Detecting Doubly Compressed JPEG Images by Factor Histogram

Jianquan Yang[a], Guopu Zhu[a,*] and Jiwu Huang[b]
[a]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, GD 518055
[b]School of Information Science and Technology, Sun Yat-sen University, Guangzhou, GD 510275
Email: jq.yang@siat.ac.cn, guopu_zhu@yahoo.com, isshjw@mail.sysu.edu.cn

***Abstract*—The detection of double compression plays an important role in JPEG image forensics and steganalysis. This paper introduces a new statistic called factor histogram, which describes the distribution of the factors being related to the quantized discrete cosine transform coefficients of JPEG image. Then, two concrete schemes based on factor histogram are presented for detecting doubly compressed images and identifying primary quality factor, respectively. Our experimental results demonstrate that both of the proposed schemes perform well, which validates the applicability of factor histogram in the detection of double compression and the estimation of primary quantization parameter.**

I. Introduction

JPEG is one of the most widely used standards for compressing image data. Most image acquisition devices and processing softwares output JPEG files, and the digital images on the web are often transmitted in this format. So JPEG images are playing an increasingly important role in our daily lives. However, with the development of image processing techniques, it is now easier than ever to tamper with digital images and the tampered images are usually stored in JPEG format for distribution. Furthermore, many steganography algorithms (such as F5 [1], Outguess [2], Yass [3], etc.) store stego images also in JPEG format. In some scenarios, fake or stego images may cause serious harms if it is failed to detect them. Therefore, the authentication of JPEG images and the detection of the secret messages in JPEG images have become important issues. And many forensics and steganalysis techniques have been developed to address these issues in recent years.

The detection of double compression (DC) and the estimation of the primary quantization matrix (PQM) of doubly compressed images are hot topics in the field of JPEG image forensics and steganalysis. Double compression refers to the procedure that a natural image is compressed firstly, then the compressed image is decompressed into spatial domain and recompressed with a different quantization matrix. DC detection is usually the first step to detect forged or stego JPEG images, while PQM estimation may help to further reveal more details, such as the location of tampered areas or the length estimation of the secret message embedded in stego images. DC detection and PQM estimation have been addressed by many researchers, who have also proposed a plenty of algorithms. And these algorithms could be roughly classified into two categories: the spatial-domain-based algorithms [4]−[6] and the frequency-domain-based algorithms [7]−[14].

Among the spatial-domain-based algorithms, Luo et al. [4] proposed a characteristic matrix for blocking artifacts to detect cropped and recompressed images. Based on the work of [4], Barni et al. [5] recently proposed two algorithms to identify cut and paste tampering in forged JPEG images. Chen and Hsu [6] proposed a blocking periodicity model to analyze blocking artifacts. The idea of the spatial-domain-based algorithms is based on detecting the abnormal blocking artifacts of intra blocks. However, if there is no shift between the JPEG grids of the two sequential compressions (i.e., recompress with aligned 8×8 grids), these algorithms will fail to detect DC.

The frequency-domain-based algorithms apply the statistics of quantized discrete cosine transform (DCT) coefficients to detect DC and/or estimate PQM, requiring that the two sequential compressions are with aligned 8×8 grids. Lukas and Fridrich [7] studied the histogram shape of the quantized DCT coefficients in doubly compressed image and proposed three schemes for estimating primary quantization step (PQS). The first two of their schemes are non-training-based, and the last one is a training-based scheme that adopts neural network. Pevny and Fridrich [8] trained a SVM with 144-D features for DC detection, and designed several SVM-based multi-classifiers for PQS estimation. Popescu and Farid [9] observed that the histogram of quantized DCT coefficients in doubly compressed image appears periodic artifacts, so they proposed a measure to evaluate the periodicity in histogram and applied it for DC detection. Fu et al. [10] presented a novel statistical model called generalized Benford's law to study JPEG compression. They reported that the distribution of the first digits of the quantized DCT coefficients in doubly compressed image no longer follows the generalized Benford's law, which could be used as a clue to detect DC. Li et al. [11] improved and extended the work in [10] by proposing mode based first digit features for DC detection and the estimation of primary quality factor (PQF). Chen and Hsu [12] combined the features in frequency domain with that in spatial domain for DC detection, which makes their proposed scheme work well in both JPEG grids aligned and non-aligned situations. Besides, several techniques have been developed to locate the tampered regions in forged JPEG images via DCT coefficient analysis [13], [14]. Since frequency-domain-based algorithms

This work was supported by the 973 Program (No. 2011CB302204) and the National Natural Science Foundation of China (No. 61003297).
* Corresponding author.

have potentials to detect DC as well as estimate PQS/PQF, they seem to attract more attentions in recent years.

This paper introduces a new statistic called factor histogram by analyzing the procedure of double quantization. Factor histogram describes the distribution of the factors being related to quantized DCT coefficients. By theoretically analyzing its characteristics, we find that factor histogram can be applied to DC detection and PQM estimation, and then design two schemes to further investigate the advantages and drawbacks of factor histogram. Our experimental results show that the proposed schemes can achieve satisfactory performance even when the tested image is quite small or is compressed by multiple times, which indicates that factor histogram can be used as a feasible and effective technique for JPEG DC detection and PQM estimation.

The rest of the paper is organized as follows. Section II derives the concept of factor histogram by analyzing the procedure of the double quantization in JPEG DC, and Section III describes the details of the proposed DC detection and PQM estimation schemes. Experimental results are presented in Section IV. Finally, we conclude our work in Section V.

II. Factor Histogram

JPEG DC makes the DCT coefficients undergo double quantization. In this section, we derive the concept of factor histogram by theoretically analyzing the procedure of double quantization.

Fig. 1 shows the procedure of JPEG DC. In order to highlight double quantization in JPEG DC, some operations, such as entropy encoding and decoding, and etc., are omitted. As shown in Fig. 1, the original coefficient c_0 is quantized by the step size q_1 to generate c_1, then the quantized coefficient c_1 undergoes a series of operations, which include dequantization, inverse DCT, rounding and/or truncating, DCT, and then becomes into c_1q_1+e, where the term e denotes the error due to the rounding and/or truncating operations. Finally, the whole term c_1q_1+e is quantized again by the step size q_2 to form the coefficient c_2. Note that the input and output data in Fig. 1 are in form of matrix in practice, but for clearly describing and analyzing the procedure of double quantization, we adopt the above-mentioned scalar notations.

According to the DC procedure shown in Fig. 1, the quantized coefficient c_2 can be expressed as

$$c_2 = \left\lfloor \frac{c_1q_1+e}{q_2} \right\rceil \tag{1}$$

where $\lfloor\cdot\rceil$ denotes the round operator, and the quantization step size q_1 and q_2 are positive integers. In order to derive a clear relationship among c_1, q_1, c_2 and q_2, we ignore the error term e. According to the rule of round operator, it can be obtained that [13],

$$c_2 - 0.5 \le c_1q_1/q_2 < c_2 + 0.5 \tag{2}$$

From (2), we can further get the value range of c_1q_1:

$$(c_2-0.5)q_2 \le c_1q_1 < (c_2+0.5)q_2 \tag{3}$$

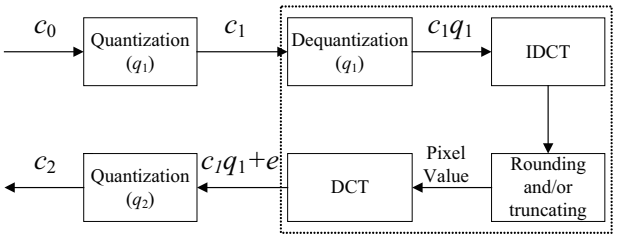


Fig. 1. Procedure of JPEG double compression. Some operations are omitted in order to highlight double quantization.

The above range of c_1q_1 includes q_2 consecutive integers, which can be collected to form a set $D(c_2,q_2)$. The set $D(c_2,q_2)$ is determined by c_2 and q_2, and can be described by

$$D(c_2,q_2) = \{\lceil (c_2-0.5)q_2\rceil + x \mid x = 0,1,\cdots,q_2-1\} \tag{4}$$

where $\lceil\cdot\rceil$ denotes the ceiling operator. Then, according to (3) and (4), we have

$$c_1q_1 \in D(c_2,q_2) \tag{5}$$

Observing that q_1 is one of the positive factors of the term c_1q_1, we factorize each integer in $D(c_2,q_2)$, and collect all of the positive factors to form the factor set $F(c_2,q_2)$, which is given by

$$F(c_2,q_2) = \{x \mid \mathrm{mod}(y,x) = 0, y \in D(c_2,q_2), x > 0\} \tag{6}$$

where $\mathrm{mod}(\cdot,\cdot)$ denotes the modulo operator. Then we can obtain that

$$q_1 \in F(c_2,q_2) \tag{7}$$

Note that the set $F(c_2,q_2)$ can be regarded as a constraint for the value range of the step size q_1. If $c_2 = 0$, according to (4), we have $0 \in D(0,q_2)$, then, according to (6), we further have $F(0,q_2) = \mathbb{Z}^+$, which means that if $c_2 = 0$, $F(0,q_2)$ does not provide any constraint on q_1. Therefore, the case that $c_2 = 0$ is omitted in our following analysis.

As mentioned above, $D(c_2,q_2)$ consists of q_2 consecutive integers, therefore, it is obvious that

$$\{1,2,\cdots,q_2\} \subseteq F(c_2,q_2) \tag{8}$$

Especially, when $q_1 > q_2$, according to (7) and (8), we further have

$$\{1,2,\cdots,q_2,q_1\} \subseteq F(c_2,q_2) \tag{9}$$

Here we give a concrete example to show the factor set property described by (9). Given that $c_1 = 1$, $q_1 = 8$, and $q_2 = 3$, according to (1), we obtain $c_2 = \lfloor c_1q_1/q_2\rceil = 3$. Further, according to (4) and (6), we have $D(c_2,q_2) = \{8,9,10\}$ and $F(c_2,q_2) = \{1,2,3,4,5,8,9,10\}$, respectively. So it is obvious that $\{1,2,\cdots,q_2,q_1\} = \{1,2,3,8\} \subseteq F(c_2,q_2)$.

We have analyzed double quantization for a single quantized coefficient above, and some conclusions (i.e., (8) and (9)) are drawn. In the following, we consider the case of a quantized coefficient sequence. Let $\mathbf{c}_2 = \left[c_2^1, c_2^2, \cdots, c_2^n\right]$ denote a nonzero coefficient sequence of length n, in which each component has been doubly quantized with step size q_1 and q_2. For each component c_2^i $(i = 1,2,\cdots,n)$ of $\mathbf{c}_2$, we calculate $F(c_2^i,q_2)$ according to (6), thus we can get n factor sets totally. All

elements of these factor sets are collected to form a factor sequence. Then we calculate the histogram of the factor sequence, and denote the resulting histogram by h_f. Since h_f describes the distribution of the elements in the factor sequence, we call it factor histogram. We also denote the histogram of the quantized coefficient sequence $\mathbf{c}_2$ by h_c, and call it quantized coefficient histogram. It is interesting that h_f can be computed based on h_c, and h_f can be expressed by

$$h_f(u) = \sum_{x=a}^{b} h_c(x)s(u,x,q_2), 1 \le u \le r \tag{10}$$

where a and b denote the minimum and maximum in $\mathbf{c}_2$, respectively, and r is a parameter that controls the interested range, and $s(u,x,q_2)$ is given by

$$s(u,x,q_2) = \begin{cases} 1, & u \in F(x,q_2) \\ 0, & u \notin F(x,q_2) \end{cases} \tag{11}$$

Then, according to the definition of h_f described by (10) and (11), we have

$$0 \le h_f(u) \le n \tag{12}$$

where $n = \sum_{x=a}^{b} h_c(x)$ is the length of the nonzero coefficient sequence $\mathbf{c}_2$.

Up to now, we have derived the concept of factor histogram. According to (10) and (11), it can be found that the calculation of factor histogram just depends on the quantized coefficient sequence $\mathbf{c}_2$ and its current quantization step size q_2. Therefore, as long as the quantized coefficient sequence and its current quantization step size are available, the factor histogram of the sequence can be calculated in practice, no matter the sequence has been singly quantized, doubly quantized or repeatedly quantized. For the images being saved in JPEG format, its quantized coefficients and the quantization step size can be read from the image file directly; while for the images that have been JPEG compressed and then been resaved in other format, such as BMP format, its quantization step size can be estimated by some existing algorithms [15], [16], and its quantized coefficients can also be calculated after knowing the quantization step size.

In the following we analyze some characteristics of the factor histogram h_f. According to (8), (10), and (11), we have

$$h_f(u) = n, u \in \{1,2,\cdots,q_2\} \tag{13}$$

Especially, when $q_1 > q_2$, according to (9), (10) and (11), we have

$$h_f(u) = n, u \in \{1,2,\cdots,q_2,q_1\} \tag{14}$$

Eq. (13) means that factor histogram will reach its maxima at positions $1,2,\cdots,q_2$, which has nothing to do with how many times the coefficient sequence have been quantized, and thus is an inherent characteristic of factor histogram. Furhtermore, for the doubly quantized sequence with $q_1 > q_2$, the factor histogram will achieve its maximum at q_1 in addition. It is worth noting that (14) describes a quite important characteristic for DC detection and PQM estimation.

Here is an example to intuitively illustrate the conclusions mentioned above. Let $\mathbf{c}_0$ denote a coefficient sequence generated by a Gaussian process with mean 0 and standard deviation 25. Then, we quantize the sequence $\mathbf{c}_0$ in three different manners, which are single quantization with step size $q_a = 3$, double quantization with step size pair $(q'_b, q_b) = (5,3)$, and double quantization with $(q'_c, q_c) = (3,5)$, then denote the three quantized coefficient sequences by $\mathbf{c}_a$, $\mathbf{c}_b$ and $\mathbf{c}_c$, respectively. Factor histograms of the three quantization versions of $\mathbf{c}_0$ are shown in Fig. 2. The three factor histograms achieve their maxima at positions $1,2,\cdots q_a$, positions $1,2,\cdots,q_b$ and positions $1,2,\cdots,q_c$, respectively, which validates (13). Since $q'_b > q_b$, Fig. 2(b) shows that the factor histogram of $\mathbf{c}_b$ also reaches its maximum at q'_b, which validates (14). Although $\mathbf{c}_c$ has also gone through double quantization, there is no additional maximum in its factor histogram due to that $q'_c < q_c$.

III. Proposed Schemes

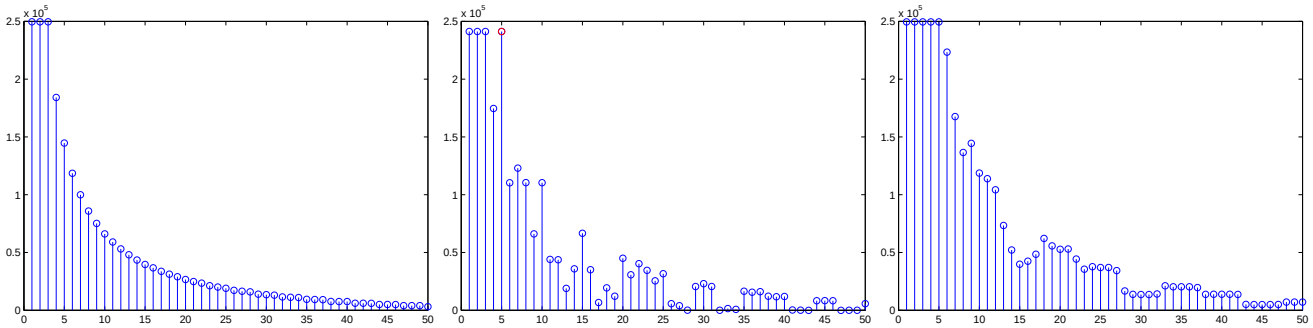
In practice, many camera manufacturers and image software developers have designed their own JPEG quantization matrixes [17], which may have some differences from each other. But, the JPEG quantization matrixes provided by the same company are usually very similar, thus can be regarded as the same quantization matrix system (QMS).

In some scenarios, the knowledge of primary QMS (i.e., the QMS used in the primary JPEG compression) can be used as auxiliary information to detect DC and estimate PQM. According to whether the knowledge of primary QMS is available or not, the problems of DC detection and PQM estimation could be classified into two categories. The first category is to detect DC and estimate PQM with the knowledge of primary QMS. Whereas the second one is to detect DC and estimate PQM without any primary QMS knowledge. If the information of the quantized coefficients in different frequencies can be integrated together by utilizing the knowledge of primary QMS appropriately, then better performance should be achieved. In this paper, we focus on solving the first category of the problems (i.e. detecting DC and estimating PQM with the knowledge of primary QMS) by using factor histogram. For simplicity, standard QMS, which is recommended by JPEG compression standard, is adopted in our analysis and experiments. And we also assume that the primary and current JPEG compression use the same QMS. The quality factor ranging from 1~100 is used to represent the compression quality. User can adjust quality factor to make a tradeoff between fidelity and compression rate. In the following, we will describe our proposed schemes in detail.

A. Detecting JPEG Double Compression

Let F denote the quality factor in standard QMS, and $\mathbf{Q} = \left[q^{ij}\right]_{8\times 8}$ denote the quantization matrix corresponding to F. For a given JPEG image, we first extract the quantized DCT coefficients and quantization matrix from the image file, and calculate the factor histogram for each frequency according to (10) and (11). We denote the factor histogram of the frequency (i,j) by h_f^{ij}. Then, we define the following statistic for DC detection

$$M(F) = \frac{\sum_{(i,j)\in L} h_f^{ij}(q^{ij})}{\sum_{(i,j)\in L} h_f^{ij}(1)}, F = 1,2,\cdots,100 \tag{15}$$



(a) (b) (c)

Fig. 2. Factor histogram of different quantized sequences (a) singly quantized with step size 3 (b) doubly quantized with step size pair (5,3) (c) doubly quantized with step size pair (3,5)

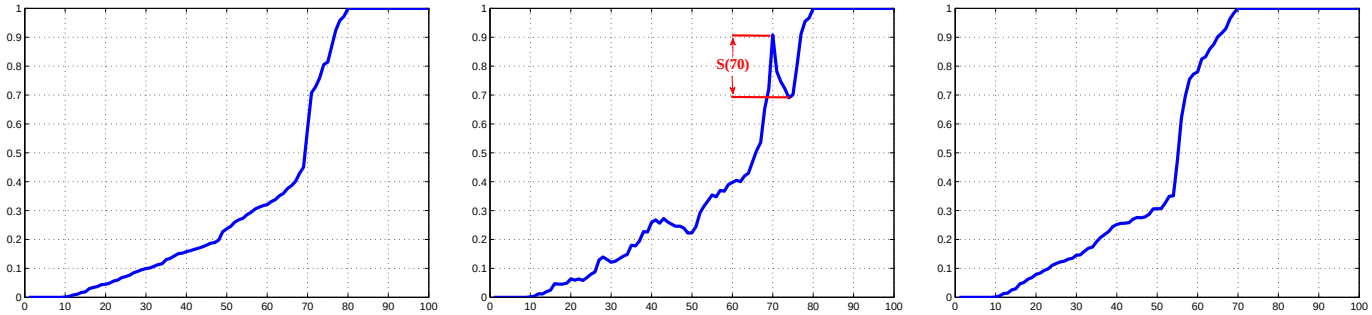
where L denotes the set of the interested frequencies. Through (15), we unify the information of the factor histograms of different frequencies together to calculate the statistic $M(F)$. We name the statistic $M(F)$ as the factor matching degree (FMD) of F for $F = 1,2,\cdots,100$, and name the curve formed by the value of $M(F)$ at each position F as FMD curve, due to $M(F)$ has the following characteristics:

1) Since $0 \le h_f^{ij}(q^{ij}) \le h_f^{ij}(1)$ for $1 \le i,j \le 8$, according to (15), we have $0 \le M(F) \le 1$.
2) Let F_c denote the quality factor of current compression, and $\mathbf{Q}_c$ denote its corresponding quantization matrix. According to (13), it is obtained that $h_f^{ij}(u) = n^{ij}$, $u \in \{1,2,\cdots,q_c^{ij}\}$, $1 \le i,j \le 8$, where n^{ij} denotes the number of nonzero coefficients at frequency (i,j). And we also notice that for standard QMS, when $F_c \le F \le 100$, $q_c^{ij} \ge q^{ij}$. Thus, for $F_c \le F \le 100$, we have
$$M(F) = \frac{\sum_{(i,j)\in L} h_f^{ij}(q^{ij})}{\sum_{(i,j)\in L} h_f^{ij}(1)} = \frac{\sum_{(i,j)\in L} n^{ij}}{\sum_{(i,j)\in L} n^{ij}} = 1 \qquad (16)$$
which means that FMD curve reaches its maxima at positions $F_c, F_c+1,\cdots,100$.
3) If the detected image is doubly compressed with quality factor pair (F_p,F_c), and $F_p < F_c$. In this case, we have $q_p^{ij} \ge q_c^{ij}$ for $1 \le i,j \le 8$. Then, according to (14), $h_f^{ij}(q_p^{ij}) = n^{ij}$. Thus we have
$$M(F_p) = \frac{\sum_{(i,j)\in L} h_f^{ij}(q_p^{ij})}{\sum_{(i,j)\in L} h_f^{ij}(1)} = \frac{\sum_{(i,j)\in L} n^{ij}}{\sum_{(i,j)\in L} n^{ij}} = 1 \qquad (17)$$
which means that the FMD curve of a doubly compressed image with $F_p < F_c$ also reaches its maximum at F_p. In summary, if a given image is doubly compressed with quality factor pair (F_p,F_c), and $F_p < F_c$, the FMD curve of the given image reaches its maxima at positions $F_p, F_c, F_c+1,\cdots,100$.

We have observed from a lot of experiments that the FMD curves of doubly compressed images show two different characteristics, which are determined by the relationship between F_p and F_c. If $F_p < F_c$, the FMD curve will have a local peak at F_p, and is quite distinct from the FMD curve of singly compressed image. However, if $F_p > F_c$, the FMD curve is similar to that of singly compressed image. For instance, given an uncompressed image, let it undergo a single compression with quality factor $F_c = 80$, a double compression with quality factor pair $(F_p,F_c) = (70,80)$, and a double compression with quality factor pair $(F_p',F_c) = (80,70)$, respectively, then we get three compressed images. The FMD curves corresponding to the three compressed images are shown in Fig. 3. All of the three FMD curves reach the maxima at their positions $F_c, F_c+1,\cdots,100$. Particularly, the FMD curve corresponding to the double compression with $F_p < F_c$ (as shown in Fig. 3(b)) has a local peak at position F_p. According to (17), the FMD curve should reach its maximum at this position. But, in practice, $M(F_p)$ suffers a slight decrease inevitably due to the rounding and truncating errors. However, the FMD curve corresponding to the double compression with $F_p' > F_c$ (as shown in Fig. 3(c)) not only has no obvious local peak, but also is very similar to that corresponding to the single compression (as shown in Fig. 3(a)), which indicates that FMD curve is not suitable for distinguishing double compression with $F_p' > F_c$ from single compression.

In the following, we will calculate the height of the above-mentioned local peak to measure the smoothness of FMD curve. We first define a statistic as follows
$$S(F) = M(F) - \min\{M(i) \mid i = F+1, F+2,\cdots,100\} \qquad (18)$$
Then, the height of local peak can be calculated by
$$S_p = \max_F \{S(F)\} \qquad (19)$$
Because that the FMD curve of singly compressed images decrease from right to left (as shown in Fig. 3(a)), thus S_p is usually zero for singly compressed image; while S_p takes positive value (as shown in Fig. 3(b)) for doubly compressed image, especially for the case that $F_p < F_c$. Then, we choose a threshold t to make a binary decision for DC detection. That is, if $S_p > t$, the detected image will be regarded as a doubly compressed image; otherwise, it will be regarded as a singly compressed one.

Based on the above descriptions, the proposed DC detection scheme can be summarized by the following steps:



(a) (b) (c)

Fig. 3. FMD curves of three compressed versions (a) single compression with quality factor 80 (b) double compression with quality factor pair (70,80) (c) double compression with quality factor pair (80,70)

1) For a given JPEG image to be detected, extract its quantized DCT coefficients and current quantization matrix to calculate the factor histogram h_f^{ij} for each frequency $(i,j) \in L$;
2) According to (15), calculate the factor matching degree $M(F)$ for $F = 1,2,\cdots,100$, then according to (18) and (19) , calculate S_p;
3) If $S_p > t$, the detected image is classified as a doubly compressed image; otherwise, it is classified as a singly compressed one.

B. Estimating Primary Quality Factor

We have explained in section III-A that the FMD curve corresponding to the doubly compression with $F_p < F_c$ usually has a local peak at position F_p. By detecting the position of the local peak, we can estimate F_p by

$$\hat{F}_p = \underset{1 \le F \le F_c, S(F)>0}{\operatorname{argmax}} M(F) \tag{20}$$

Note that $\hat{F}_p$ can be used as the estimation of F_p only when $F_p < F_c$. In practice, we may need to confirm whether the condition $F_p < F_c$ is satisfied or not when we use the proposed scheme to estimate PQF. This can be performed by checking whether $M(\hat{F}_p)$ is close to 1 or not.

For simplifying our analysis and description, we have assumed that the primary and current JPEG compression adopt the same QMS. However, the proposed schemes with minor modifications can also perform well even if the two consequent compressions adopt different QMS.

IV. Experimental Results

In order to test the performances of the proposed schemes, we setup several image sets for our experiments. First, 2000 color images are randomly selected from COREL and NRCS image sets. Then these color images are converted to grayscale, and center-cropped into smaller images of sizes 512×512, 256×256,$\cdots$, 8×8. Thus, there are totally 7 natural image sets with different sizes. In our experiments, we adopt Matlab image processing toolbox to implement JPEG compression, and use JPEG toolbox [18] to extract the quantized DCT coefficients and the quantization matrix for calculating factor histogram. And the parameters of the proposed schemes are set as $a = -128$, $b = 128$, $r = 50$, and L takes the first 30 frequencies in Zigzag order.

The first scheme proposed by Lukas and Fridrich in [7] is used for comparison. Lukas and Fridrich's (shortened as L&F's) scheme estimates primary quantization steps by producing a bench image with the cropping and recompressing operations and then performing compatibility test. As the authors of [7] said, it can be easily extended to detect DC and estimate PQM with the knowledge of primary QMS.

A. Experimental Results on DC Detection

We first compare the performance between the proposed and L&F's schemes on DC detection. For obtaining appropriate thresholds for the proposed scheme, each of the 7 natural image sets is divided into two image subsets of the same number (i.e., each subset has 1000 images). The first image subset is used to determine the threshold t, while the second one is used to test the performances of the two compared schemes. For each image in the first image subset, we randomly selected two quality factors F_p and F_c from $\{50,51,\cdots,95\}$ with the constraint $|F_p - F_c| \ge 5$, then compress the image by a double compression with quality factor pair (F_p,F_c) and a single compression with quality factor F_c, respectively. After the feature S_p of the proposed DC detection scheme is extracted from each image of the first image subset, we then determine the thresholds by minimizing the classification error. Finally, we get 7 thresholds corresponding to the 7 different image sizes.

In the test stage, each image in the second image subset is singly compressed with quality factor F_c and doubly compressed with quality factor pair (F_p,F_c), respectively. (F_p,F_c) is randomly selected from $\{50,51,\cdots,95\}$with the restriction $|F_p - F_c| \ge 5$. We classify the feature S_p of these singly and doubly compressed images by the determined thresholds and then calculate the accuracy of classification. For showing the performance of the two compared schemes in more detail, we present the experimental results, respectively, for the two cases that $F_p < F_c$ and $F_p > F_c$.

It is clearly shown from Table I that the proposed scheme performs quite well in the case that $F_p < F_c$. When the size of the test image is relatively small, the proposed scheme can obtain higher detection accuracy than L&F's scheme.

TABLE I
DETECTION ACCURACY (%) OF DOUBLE COMPRESSION
(THE BETTER RESULTS ARE HIGHLIGHTED BY RED AND BOLD FONT)

Size of image		512×512	256×256	128×128	64×64	32×32	16×16	8×8
L&F's	$F_p < F_c$	**98.8**	**97.2**	**95.6**	88.8	81.7	75.0	-
scheme	$F_p > F_c$	**98.1**	**92.5**	**73.8**	**63.8**	**54.1**	48.3	-
Proposed	$F_p < F_c$	97.1	96.6	94.4	**90.0**	**87.5**	**80.7**	**76.0**
scheme	$F_p > F_c$	67.4	64.1	54.4	50.2	49.9	**48.7**	**48.5**

TABLE II
ESTIMATING ACCURACY (%) OF THE PRIMARY QUALITY FACTOR OF DOUBLE COMPRESSION
(THE BETTER RESULTS ARE HIGHLIGHTED BY RED AND BOLD FONT)

Size of image		512×512	256×256	128×128	64×64	32×32	16×16	8×8
L&F's	$A(\pm 0)$	**90.5**	**90.8**	87.1	81.6	61.9	18.8	-
scheme	$A(\pm 1)$	98.9	98.7	96.1	91.6	80.2	36.3	-
Proposed	$A(\pm 0)$	89.5	89.3	**88.0**	**84.3**	**76.5**	**62.6**	**40.1**
scheme	$A(\pm 1)$	**99.0**	**98.8**	**98.0**	**95.8**	**89.6**	**79.0**	**56.4**

TABLE III
ESTIMATION ACCURACY (%) OF THE SECONDARY QUALITY FACTOR OF TRIPLE COMPRESSION
(THE BETTER RESULTS ARE HIGHLIGHTED BY RED AND BOLD FONT)

Size of image		512×512	256×256	128×128	64×64	32×32	16×16	8×8
L&F's	$A(\pm 0)$	65.5	65.3	63.4	55.9	38.9	11.3	-
scheme	$A(\pm 1)$	74.3	74.6	73.6	67.7	54.2	23.4	-
Proposed	$A(\pm 0)$	**86.2**	**85.5**	**84.2**	**80.8**	**72.7**	**57.2**	**36.9**
scheme	$A(\pm 1)$	**95.5**	**95.3**	**94.7**	**92.7**	**87.6**	**74.6**	**52.2**

Especially, when the size of the test image decreases to 8×8, L&F's scheme fails to work because that the cropping and recompressing operations cannot be applied to a single 8×8 image block, but the detection accuracy of the proposed scheme is up to 76%. As mentioned in section III-A, the proposed scheme is likely unable to detect DC for the case that $F_p > F_c$. But we can found from Table I that the proposed scheme can also detect the doubly compressed image with $F_p > F_c$ to some extent. In summary, the proposed DC detection scheme is effective in distinguishing double compression with $F_p < F_c$ from single compression, even when the size of the detected image is very small.

B. Experimental Results on PQF Estimation

In this sub-section, we evaluate the performance of the proposed scheme on PQF estimation. For each image in our image sets, we doubly compress it by (F_p, F_c) , where F_p and F_c are randomly selected from $\{50, 51, \cdots, 95\}$ with the constriction of $F_c - F_p \geq 5$. In order to make a fair comparison, we modify L&F's scheme to make it use the prior knowledge of $F_p < F_c$. Here, $A(\pm 0)$ and $A(\pm 1)$ are introduced to measure the accuracy in estimating PQF. $A(\pm 0)$ denotes the percentage of the cases that the estimated PQF is equal to the true PQF, while $A(\pm 1)$ denotes the percentage of the cases that the absolute difference between the estimated PQF and the true PQF is not greater than 1.

As shown in Table II, when the size of the detected image is relatively large, the two compared schemes have almost the same performance on estimating PQF. With the size of the detected image decreases, the proposed PQF estimation scheme obtains higher estimation accuracy than L&F's. Especially, for the images of size 16×16, the $A(\pm 1)$ of L&F's scheme is only 36.3%, while the $A(\pm 1)$ of the proposed scheme is 79%, and for the images of size 8×8, L&F's scheme fails to work, but the proposed scheme can still work and its $A(\pm 1)$ reaches 56.4%. These experimental results show that the local peak of FMD curve is a quite robust feature for the PQF estimation in the case that $F_p < F_c$.

To further show the good performance of the proposed PQF estimation scheme, we set up 7 image sets consisting of triply compressed images, and then attempt to estimate the secondary quality factors of the triply compressed images. It should be pointed out that, for a triple compression, if we regard the first compression as a disturbing operation, then the last two compressions can be regarded as a double compression disturbed by the first compression, and thus the quality factor using in the second compression can be also regarded as the PQF of the disturbed double compression. In this experiment, we first randomly select a pair of quality factor (F_p, F_c) from $\{50, 51, \cdots, 95\}$ with the constraint $F_c - F_p \geq 5$ and a parameter d from $\{1, 2, \cdots, 20\}$, then let $F'_p = F_p - d$, finally we compress each image in our testing image sets triply with quality factor group (F'_p, F_p, F_c). Note that the quality factor group (F'_p, F_p, F_c) in this experiment satisfies the condition that $F'_p < F_p < F_c$.

By comparing the experimental results in Table III with

that in Table II, it can be found that the first compression disturbs the estimation of the secondary quality factor in some degree. The estimation accuracy of L&F's scheme decreases significantly, however, that of the proposed scheme does not decrease too much. For the images of size 512×512, the $A(\pm 0)$ of L&F's scheme reaches 90.5% in the case of double compression, while is only 65.5% in the case of triple compression, which indicates that L&F's scheme is very sensitive to disturbances, such as a pre-compression before double compression. The proposed scheme performs much better than L&F's scheme in the case of triple compression. Even for the images of size 64×64, the $A(\pm 0)$ of the proposed scheme also reaches up to 80%, which shows that the feature extracted from factor histogram is robust to serious disturbances.

V. Conclusions

In this paper, we have derived the concept of factor histogram by theoretically analyzing the procedure of the double quantization in JPEG double compression. By investigating the characteristics of factor histogram, we found that the factor histogram of a doubly quantized sequence with primary quantization step size larger than current quantization step size likely has obvious artifact, which can be used as a clue to detect double quantization and to estimate primary quantization parameter. We have proposed two schemes to validate the applicability of factor histogram. The experimental results show that the proposed schemes have satisfactory performances even when the detected images are of small size and are compressed several times, which means that factor histogram has great potential in detecting double compression and estimating primary quality factor.

References

[1] A. Westfeld, "F5 - a steganographic algorithm: high capacity despite better steganalysis," Information Hiding, 4th International Workshop, ser. Lecture Notes Comput. Sci., 2001, vol. 2137, pp. 289–302.

[2] N. Provos, "Defending against statistical steganalysis," in Proc. 10th USENIX Security Symp., 2001, pp. 323–335.

[3] K. Solanki, A. Sarkar, and B. S. Manjunath, "YASS: yet another steganographic scheme that resists blind steganalysis," in Proc. 9th Int. Workshop on Information Hiding, Saint Malo, France, 2007, vol. 4567, pp. 16–31.

[4] W. Luo, Z. Qu, J. Huang, and G. Qiu, "A novel method for detecting cropped and recompressed image block," Proc. ICASSP, vol.2, April 2007, pp. 217-220.

[5] M. Barni, A. Costanzo, and L. Sabatini, "Identification of cut & paste tampering by means of double-JPEG detection and image segmentation," in Proc. of ISCAS 2010, 2010, pp. 1687-1690.

[6] Y. L. Chen and C. T. Hsu, "Image tampering detection by blocking periodicity analysis in JPEG compressed images," Proc. MMSP, 2008.

[7] J. Lukas and J. Fridrich, "Estimation of primary quantization matrix in double compressed JPEG images," in Proc. Digital Forensic Research Workshop, Cleveland, Ohio, August 2003.

[8] T. Pevny and J. Fridrich, "Detection of double-compression in JPEG images for application in steganography," IEEE Trans. Inf. Forensics Security, June 2008, vol. 3, no. 2, pp. 247-258.

[9] A. C. Popescu, "Statistical tools for digital image forensics," Ph.D. dissertation, Dartmouth College, Hanover, NH, Dec. 2004.

[10] D. Fu, Y.Q. Shi, and W. Su, "A generalized benford's law for JPEG coefficients and its applications in image forensics," in Proc. SPIE, Security, Steganography and Watermarking of Multimedia Contents IX, San Jose, USA, January 2007.

[11] B. Li, Y. Q. Shi, and J. Huang, "Detecting doubly compressed JPEG images by mode based first digit features," Proc. MMSP, 2008.

[12] Y. L. Chen and C. T. Hsu, "Detecting recompression of JPEG images via periodicity analysis of compression artifacts for tampering detection," IEEE Trans. Inf. Forensics Security, June 2011, vol. 6, no. 2, pp. 396-406.

[13] J. He, Z. Lin, L. Wang, and X. Tang, "Detecting doctored JPEG images via DCT coefficients analysis," Proc. European Conference on Computer Vision, Graz, Austria, 2006.

[14] T. Bianchi, A. D. Rosa, and A. Piva, "Improved DCT coefficient analysis for forgery localization in JPEG images," in Proc. of ICASSP 2011, May 2011, pp. 2444-2447.

[15] W. Luo, J. Huang, and G. Qiu, "JPEG error analysis and its applications to digital image forensics," IEEE Trans. Inf. forensics and security, vol. 5, no. 3, Sep. 2010, pp. 480-491.

[16] Z. Fan and R. de Queiroz, "Identification of bitmap compression history: JPEG detection and quantizer estimation", IEEE Trans. Image Processing, vol. 12, no. 2, Feb. 2003, pp. 230-235.

[17] Predefined quantization matrix used in JPEG compression. [Online], Available: http://www.impulseadventure.com/photo/jpeg-quantization.html

[18] P. Sallee, Matlab JPEG toolbox 1.4, [online], Available: http://www.philsallee.com/jpegtbx/index.html.