FIRST JPEG QUANTIZATION MATRIX ESTIMATION BASED ON HISTOGRAM ANALYSIS

Giovanni Puglisi[⋆] *Arcangelo Ranieri Bruna*[⋆] *Fausto Galvan*[†] *Sebastiano Battiato*[⋆]

[⋆]University of Catania, Italy
[†]University of Udine, Italy
[⋆†]Image Processing Lab - http://iplab.dmi.unict.it
{puglisi, bruna, battiato}@dmi.unict.it fausto.galvan@uniud.it

ABSTRACT

To assess if a digital image has been (or not) doubly compressed is a challenging issue especially in forensics domain where could be fundamental clarify if, in addition to the compression at the time of shooting, the picture was decompressed (in some way) and then resaved. This is not a clear indication of forgery, but it guarantees that the image, probably, is not the original one. In this paper we propose a novel technique able to recover the coefficients of the first compression in a double compressed JPEG image under some assumptions. The proposed approach exploits how successive quantizations followed by dequantizations introduce some regularities (e.g., sequence of zero and not zero values) on the histograms of coefficient distributions that could be analyzed to recover the original compression parameters. Experimental results and comparisons with state of the art methods confirm the effectiveness of the proposed approach.

Index Terms— Double JPEG Compression, Forgery Identification

1. INTRODUCTION

The pipeline which leads to ascertain whether an image has undergone to some kind of forgery leads through the following steps: determine whether the image is ”original” and, in the case where the previous step has given negative results, try to understand the past history of the image. Discovery that the input image has been manipulated or not is a prelude to any other type of investigation. Regarding the first stage, the EXIF metadata could be examined, but they are not so robust to tampering, so that they can provide indicative but not certain results. To discover image manipulations, many approaches have been proposed in literature (JPEG blocking artefacts analysis [1], hash functions [2], etc.). A lot of works have proved that analyzing the statistical distribution of the values assumed by the DCT coefficients is widely more reliable [3, 4]. In this regard, as reported in [5], [6] and [7], by checking the involved histogram it is possible not only to determine whether the image was or not doubly saved, but also if the quantization coefficient, in this case, was greater (or less) than the one used in the first compression. In [8], the authors suggest a method to determine whether an image has been subjected to double compression with the same compression factor. The part of this theory not yet fully developed, at least in our knowledge, concerns the determination of the values of the coefficients of the first quantization. In [7] the authors expose some ideas to retrieve the coefficients of the first quantization just analysing the normalized histograms. They then focused on a method that uses a Neural Network as a classifier. Their approach, however, is not robust with respect to medium and high frequencies, and it has been proven only for a specific subset of the AC terms. The works in [9, 10] also estimate first quantization coefficients but only to locate forgeries without providing exhaustive results related to its estimation.

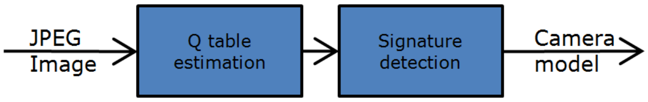

Fig. 1. Block based schema of the pipeline used to retrieve the camera model from an image.

In this paper we focused on the determination of first quantization coefficients when the second quantization factor is lower than the first one. The proposed approach analyses histograms of quantized DCT coefficients of double compressed images exploiting their peculiarities. In particular when the second compression is lighter than the first one it is possible to retrieve a pool of possible candidates for the first quantization factor making use of a proper simulation of the double compressed coefficient distribution with different first quantization values. The proposed method can be used as stand-alone module (just to detect forgeries) or combined with other methods. As example, in the Fig. 1 the combination with the ”Signature detection” algorithm proposed in [11] allows retrieving the camera model used to shoot the image.

This paper is structured as follows: in Section 2 JPEG

978-1-4799-2341-0/13/$31.00 ©2013 IEEE

compression algorithm together with some properties of double compressed images are reviewed. Section 3 presents the proposed approach whereas in Section 4 the effectiveness of the proposed solution has been tested considering real data (double JPEG compressed images). Finally, in Section 5 we report our conclusions and our prospects for future works.

2. SCIENTIFIC BACKGROUND

The first step in the JPEG compression engine [12] consists in a partition of the input image into 8×8 non-overlapping blocks (for both luminance and chrominance channels). A DCT transform is then applied to each block; next a proper dead-zone quantization is employed just using for each coefficient a corresponding integer value belonging to a 8×8 quantization matrix [13]. The quantized coefficients obtained just rounding the results of the ratio between the original DCT coefficients and the corresponding quantization values are then transformed into a data stream by mean of a classic entropy coding (i.e., run length/variable length). Coding parameters and other metadata are usually inserted into the header of the JPEG file to allow a proper decoding. If an image has been compressed twice (i.e., after having introduced some malicious manipulation) although, of course, the current quantization values are available, the original (initial) quantization factors are lost. It is worth noting that in forensics application, this information could be fundamental to assess the integrity of the input image or to reconstruct some information about the embedded manipulation [14].

For sake of simplicity, considering a single DCT coefficient c and the related quantization factors q_1 (first quantization) and q_2 (second quantization), the value of each coefficient is given by:

$$c_{DQ} = round((round(\frac{c}{q_1}) * q_1 + e) * \frac{1}{q_2}) \qquad (1)$$

where e is the error introduced by several operations, such as color conversions (YCrCb to RGB and vice versa), rounding and truncation of the values to eight bit integers, etc. Note that the factor e, often omitted in previous published works, if not properly managed can limit the effectiveness of the proposed methodology.

As already described in [5] double quantization introduces artefacts in the DCT coefficient histograms. These artefacts can be then exploited to recover first quantization factor q_1 (q_2 is already present in the header). Considering as example an image I and the corresponding histogram of a generic AC coefficient c_{fj}, the histograms h_{DQi} obtained applying a double compression with several q_{1i} and a given q_2 are different and present predictable sequences (pattern) of zero (not zero) values as depicted in Fig. 2.

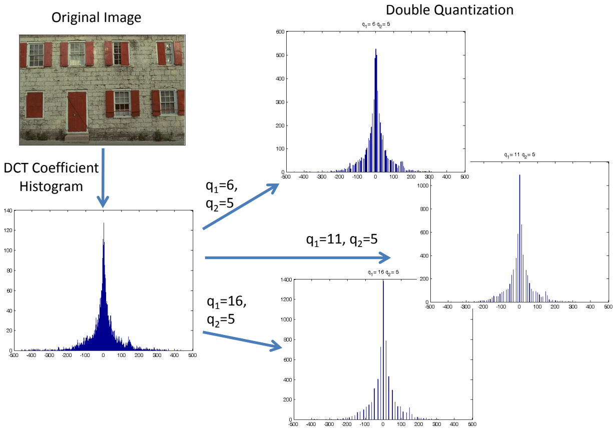

Fig. 2. Examples of DCT coefficient histograms relative to double quantized images. Specifically, an image I has been double compressed with several q_{1i} and a fixed q_2. The final histograms show several differences and their analysis can be useful to recover compression history.

3. PROPOSED APPROACH

As noted above, double JPEG compression modifies the histograms of the DCT coefficients with a function depending on both first and second quantization factor (q_1 and q_2). Moreover, as can be easily seen in Fig. 2 the sequence of zero (and not zero) values of the histogram related to a double compressed image I_{DQ} provides useful information for the estimation of the first quantization factor q_1. Starting from the same initial distribution of DCT coefficients c_i, different histograms have been generated just considering several values of q_{1i} followed by a further quantization with q_2.

This paper exploits a binary representation based on the pattern of zero (and not zero) values of the histograms to perform a first selection of a pool of q_1 candidates. Specifically, considered a wide range of initial possible values ($q_1 \in [q_{2+1}, ..., q_{max}]$), such selection method provides just a few candidates. The overall schema of the algorithm is reported in Fig. 3. Starting from a double quantized image I_{DQ}, DCT coefficients c_{DQ} are extracted and, for each frequency f_j with $j \in [1, 2, ..., 64]$, the selection algorithm is applied. First, the histogram of the absolute value of DCT coefficients c_{f_j} is computed and refined, filtering out some unreliable values by simple trying to preserve the monotonicity of the distribution (this is applied only to AC coefficients usually characterized by Laplace distribution [15]). The histogram of the DC coefficient is filtered according to a threshold depending on the mean value of all not null bins. Later, a binary vector is computed just considering the sequence of zero and not zero values of the filtered histogram.

A set of binary representations is then built for each $q_{1i_{f_j}}$

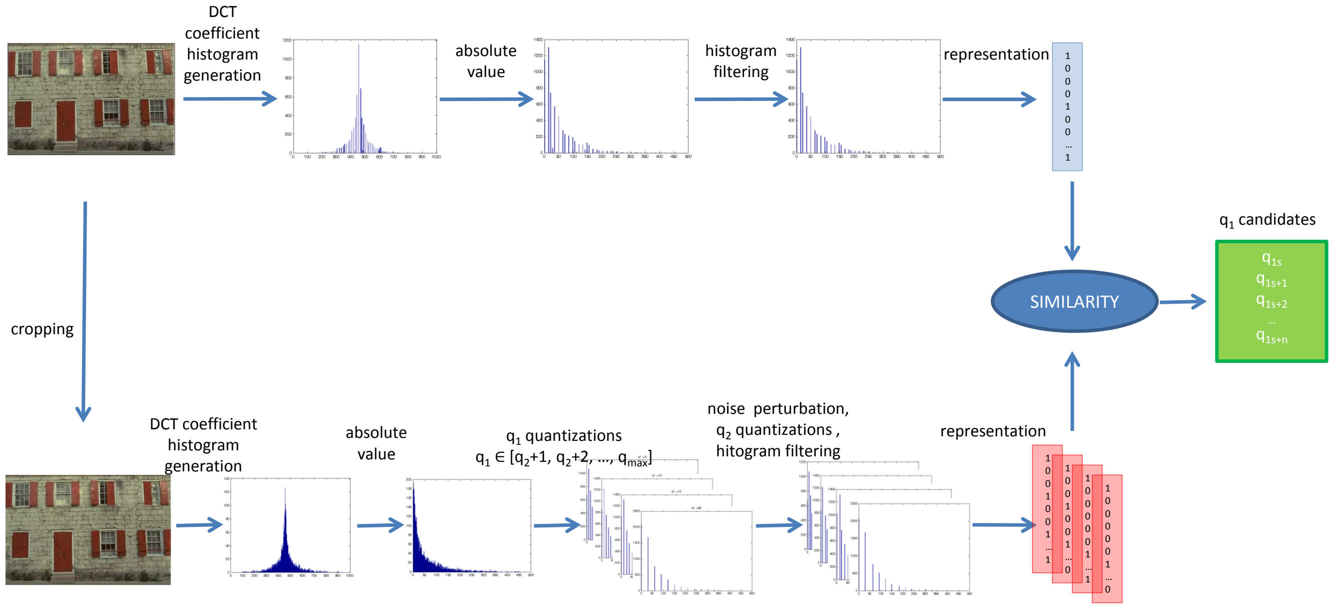

Fig. 3. Candidate selection strategy. Starting form a wide set of q_1 candidates the proposed algorithm, by proper analysing histogram properties, selects a short list of elements.

value exploiting information coming from the input (double compressed) image I_{DQ}. Specifically, as already proposed in [7], by performing a proper cropping of the double compressed image, an estimation of the original DCT coefficients can be obtained ($\widehat{c_{f_j}}$). These coefficients are then used as input of a double compression procedure where the first quantization is performed by using $q_{1i_{f_j}}$ and the second one by simply using the already known values of the second quantization coefficients ($q_{2_{f_j}}$ values are present in the header data). To better mimic double JPEG compression, some additional Gaussian noise [10] is added before the second quantization step. Double quantized histograms are then refined to remove unreliable bins (trying to preserve the monotonicity of the distribution) and the binary representation is generated. Based on the similarity between the generated representations and the one of the I_{DQ}, a set of $q_{1i_{f_j}}$ candidates are then selected (C_{f_j}). In particular for each $q_{1i_{f_j}}$ the sequence of zero (not zero) values represented as a binary sequence is compared by using the following similarity function:

$$F_s = \sum_{i=1}^{N} (B_{real}(i) * B_{sim}(i)) - \sum_{i=1}^{N} (B_{real}(i) \oplus B_{sim}(i)) \quad (2)$$

where B_{real} and B_{sim} are the N-bin binary representations related to the original input image I_{DQ} and the one corresponding to the simulated double compression respectively. Moreover $*$ and $\oplus$ represent logical AND and XOR operators.

The second step of the proposed approach tries to estimate the correct q_1 from the candidate set. To perform this task the whole histogram information is used instead of considering only the binary representation. Specifically, the refined histograms $H_{q_{1s}}$ related to the simulated double quantization with $q_{1s} \in C_{f_j}$ are compared with H_{real} obtained from I_{DQ} and the closest one is selected as follows:

$$\widehat{q_{1f_j}} = \min_{q_{1s} \in C_{f_j}} \sum_{i=1}^{N} \min(max_{diff}, |H_{real}(i) - H_{q_{1s}}(i)|) \quad (3)$$

where N is the number of bins of the histograms and max_{diff} is a threshold used to limit the contribution of a single difference in the overall distance computation.

4. EXPERIMENTAL RESULTS

In order to prove the effectiveness of the proposed approach several tests and comparisons have been performed considering real double compressed images. Starting from a set of 24 uncompressed images [16], by using the JPEG encoder included in Matlab, a dataset of double compressed images has been built just considering quality factors (QF_1, QF_2) in the range 50 to 100 at step of 10. Taking into account the condition $q_1 > q_2$, the final dataset contains 360 images. All the tests have been performed with the same parameter setting (experimentally found). Specifically, the selection procedure (see Fig. 3) considers the best 8 candidates with respect to the similarity function (2). Moreover, q_{max} and max_{diff} have been set to 30 and 50 respectively.

Table 1. Percentage of erroneously estimated q_1 values at varying of quality factor (QF_1, QF_2) relative to the first 3 DCT coefficients in zig-zag order.

		QF_2				
		60	70	80	90	100
QF_1	50	0.00%	0.00%	0.00%	0.00%	0.00%
	60	-	0.00%	0.00%	0.00%	0.00%
	70	-	-	0.00%	0.00%	0.00%
	80	-	-	-	0.00%	0.00%
	90	-	-	-	-	0.00%

Table 2. Percentage of erroneously estimated q_1 values at varying of quality factor (QF_1, QF_2) relative to the first 6 DCT coefficients in zig-zag order.

		QF_2				
		60	70	80	90	100
QF_1	50	1.39%	0.69%	0.00%	0.00%	0.00%
	60	-	0.00%	0.69%	0.00%	0.00%
	70	-	-	0.00%	0.00%	0.00%
	80	-	-	-	0.00%	0.00%
	90	-	-	-	-	0.00%

Table 3. Percentage of erroneously estimated q_1 values at varying of quality factor (QF_1, QF_2) relative to the first 10 DCT coefficients in zig-zag order.

		QF_2				
		60	70	80	90	100
QF_1	50	2.08%	4.58%	0.00%	0.00%	0.00%
	60	-	1.25%	2.92%	0.00%	0.00%
	70	-	-	0.00%	0.00%	0.00%
	80	-	-	-	0.00%	0.00%
	90	-	-	-	-	0.00%

Table 4. Percentage of erroneously estimated q_1 values at varying of quality factor (QF_1, QF_2) relative to the first 15 DCT coefficients in zig-zag order.

		QF_2				
		60	70	80	90	100
QF_1	50	4.44%	7.78%	1.94%	1.11%	0.00%
	60	-	5.56%	8.06%	0.28%	0.00%
	70	-	-	0.00%	0.28%	0.00%
	80	-	-	-	0.00%	0.00%
	90	-	-	-	-	0.00%

Tables 1, 2, 3 and 4 report the average percentage of erroneously estimated q_1 values at varying of quality factor relative to the first 3, 6, 10 and 15 DCT coefficients. These values have been averaged over all images [16]. The coefficients were considered in zig-zag order. This order, used in JPEG standard [17], allows sorting the coefficients from the lowest frequency (DC) to the highest frequencies in a 1D vector. As expected, better results are usually obtained for higher QF_1 and QF_2 quality factor corresponding to lower quantization.

Further analyses have been conducted in order to study the performance of the proposed approach with respect to each specific DCT coefficient. In Fig. 4 is reported the average percentage of erroneously estimated q_1 values at varying of the DCT coefficients (from low to high frequencies). These values are obtained averaging over all (QF_1,QF_2). As expected the performance of the proposed solution degrades with DCT coefficients corresponding to highest frequencies.

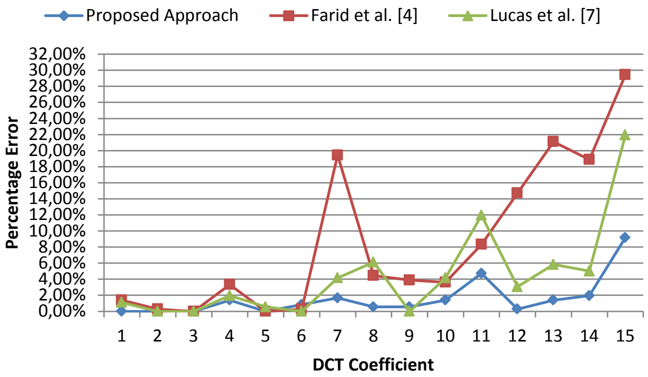

Fig. 4. Percentage of erroneously estimated q_1 values at varying of DCT coefficient position considering the proposed approach and state of the art methods ([4], [7]). These values are obtained averaging over all (QF_1,QF_2) and images [16].

To further assess the effectiveness of the proposed approach, several comparisons with state of the art techniques have been performed. Methods based on histogram comparisons [7] and multiple DCT coefficient compression properties [4] have been selected. Specifically, the properties of the error function proposed in [4] have been exploited to build up a method for q_1 estimation. As can be easily seen from Fig. 4, the proposed approach provides satisfactory results outperforming the considered state-of-the-art approaches.

5. CONCLUSION

In this paper we proposed a novel algorithm for the estimation of the first quantization coefficient from double compressed JPEG images. To confirm the effectiveness of the proposed solution several tests have been conducted on real double compressed images. Comparisons have been also performed with respect to the state of the art approaches ([4], [7]). Future work will be devoted to find smart solutions able to retrieve q_1 also in the case where $q_1 \leq q_2$. Moreover additional experiments related to the recovering of the overall initial quantization matrix, considering a double compression process achieved by applying actual quantization tables used by camera devices and common photo-retouching software (e.g., Photoshop, Gimp, etc.) will be performed.

6. ACKNOWLEDGMENTS

We thank AMPED SRL [18] for providing us useful code for DCT coefficient manipulation.

7. REFERENCES

[1] A. R. Bruna, G. Messina, and S. Battiato, "Crop detection through blocking artefacts analysis," in *Proceedings of International Conference on Image Analysis and Processing (ICIAP 2011)*, 2011, vol. 6978 of *Lecture Notes in Computer Science*, pp. 650–659.

[2] S. Battiato, G. M. Farinella, E. Messina, and G. Puglisi, "Robust image alignment for tampering detection," *IEEE Transactions on Information Forensics & Security*, vol. 7, no. 4, 2012.

[3] S. Battiato and G. Messina, "Digital forgery estimation into DCT domain: a critical analysis," in *Proceedings of the First ACM workshop on Multimedia in forensics (MiFor 2009)*, 2009, pp. 37–42.

[4] H. Farid, "Exposing digital forgeries from JPEG ghosts," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 4, pp. 154–160, 2009.

[5] A.C. Popescu and H. Farid, "Statistical tools for digital forensics," in *Proceeding of the 6th International Workshop on Information Hiding*, Toronto, Canada, 2004.

[6] J. Redi, W. Taktak, and J. Dugelay, "Digital image forensics: a booklet for beginners," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 133–162, 2011.

[7] J. Lukas and J. Fridrich, "Estimation of primary quantization matrix in double compressed JPEG images," in *Proceedings of Digital Forensic Research Workshop (DFRWS)*, 2003.

[8] F. Huang, J. Huang, and Y. Q. Shi, "Detecting double JPEG compression with the same quantization matrix," *Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 848–856, Dec. 2010.

[9] W. Wang, J. Dong, and T. Tan, "Exploring DCT coefficient quantization effect for image tampering localization," in *Workshop on Information Forensics and Security*, 2011.

[10] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 1003–1017, 2012.

[11] H. Farid, "Digital image ballistics from JPEG quantization: A followup study," Tech. Rep. TR2008-638, Department of Computer Science, Dartmouth College, 2008.

[12] G. K. Wallace, "The JPEG still picture compression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[13] S. Battiato, M. Mancuso, A. Bosco, and M. Guarnera, "Psychovisual and statistical optimization of quantization tables for DCT compression engines," in *Proceedings of the 11th International Conference on Image Analysis and Processing (ICIAP 2001)*, 2001, pp. 602–606.

[14] E. Kee, M. K. Johnson, and H. Farid, "Digital image authentication from JPEG headers," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1066–1075, 2011.

[15] E. Lam and J. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1661–1666, 2000.

[16] "Dataset Eastman Kodak Company: PhotoCD PCD0992, http://r0k.us/graphics/kodak/," 2013.

[17] "ITU T.81 (09/92) - "Information Technology - Digital compression and coding of countinuous-tone still images - requirements and guidelines"," .

[18] "Amped srl, http://ampedsoftware.com/company," 2013.