一种估计 JPEG 双重压缩原始量化步长的新方法

王俊文[①] 刘光杰[①] 戴跃伟[①] 周琳娜[②] 郭云彪[②]
[①](南京理工大学自动化学院 南京 210094)
[②](北京电子技术应用研究所 北京 100091)

摘 要：该文提出了一种双重压缩后 JPEG 图像的原始量化步长的估计方法。该方法根据两次量化步长之间的大小关系分 3 种情况进行讨论。当原始量化步长大于第 2 次量化步长，提出了直接利用直方图计算的新方法；为解决原始量化步长是第 2 次量化步长因子和傅里叶频谱分析中的多值问题，提出了采用 0.98 缩放来近似未压缩图像的方法。本文方法能给出第二次量化步长为一次量化步长倍数时的估计，并利用频谱分析的结果降低了计算的复杂度，实验结果表明本文方法有较高的估计准确度。

关键词：图像处理；双重量化；直方图；重构；匹配
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2009)04-0836-04

A New Method for Estimating the Primary Quantization Step of JPEG Double-Compression

Wang Jun-wen[①] Liu Guang-jie[①] Dai Yue-wei[①] Zhou Lin-na[②] Guo Yun-biao[②]
[①](*School of Automation*, *Nanjing University of Science & Technology*, *Nanjing* 210094, *China*)
[②](*Beijing Application Institute of Electronic Technology*, *Beijing* 100091, *China*)

Abstract: This paper presents a method for estimating the primary quantization step of double compressed JPEG image. The method in accordance with the relationship between the two quantization steps contains three situations. When the primary quantization step is larger than the second quantization step, the histogram can be used to calculate directly. To resolve the issue of multi-value caused by Fourier spectrum analysis and the problem that the primary quantization step is the factor of the second quantization step, the uncompressed image can be approximated by narrowing the image by 98%. This method can resolve the problem that the primary quantization step is the factor of the second quantization step, and uses spectrum analysis to reduce the complexity. The experimental results show that this method has higher accuracy of the estimation.

Key words: Image processing; Double quantization; Histogram; Reconstruction; Matching

1 引言

高质量数码相机的普及和功能日益强大的图像处理软件的广泛应用，使得人们不需要特殊的专业技术即可对数字图像进行非常逼真的篡改，且篡改和伪造的效果很难通过人眼分辨。伪造的数字图像可能成为事实证据用于法庭举证、新闻报道、学术论文发表等场合，其所导致的误判、误报道和欺诈等问题会引起难以估量的损失。因此，数字图像的真实性和完整性鉴别是迫切需要的技术。

传统的数字水印和数字签名技术，通过在图像中或图像外附加额外信息以辅助对数据图像的完整性和真实性认证，可以看作为一类“主动”的策略。但由于水印与签名技术均需要图像采集设备的支持，且存在着水印容易受到攻击，签名容易被丢弃等问题，因此仅依赖数字图像本身进行认证的取证技术成为另外一种更实用的选择。当前这一技术取得了业界的广泛重视，已经报道了许多新颖、有效的取证方法。读者可参考 Sencar 和 Memon[1]以及黄继武等人[2]对取证研究现状的综述。

数字图像多以JPEG这种事实的图像压缩标准进行存储，针对JPEG图像的空域篡改处理后往往会再进行JPEG压缩，以保存成JPEG格式。Fan和Fridrich对BMP图像是否经过JPEG压缩进行了分析[3,4]。在文献[5,6]中，Fridrich提出了判断图像是否经过双重压缩并估计量化因子的算法，Lukas和Fridrich对文献[5,6]的方法进行了改进[7]，改进的方法通过试探的方式重压缩剪切后的JPEG图像，并根据直方图的比较获得JPEG量化矩阵的估计。该方法需要在一个较大的搜索空间内遍历所有可能的量化矩阵，因此计算量较大，算法复杂度较高。同时由于使用裁减的方式获得原有DCT系数分布的估计分布，也会导致估计的准确率下降。Lukas和Fridrich同时提出了用神经网络来估计量化因子的方法[7]，该方法需要大量的样本来训练神经网络参数，不能

2007-12-25 收到，2008-05-12 改回
江苏省自然科学基金(BK2008403)，中国博士后基金(20070421017)，中国信息安全产品评测认证中心“115”基金和江苏省“青蓝工程”中青年学术带头人培养基金资助课题

快速方便的进行估计。Popescu和Farid[8]通过数学推理分析了双重压缩的周期效应,利用DCT系数直方图的Fourier变换实现原始量化步长的估计,但Popescu的方法无法实现第2次量化步长是第1次量化步长的倍数的情况,并且不同的第1次量化步长,相同的第2次量化步长可能会产生相同周期,只用直方图频谱来估计会出现错误。He和Lin等[9]分析双重压缩效应并利用双重压缩实现图像篡改的检测,但他们同时提出当前缺乏一种对图像的原始量化步长的简单估计算法。

为此,本文提出了一种简单快速的JPEG原始量化步长估计方法。该方法依赖直方图特性来确定所使用的具体估计方法。当直方图幅值出现间断的零值时,按照原始量化步长小于第2次量化步长的方式估计;当直方图幅值无间断零值出现时,按照直方图是否周期振荡进行分类估计。所提方法能简单快速地实现原始量化步长的估计,实验结果准确率高。

2　问题描述

原始量化步长的估计一般是服务于图像的被动认证需求,很多情况下只需要估计出若干低频位置上所使用的量化步长。

假设 JPEG 图像大小为 $M\times N$,由于每块的 64 个 DCT 经过初次量化过后,大部分的高频系数基本是 0,准确估计高频位置上的量化步长较为困难。本文选取每块 (I,J) 位的低频交流系数组成矩阵 $\boldsymbol{D}_{IJ}^2$,其大小为 $\lceil M/8\rceil\times\lceil N/8\rceil$,($\lceil\ \rceil$ 为上取整)。矩阵中的元素为 d_{ij}^2 。(1,2)位第 2 次 JPEG 量化步长为 Q^2 ,可以通过读取 JPEG 文件中存储的量化表获得。

设 JPEG 压缩过程的量化采用四舍五入的取整方式,则可知

$$d_{ij}^2=\left\lfloor\frac{Q^1d_{ij}^1}{Q^2}\right\rceil \tag{1}$$

其中[·]表示四舍五入取整,原始量化步长 Q^1 的估计问题归结为:

已知 $\boldsymbol{D}_{IJ}^2$, Q^2 , 求: Q^1 (2)

单纯依赖一个量化值 d_{ij}^2 估计出 Q^2 是不可能的,但是通过 $\boldsymbol{D}_{IJ}^2$ 的统计信息可估计出原始的量化步长。本文使用 $\boldsymbol{D}_{IJ}^2$ 的一阶统计信息——量化系数的直方图构造原始步长的估计方法。

3　基于直方图特征的估计方法

3.1 当 $Q^1>Q^2$

定理 1　当 $Q^1>Q^2$ 时,如式(1)的双重压缩产生的 $\boldsymbol{D}_{IJ}^2$ 的直方图 H^2 满足式(3):

$$\sum_{u=0}^{Q^1}\mathrm{sgn}\left(H_{IJ}^2[u]\right)=Q^2+1 \tag{3}$$

其中函数 $\mathrm{sgn}(x)$ 为符号函数(当 $x>0$, $\mathrm{sgn}(x)=1$;当 $x\le 0$, $\mathrm{sgn}(x)=0$)。

证明　根据式(1), $d_{ij}^2=[k\cdot(Q^1/Q^2)]$ (k=0,1,2,…),当 k 在 $[0,Q^2]\cap\boldsymbol{Z}$ 上取值时, d_{ij}^2 在 $[0,Q^1]\cap\boldsymbol{Z}$ 上取值,且当 $k=Q^2$ 时, $d_{ij}^2=Q_1$ 。由于 $Q^1>Q^2$, $[0,Q^1]\cap\boldsymbol{Z}$ 上的 d_{ij}^2 必然有 Q^1-Q^2 个整数取不到值。由此可知 d_{ij}^2 对应的直方图 $H_{IJ}^2[u]$ 在 $u\in[0,Q^1]\cap\boldsymbol{Z}$ 上必有 Q^2+1 处有值。故式(3)成立。

以 512×512 的灰度 lena 图像为例,图 1 给出了 $Q^1=9$, $Q^2=5$ 的双重压缩实验结果。

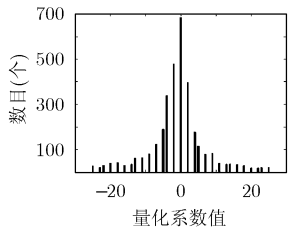

图 1　$Q^1=9,Q^2=5$ 双重量化低频系数直方图

根据定理 1,我们给出当 $Q^1>Q^2$ 时的原始量化步长的估计方法如下(设估计位置为低频的(1,2),其他位置的量化步长可依此处理):(1)根据获取的双重压缩 JPEG 图像,统计系数 D_{12}^2 的直方图 H^2 ;(2)根据 H^2 ,若 $\sum_{u=0}^{Q^2}\mathrm{sgn}\left(H_{IJ}^2[u]\right)<Q^2+1$ 则进入步骤(3),否则按照 $Q^1<Q^2$ 情形处理;(3)计算满足公式 $\sum_{u=0}^{Q^X}\mathrm{sgn}\left(H_{IJ}^2[u]\right)=Q^2+1$ 的 Q^X ,令 $Q^1=Q^X$ 。

3.2 当 $Q^1>Q^2$

当 $Q^2>Q^1$,双重压缩后的 DCT 系数在直方图上不会出现某些“值缺失”现象,因此没有类似于定理 1 那样明确的原始量化步长的估计方法。但是当 Q^2 不是 Q^1 的整数倍的时候,直方图会显现出一定的周期性,依赖这一特征可大致估算出可能的原始量化步长,并进一步通过试探估计原始量化步长。当 Q^2 是 Q^1 的整数倍的时候可取 Q^2 的因子进行试探计算以估计原始量化步长。下面分别对这两种情况进行讨论。

3.2.1 当 Q^2 为 Q^1 的整数倍　当 $Q^2=kQ^1$,先 Q^1 量化后 Q^2 得到的直方图与只进行一次 Q^2 量化的直方图相比,两者有明显的差别,双重量化的直方图更为平缓。这主要是因为 Q^1 量化过后,再进行 Q^1 反量化,会放大了一些系数的值。假设原始 DCT 系数 0,1,2,3,4,进行 $Q^1=4$ 的量化则 0,1 量化为 0;而 2,3,4 量化为 1,进行 $Q^1=4$ 反量化后进行 $Q^2=8$ 的量化,则原始系数中 0,1 量化为 0,而 2,3,4 量化为 1,如果是将原始系数直接用 $Q^2=8$ 量化,则 0,1,2,3 量化为 0。而只有 4 量化为 1。整数倍情况下的这种双重量化直方图趋缓效应可用来估计原始的量化步长。

但是,由于 Q^2 是 Q^1 的整数倍,因此仅仅从双重量化后直方图无法获取原始量化步长的信息。这里我们利用提出一种获取原始未压缩图像近似版本的方法,首先将待估计的图像转化成 BMP 图像,对 BMP 图像进行 0.98 比例的缩小。

对经过缩小的 BMP 图像进行 Q^X 接连 Q^2 的双重压缩，并将统计得到的 JPEG 图像的 DCT 系数 D'_{IJ} 的直方图 H'，通过将得到的直方图同从待估计的 JPEG 图像获取的直方图 H^2 进行比较可判断 Q^X 是否是原始的量化步长。

利用上面描述的方法进行原始量化步长的估计建立在 BMP 与原始未压缩图像的近似程度。图 2 给出了这种技术的实际效果。图 2(a)是原始 512×512 的灰度 Lena 图直接进行 $Q=8$ 量化的直方图。图 2(b)原始图进行 $Q^1=4,Q^2=8$ 量化，再将图像进行 0.98 的缩小，对其进行 $Q^2=8$ 的量化得到的直方图。从图中可以看到 0.98 缩小的处理结果是一种较好的对原始未压缩图像的近似。

为描述直方图之间的相似性，本文定义了如下的直方图相似性函数。

$$\mathrm{Sim}(H_1,H_2)=\frac{\sum_{u\in R}\left|H_1[u]-H_2[u]\right|}{\frac{1}{2}\sum_{u\in R}\left|H_1[u]+H_2[u]\right|} \tag{4}$$

基于直方图相似性函数，本文对 100 幅图像进行了实验，图 3 给出了 100 幅图像的平均结果，这 100 幅图像的(1,2)低频交流分量的初始量化步长 Q^1 随机选取(取值范围 2–9)，Q^2 分别取 $3,\cdots,9$。H_1 是图像经过 $Q^1\to Q^2\to 0.98\text{resize}\to Q^2$ 后的统计结果，H^2 是图像直接经过 Q^2 的统计结果。图 3 中的值是不同的 Q^2 取值情况下的 $\mathrm{Sim}(H_1,H_2)$ 对 100 幅图像的均值。从图 3 我们可以看出均值维持在 0.08 的水平，这表明 0.98resize 可较好的近似原始的未压缩图像。

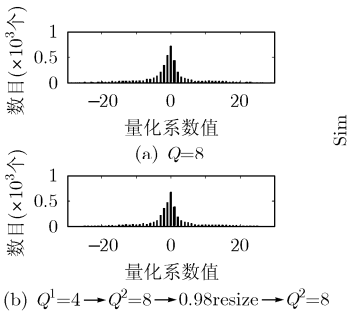

图 2

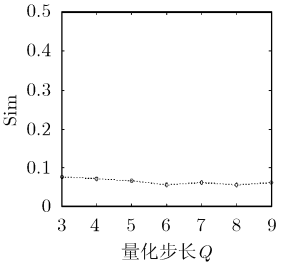

图 3 相似性统计

利用 0.98resize 的重构方法，生成原始图像只进行了一次 Q^2 量化的近似直方图。然后计算该直方图与双重量化的直方图之间的 Sim 值，如果 Sim 值大于一个给定的阈值(如 0.15)就可认为存在双重压缩。图 4(a)是原始图进行 $Q^1=4$, $Q^2=8$ 双重量化直方图，图 4(b)为重构方法得到的 $Q^2=8$ 的量化直方图。

进一步地，可根据直方图匹配来估计原始量化步长 Q^1，估计步骤如下：(1)计算待检测的图像的 DCT 系数直方图，记为 H_1；(2)根据 Q^2 列举满足 $Q^2=kq$ 的 q，k 为整数，分别用 q，Q^2 对 0.98resize 图像进行双重量化，得到的系数直方图记为：H_2^q；(3)则 $Q_1=\arg\min_q\left[\mathrm{sim}(H_2^q,H_1)\right]$。

作为示例，图 5(a)给出待检测图像的 DCT 系数直方图 $Q^1=4,Q^2=8$；图 5(b)为缩小图像进行 $Q^1=4,Q^2=8$ 量化的 DCT 系数直方图；图 5(c)为缩小图像进行 $Q^1=2$, $Q^2=8$ 量化的 DCT 系数直方图。可以看到缩小图像进行 $Q^1=4,Q^2=8$ 得到的结果和待检测图像本身的系数分布十分接近。

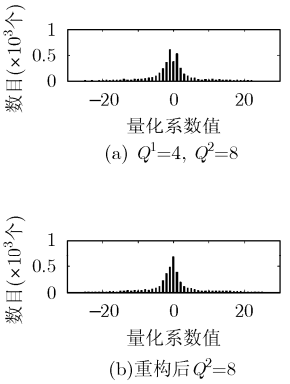

图 4

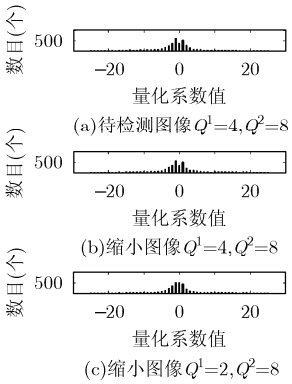

图 5

3.2.2 $Q^2\neq kQ^1$ 当 Q^2 不是 Q^1 的整数倍时，直方图存在着周期振荡性，Farid 提出了利用对直方图进行傅里叶变换，得出周期，从而估计 Q^1 [8]。He 等[9]阐述出直方图的周期振荡的原因，并推导出 $p=Q^1/\gcd(Q^1,Q^2)$，$\gcd(Q^1,Q^2)$ 表示 Q^1,Q^2 的最大公约数，所以不同的 Q^1,Q^2 可以得到相同的 p, 如 $Q^1=3,Q^2=8$ 与 $Q^1=6,Q^2=8$, 得到 p=3,则直接利用文献[8]中的傅里叶变换频谱局部极值寻找周期来估计 Q^1 就会有多值可能。图 6 给出了 $Q^1=3,Q^2=8$，$Q^1=6,Q^2=8$ 双重量化系数直方图的频谱图，从图 6 可见在出现多值情况时，原有方法无法准确判断具体的量化步长。这里我们利用上面介绍的 0.98resize 的方法给出了如下的估计方法。

估计步骤：(1)计算待检测图像的 DCT 系数直方图，计为 H_1，对其进行傅里叶变换；(2)求出相邻两峰值对应的频率差 Δf，则 $p=1/\Delta f$，列举小于 Q^2 且不是 Q^2 因子的系数集 $\Theta=\{q\mid q<Q^2,\mathrm{mod}(Q^2,q)\neq 0,q\in Z^+\}$，计算 $\Pi=\{q\mid p=q/\gcd(q,Q^2),q\in\Theta\}$；(3)若 $|\Pi|=1$，则 Π 中的元素即为要估计的 Q^1，否则利用 0.98resize 来处理待检测图像，并分别用 $q(q\in\Pi)$，Q^2 对 0.98resize 图像进行双重量化，得到的系数直方图记为：H_2^q；(4)利用 $Q^1=\arg\min_{q\in\Pi}[\mathrm{sim}(H_2^q,H_1)]$ 得到最终的估计值。

图 7 给出了 $Q^1=3,Q^2=8$ 和 $Q^1=6,Q^2=8$ 的实验结果。

用频谱来分析周期时，当两次量化步长差距很大或非常接近时，直方图的周期性不明显，很难用频谱方法准确估计。估计方法：列举小于 Q^2 的 Q^1，依次用 Q^1, Q^2 来量化前面讲述的缩小图像，再与待检测图像的直方图进行匹配，误差最小的就是要估计的 Q^1。图 8 给出了 3 幅分别进行了 $Q^1=3$, $Q^2=7$，$Q^1=4,Q^2=7$，$Q^1=5,Q^2=7$ 双重量化的 JPEG

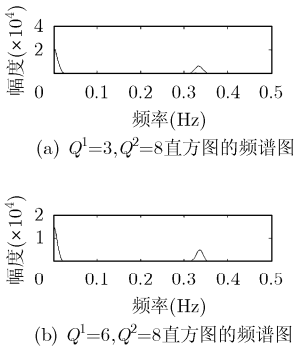

图 6

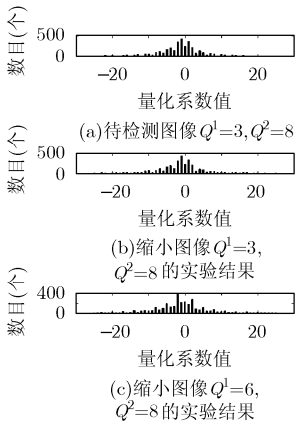

图 7

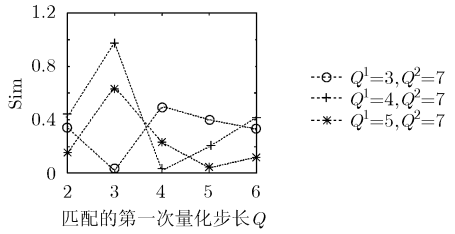

图 8 双重量化 $Q^1=3,Q^2=7$；$Q^1=4,Q^2=7$；$Q^1=5,Q^2=7$ 相似性统计

图像，分别用该方法来计算所得到的直方图匹配误差情况。

4　实验结果

取 100 幅灰度 bmp 图像，用 Q^1,Q^2 所对应的量化因子进行双重 JPEG 量化，再用文中所述的方案进行估计，每种情况正确估计出的结果如表 1 所示，从结果看，当 $Q^1>Q^2$ 时，准确率高，主要是因为当 $Q^1<Q^2$ 时，采取缩小与直方图匹配，误差造成估计准确率的下降。

5　结束语

本文利用 JPEG 图像双重压缩对系数分布产生的影响，提出了一种双重压缩后 JPEG 图像的原始量化步长的估计方法。该方法根据两次量化步长之间的大小关系分 3 种情况进行讨论。当原始量化步长大于第 2 次量化步长，提出了直接利用直方图计算的新方法；为解决原始量化步长是第 2 次量化步长因子和傅里叶频谱分析中的多值问题，提出了采用 0.98 缩放来近似未压缩图像的方法。本文方法能给出第 2 次量化步长为第 1 次量化步长倍数时的估计，并利用频谱分析的结果降低了计算的复杂度，实验结果表明本文方法有较高的估计准确度。值得注意的是当两次量化步长比较接近或差距较大时估计的结果都不是很好，这主要是由于在这两种情况下二次量化对直方图分布的影响的所能提供给估计器的信息不够充分，这在另一方面也表明单纯依赖一阶的直方图分布很难得到较好的估计结果，需要借用更能揭示双重量化本质的统计特征来实现更为精确的估计。

表 1 100 幅图像上的实验结果

Q^2	Q^1							
	2	3	4	5	6	7	8	9
2	92	98	99	97	98	96	99	97
3	95	94	96	96	99	98	97	98
4	96	95	94	97	97	98	99	98
5	95	98	97	95	97	98	97	98
6	93	94	96	93	92	98	97	97
7	92	92	94	93	93	95	96	98
8	93	91	95	90	92	87	92	91
9	86	93	92	91	96	89	85	90

参 考 文 献

[1] Sencar H T and Memon Nasir. Overview of state-of-art in digital image forensics. Part of Indian Statistical Institute Platinum Jubilee Monograph series titled 'Statistical Science and Interdisciplinary Research,' World Scientific Press, 2008.

[2] Luo Weiqi, Qu Zhenhua, Pan Feng, and Huang Jiwu. A survey of passive technology for digital image forensics. *Front. Comute. Sci. China*, 2007, 1(2): 166–179.

[3] Fan Z and de Queiroz R L. Identification of bitmap compression history: JPEG detection and quantizer estimation. *IEEE Trans. on Image Processing*, 2003, 12(2): 230–235.

[4] Fridrich J, Goljan M, and Du R. Steganalysis based on JPEG compatibility. Special session on Theoretical and Practical Issues in Digital Watermarking and Data Hiding, Multimedia Systems and Applications IV, Denver, Co, USA, August 19-24, 2001: 275–280.

[5] Fridrich J, Goljan M, and Hogea D. Attacking the outguess. Proc. Multimedia and Security, Workshop at ACM Multimedia, Juan-les-Pins, France, December 6, 2002: 3–6.

[6] Fridrich J, Goljan M, and Hogea D. Steganalysis of JPEG images: Breaking the F5 algorithm. Proc. 5th International Workshop on Information Hiding, Noordwijkerhout, The Nederlands, October 7–9, 2002: 310–323.

[7] Lukas J and Fridrich J. Estimation of primary quantization matrix in double compressed JPEG images. Proc. Of DFRWS, Cleveland, OH, USA, August 6-8, 2003: 67–84.

[8] Popescu A and Farid H. Statistical tools for digital forensics. 6th International Workshop on Information Hiding, Toronto, Canada, May, 2004: 128–147.

[9] He J, Lin Z, Wang L, and Tang X. Detecting doctored JPEG images via DCT coefficient analysis. Proc. of ECCV, Berlin, 2006: 423–435.

王俊文：　男，1984年生，博士生，研究方向为信息安全、数字取证.

刘光杰：　男，1980年生，博士，讲师，研究方向为隐写分析、数字图像认证.

戴跃伟：　男，1962年生，教授，博士生导师，研究方向为多媒体信息安全、数字水印.