

JPEG 二次压缩的分析与检测

戴蒙¹⁾ 赵学州²⁾

¹⁾ (上海应用技术学院计算机与信息工程系, 上海 200235) ²⁾ (上海无线电设备研究所, 上海 200090)

摘要 为了对 JPEG 二次压缩后的图像进行有效检测, 通过研究 JPEG 图像经二次压缩后所引起的离散余弦变换 (DCT) 统计特性的变化及二次压缩前后的量化表 q_1, q_2 对二次压缩后图像的 DCT 统计所引起的变化的分析, 提出了针对不同 q_1, q_2 JPEG 二次压缩的检测方法。实验结果表明, 该检测方法是有效的。

关键词 JPEG 双压缩 周期模式 直方图

中图法分类号: TN918.91 文献标识码: A 文章编号: 1006-8961(2009)12-2599-04

The Analysis and Detection of Double JPEG Compression

DAI Meng¹⁾, ZHAO Xue-zhou²⁾

¹⁾ (Computer Science & Information Engineering Department, Shanghai Institute of Technology, Shanghai 200235)

²⁾ (The Institute of Shanghai Radio Equipment, Shanghai 200090)

Abstract To assess images applied with JPEG double compressed, the DCT statistic of the JPEG image will be changed, which is the focus in this paper. The distributions of DCT coefficients are varied with the double compression quantization tables: q_1, q_2 . The difference of the distributions is analyzed in this paper. The different detection methods are proposed based on different q_1, q_2 , which are effective proved by the experiments in this paper.

Keywords JPEG double compressed, periodic mode, Histogram

1 引言

JPEG 格式图片在 F5^[1] 和 OutGuess^[2] 隐写算法流程中, 均有将原始图片转换成 JPEG 格式图片的过程。如果原图是以 JPEG 格式保存的, 那么 F5 和 OutGuess 隐写算法先将 JPEG 图片解压缩, 再以用户定义的质量因子进行再压缩。这个过程就称为 JPEG 二次压缩 (简称双压)。因此, 研究检测图片是否被双压缩 (简称双压) 可以作为图片是否被隐写的手段之一。

文献[3]提到了 JPEG 格式文件经过双压后, 图像的统计特性发生了改变, 并以此作为图片是

否被篡改的衡量指标。通过实验发现, 文献[3]中所提到的周期模式是在一定条件下才出现的。所谓周期模式是指图像的 AC (alternating current) 系数直方图具有周期性的特征, 其与自然图像 (经过一次压缩) 的 AC 系数直方图满足 Laplacian 分布有着明显的区别。当第 2 次压缩采用的质量因子小于第 1 次压缩的质量因子时, 双压后的 DCT (离散余弦变换) 系数直方图仍然近似满足 Laplacian 分布 (详见第 2 节), 没有出现上述周期模式。本文在已有工作的基础上 (文献[4]) 进一步研究, 提出了一种新的检测 JPEG 二次压缩的检测方法。

基金项目: 上海高校选拔培养优秀青年教师科研专项基金 (yyy-07007)

收稿日期: 2009-06-20; 改回日期: 2009-09-24

第一作者简介: 戴蒙 (1979 ~), 女, 讲师。华东理工大学检测技术与自动化专业博士研究生。主要研究信息安全、图像隐写分析、数字图像处理等。E-mail: damon_1111@163.com

2 隐写算法的 JPEG 双压缩流程

JPEG 压缩流程是首先将输入的图像分成若干个 8×8 的子块,然后对每一个子块进行 DCT 变换,将变换后的系数用量化表量化,量化在 JPEG 压缩中起着关键作用,大量的冗余信息在量化后被丢掉。最后是无损编码阶段,这一阶段使用熵编码,不会造成信息的丢失。

JPEG 双压缩流程是指使用不同量化表 q_1, q_2 对图片进行压缩,图 1 为 JPEG 双压流程图,其中 BMP 表示位图, J_1, J_2 表示 JPEG 图像。

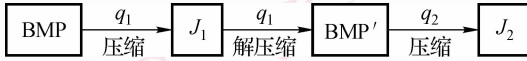


图 1 JPEG 双压缩过程

Fig. 1 Procedure of double JPEG compression

3 JPEG 双压缩量化表分析

K_1 为经过一次压缩后的 AC 系数, K_2 为 K_1 经过再压缩后的 AC 系数,它们的量化表分别为 q_1, q_2 (如无特殊说明,文中 q_1, q_2 与标准量化表存在比例关系),则 K_1, K_2 满足

$$K_2 = \text{integer_round} \left[\frac{K_1 \cdot q_1}{q_2} \right] \quad (1)$$

integer_round 为四舍五入运算,令 $r = \frac{q_1}{q_2}$ 。在 K_1 给定的前提下, K_2 的值由 r 决定。因为 AC 系数的能量主要集中在低频部分,所以只考察区间 $[0, 10]$ 的 AC 系数直方图。

式(1)等价于

$$K_2 - 0.5 \leq rK_1 < K_2 + 0.5 \quad (2)$$

若 $K_2 = K_1$ 等价于 $\text{integer_round} [rK_1] = K_1$, 则下式必定成立:

$$K_1 - 0.5 \leq rK_1 < K_1 + 0.5 \quad (3)$$

在考察区间内上式中 r 的取值范围为 $[0.95, 1.05)$ 。因此引起 K_2 不同性质改变的 r 的取值范围分别为 $r_1 \in (0, 0.95), r_2 \in [0.95, 1.05), r_3 \in [1.05, 100)$ 。表 1 为不同区间内 AC 值。

表 1 不同区间内 AC 值

Tab. 1 The value of AC in different areas

r	AC										
1	0	1	2	3	4	5	6	7	8	9	10
0.1	0	0	0	0	0	1	1	1	1	1	1
0.5	0	1	1	2	2	3	3	4	4	5	5
0.75	0	1	2	2	3	4	5	5	6	7	8
0.9	0	1	2	3	4	5	5	6	7	8	9
1.1	0	1	2	3	4	6	7	8	9	10	11
1.5	0	2	3	5	6	8	9	11	12	14	15

下面分别讨论 r 在这 3 个区间内图片的检测方法。

3.1 $r = r_1$

当 r 的取值范围为 $(0, 0.95)$ 时, K_2 的直方图与 K_1 直方图相比变高变宽,但与采用 q_2 量化表一次压缩图片的直方图无明显变化(见图 2)。图 2 中 $r = 0.4$ 。隐藏图片的载体图片为使用量化表 q_1 压缩的 JPEG 图片,使用 OutGuess 隐写软件并采用用户自定义量化表 q_2 经过第 2 次压缩,得到的隐藏图片的 AC 系数的最大值 $(K_2)_{\max}$ 和最小值 $(K_2)_{\min}$ 分别为 56 和 -51。使用量化表 q_2 经过一次压缩后的 AC 系数的最大值 $(K_1)_{\max}$ 和最小值 $(K_1)_{\min}$ 分别为 56 和 -51。从图 2 可以看出,双压图片(“ Δ ”表示)和一次压缩图片(“+”表示)在直方图上几乎完全一致,所以不能直接从二次压

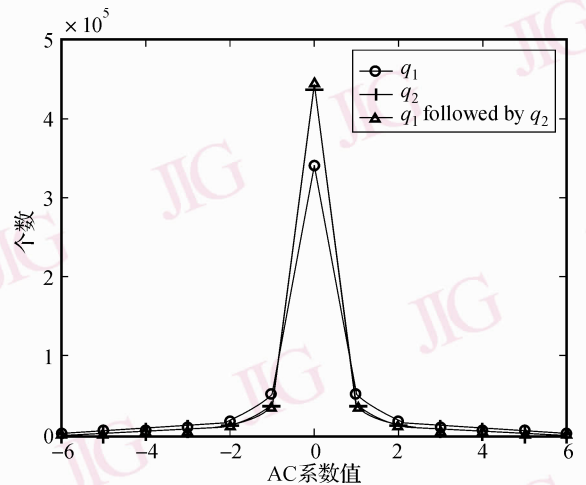


图 2 一压与双压效果对比图 ($r = 0.4$)

Fig. 2 The comparison between one and double compressed image ($r = 0.4$)

缩后的隐藏图象的直方图判断待检图片是否被
 隐写。

3.1.1 特征选取

图 3 为特征提取的示意图。

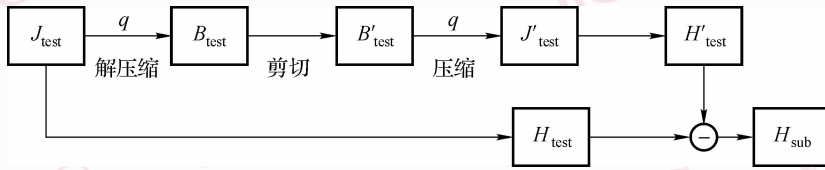


图 3 特征提取示意图

Fig. 3 Features extraction diagram

其中, q 可以直接从 JPEG 文件头中读取, 对于一压的图片 $q = q_1$, 双压图片 $q = q_2$; 剪切流程为剪切掉各方向 4 个像素; 再用 q 压缩产生的校准图片 (calibrated^[6]) J'_{test} 宏观上的特征与只经过量化表 q 一次量化的图片相似, 所以它的 DCT 系数与原图具有相似的统计特性^[6]。 H_{test} 表示待检图片的直方图, H'_{test} 表示待检图片校正后的直方图, 考察下式

$$H_{sub} = \text{abs}(H_{test} - H'_{test}) \quad (4)$$

由表 1 中的数据可见, 当 $r \in (0, 0.75)$ 时, H^0 (上角 0 表示 $AC=0$, 下同) 和 H^1 与一次压缩都有了改变; 当 $r \in [0.75, 0.95)$ 时 H^0, H^1 与一次压缩相比没有变化。所以 $r \in (0, 0.75)$ 时, 考查

$$f_0 = H_{sub}^0 / H_{sub}^1 \quad (5)$$

$r \in [0.75, 0.95)$ 时, 考查

$$f_i = H_{sub}^i / H_{sub}^{i+1} \quad i = 1, 2, \dots \quad (6)$$

3.1.2 分类检测

训练及测试使用的原图为使用扫描仪得到的位图, 再使用用户自定义的压缩因子进行一次压缩得到一次压缩后的原始图像, 训练及测试所使用的隐藏图像为采用符合 $r = r_2$ 情况的压缩因子对一次压缩后的原始图像进行二次压缩。400 幅训练图片中, 200 幅为一次压缩原图, 200 幅为符合这两种情况的双压图片。样本的所有特征统计值放入训练器中进行训练得到阈值。测试图片中一次压缩和双压图片各 162 幅。分类测试结果如表 2 所示, 其中 detection 表示检测正确率。

表 2 分类结果

Tab. 2 Classification of images

	r	检测正确率 (%)
一次压缩		89.51
二次压缩	(0, 0.75)	82.10
一次压缩		88.27
二次压缩	(0.75, 0.95)	98.77

3.2 $r = r_2$

当 r 的取值范围为 $[0.95, 1.05)$ 时, K_2 的取值与 K_1 相等, 直方图无任何变化。如果一次压缩和二次压缩采用的量化表满足此条件, 则无法从判断 JPEG 图像是否被二次压缩的方法来判断待测图片是否经过隐写。对于这类隐写图像的检测, 文献 [5] 和文献 [6] 分别给出了两种检测方法: 高阶统计量和基于特征的隐藏信息检测。

3.3 $r = r_3$

令 $k_1^{d_1}, k_1^{d_2} \in K_1$, 且满足 $k_1^{d_2} = k_1^{d_1} + 1$; 经过双压后, $k_1^{d_1}, k_1^{d_2}$ 分别变换成 $k_2^{d_1}, k_2^{d_2}$, 即

$$k_2^{d_1} = \text{integer_round}[rk_1^{d_1}]$$

$$k_2^{d_2} = \text{integer_round}[rk_1^{d_2}]$$

$$\begin{aligned} k_2^{d_2} - k_2^{d_1} &= \text{integer_round}[rk_1^{d_2}] - \text{integer_round}[rk_1^{d_1}] \\ &= \text{integer_round}[rk_1^{d_1} + r] - \text{integer_round}[rk_1^{d_1}] \\ &= rk_1^{d_1} + r + p_1 - rk_1^{d_1} - p_2 \\ &= r + (p_1 - p_2) \end{aligned}$$

式中, $p_1, p_2 \in [-0.5, 0.5)$, $r \geq 1.05$, 所以存在 $k_1^{d_1}, k_1^{d_2}$ 使得 $p_1 - p_2 \geq 0.95$, 即

$$k_2^{d_2} - k_2^{d_1} \geq 2 \quad (7)$$

式(7)表明双压后不存在处于区间 $(k_2^{d_1}, k_2^{d_2})$ 的 AC 系数, 即区间 $(k_2^{d_1}, k_2^{d_2})$ 的 AC 系数的个数为 0。所以, 对于 $r = r_3$ 这种情况下的双压图片, 其 AC 系数

直方图必有 0 点。

实验用图 600 幅,均为满足 $r \in [1.05, 100)$ 的双压图片。检测正确率达到 100%。图 4 为其中一幅的 AC 直方图比较。

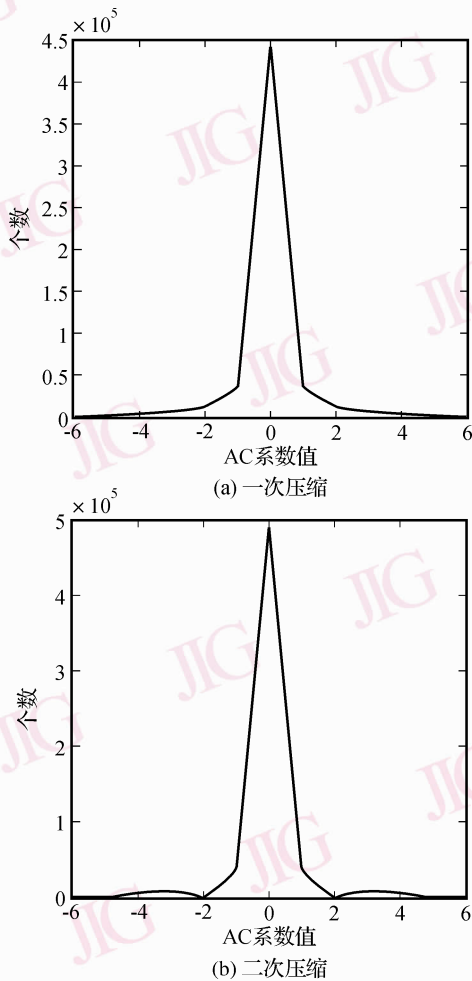


图 4 一次压缩和二次压缩图像直方图比较 ($r = 1.4$)

Fig. 4 Comparison of the histogram between one compressed and double compressed ($r = 1.4$)

4 结论

在实际运用的图片中,出现较多情况的是 $r \in (0, 0.75)$ 和 $r \in [1.05, 100)$ 。本文所提出的检测方法在 $r \in [0.75, 0.95)$ 区间内越趋近于 0.95 时检测所需的特征越多。本文提出的基于 JPEG 二次压缩的检测方法可以作为隐藏信息图片的检测手段之一。

实验结果表明,本文提出的 JPEG 图片双压缩的检测方法是有效的,该检测方法还可以用于检测图片是否被篡改。下一步将进行能表征净图的其他特征的研究工作。

参考文献 (References)

- 1 Andreas Westfeld. F5-a steganographic algorithm: High capacity despite better steganalysis [A]. In: Proceeding of 4th International Workshop on Information Hiding [C], Pittsburgh, PA, USA, 2001: 289-302.
- 2 Niels Provos. Defending against statistical steganalysis [A]. In: Proceedings of the 10th USENIX Security Symposium [C], Washington, DC, USA, 2001: 323-335.
- 3 Popescu Alin C, Farid Hany. Statistical tools for digital forensics [A]. In: Proceeding 6th International Workshop on Information Hiding [C], Toronto, Canada, 2004: 128-147.
- 4 Dai Meng, Lin Jia-jun, Mao Jia-fa. The analysis and detection of double JPEG compression [J]. Journal of Image and Graphics, 2006, 11(11): 1692-1695. [戴蒙,林家骏,毛家发. JPEG 二次压缩的分析与检测 [J]. 中国图象图形学报, 2006, 11(11): 1692-1695.]
- 5 Farid Hany, Siwei L. Detecting hidden messages using higher-order statistics and support vector machines [A]. In: Proceeding 5th International Workshop on Information hiding [C], Noordwijkerhout, the Netherlands, 2002: 340-354.
- 6 Jessica. Fridrich. Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes [A]. In: Proceeding 6th Information Hiding Workshop [C], Toronto, Canada, 2004: 67-81.