform quantization step $\delta_2 = 1/100$

Relationship of δ_1 and δ_2 for Wavelet Transform Fig. 5(b) shows the success rate for varying values of δ_1, where $\delta_2 = 1/100$. The values listed on the horizontal axis refers to the quantization step size of δ_1 from 1 to 1/100. When $\delta_1 = 1/50$ and 1/100, the percentage of zeros among all coefficients is 85.04% and 83.88% respectively. Note that these percentage reflect the compression rate [11]. As we can see from the figure, the success rate is fairly high when the compression parameter δ_2 is significantly more than δ_1. Note the interesting zig-zag shape of the curve. We suspect that the success rate depends on whether $(1/\delta_1)$ is an integer multiple of $(1/\delta_2)$. Further investigations are required.

4.4 Mobile Phone Camera Test

A postal box image was taken with a Nokia 6125 mobile phone ("normal" JPEG compression quality, image size at 640×480, grey scale effect). This image is then redacted and compressed with quality $\delta_2 = 90$ as shown in Fig. 6(a). The redacted text in the top and bottom left is "10-335" and "10-339" respectively. We assume that the adversary knows the first compression quality δ_1, and he knows that the text is one of the five candidates indicated in Fig. 7.

To prepare the templates, high quality 5 megapixels images of similar postal boxes were taken with a FujiFilm FinePix 31fd digital camera. The high quality images were then digitally adjusted to estimate the templates as shown in Fig. 6(b). Note that all the templates in Fig. 6(b) are derived from the images taken by the FujiFilm camera.

A test using the second method was carried out to recover the redacted information at the top and bottom left black boxes, and the results is tabulated

(*a*) (*b*)

Fig. 6. (a) Image captured by a Nokia 6125 mobile phone and then redacted. (b) Templates of postal boxes.

in Fig. 7. From the left table in Fig. 7, the adversary can narrow the candidate down to "10-335" and "10-339". In the right table, the correct template "10-339" gives significantly lower errors.

Results for Top Left Box

Data Name	Quantization Error
Random Templates	123.0
10-335	92.6
10-339	92.2
08-331	95.0
11-335	96.9
11-339	97.3

Results for Bottom Left Box

Data Name	Quantization Error
Random Templates	104.9
10-335	69.1
10-339	67.1
08-331	71.7
11-335	72.8
11-339	73.7

Fig. 7. Results of second method on the redacted image in Fig. 6(a).

5 Counter Measure

Since JPEG quantizes the block independently, by removing the whole 8×8 pixel block, all compression artifacts will be purged. If the above measures are not possible, then the image should be compressed in a lower bit rate after redaction. Alternatively, noise can be added to the redacted regions and its surrounding regions before the second compression. Additional studies are required to determine the level of noise required to prevent leakage of information in the redacted images.

6 Conclusion

In this paper, we argue that information leftover in the compression artifacts may contain sufficient information to recover the redacted secret. We studied the redaction process and identified a few parameters that affect the success rate of the adversary. Experiment results show that it is possible to recover the secret hidden within the compression artifacts, albeit effective only under stringent conditions, in particular the redacted image is compressed in higher bit rate than the original image. Although the requirements are stringent, nevertheless, such subtle attack must still be taken into consideration when redacting sensitive images. Furthermore, as mobile camera phones are gaining popularity, there could be more publicly available images which are first compressed with lower quality before they are redacted. It would also be interesting to further explore other types of image processing artifacts to determine which of them can also be exploited to reveal hidden information.

References

[1] G.B. Anderson, B.P. Gross, J.W. Marlin, and V. D. Tucker. Method for storing and retrieving annotations and redactions in final form documents. *US Patent*, (5581682), 1996.

[2] S. Berger, R. Kjeldsen, C. Pinhanez, M. Podlaseck, C. Narayanaswami, and M. Raghunath. Using symbiotic displays to view sensitive information in public. *IEEE International Conference on Pervasive Computing and Communications*, pages 139–148, 2005.

[3] D. Butler. US intelligence exposed as student decodes Iraq memo. *Nature*, 429:116, 2004.

[4] I. Daubechies. *Ten Lecture Notes on Wavelets*. SIAM, Philadelphia, Pennsylvania, 1992.

[5] Z. Fan and R. de Queiroz. Identification of bitmap compression history: Jpeg detection and quantizer estimation. *IEEE Transaction of Image Processing*, 12:230–235, 2003.

[6] J. Fridrich, M. Goljan, and R. Du. Steganalysis based on jpeg compatibility. *SPIE Multimedia Systems and Applications*, pages 275–280, 2001.

[7] S. Goldwasser and S. Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28:270–299, 1984.

[8] D. Johnson. Redacting pdf files: A survey of tools. *Adobe Acrobat User community Newsletter*, (http://www.acrobatusers.com/), 2006.

[9] D. Kelly and B. Foster. A process for electronic document redaction. *WO Patent*, (WO/2006/041318), 2006.

[10] D. Lopresti and A. L. Spitz. Quantifying information leakage in document redaction. *1st ACM workshop on Hardcopy Document Processing*, pages 63–69, 2004.

[11] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

[12] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek. An overview of jpeg-2000. *IEEE Data Compression Conference*, pages 523–541, 2000.

[13] A.C. Popescu. Statistical tools for digital image forensics. 2004.

[14] L. Quaeler, E. Charnock, and N. Dhakouani. Method and apparatus to provide a unified redaction system. *United States Patent and Trademark Office*, (Application number:20070030528), 2007.

[15] P. Sallee. Matlab jpeg toolbox. (`http://www.philsallee.com/jpegtbx/index.html`).
[16] S. Ye, Q. Sun, and E.-C. Chang. Detecting digital image forgeries by measuring inconsistencies of blocking artifact. *IEEE International Conference on Multimedia and Expo*, 2007.

A Quantization Matrices used in JPEG Compressions

The matrices shown in Fig. 8 are some of the quantization matrices generated using the Matlab JPEG Toolbox by P. Sallee. The Matlab JPEG Toolbox was used because it adheres to the JPEG specification, section K.1, thus giving the closest lossy compression behavior to any generic JPEG image file.

80	55	50	80	120	100	255	305
60	60	70	95	130	290	300	275
70	65	80	120	200	285	345	280
70	85	110	145	255	435	400	310
90	110	185	280	340	545	515	385
120	175	275	320	405	520	565	460
245	320	390	435	515	605	600	505
360	460	475	490	560	500	515	495

(a)

27	18	17	27	40	67	85	102
20	20	23	32	43	97	100	92
23	22	27	40	67	95	115	93
23	28	37	48	85	145	133	103
30	37	62	93	113	182	172	128
40	58	92	107	135	173	188	153
82	107	130	145	172	202	200	168
120	153	158	163	187	167	172	165

(b)

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

(c)

1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1

(d)

Fig. 8. Quantization matrices used in JPEG compression according to the JPEG specifications: (a) Quality = 10% (b) Quality = 30% (c) Quality = 50% (d) Quality = 100%