Proceedings of the IASETD International Conference
Communiaction, Network, and Information Security
November 14-16, 2005, Phoenix, AZ, USA

THE ACHILLES' HEEL OF JPEG-BASED IMAGE AUTHENTICATION

Mathias Schlauweg, Torsten Palfner, Dima Pröfrock and Erika Müller
Institute of Communications Engineering
Engineering Faculty, University of Rostock
Rostock 18119, Germany
{mathias.schlauweg, torsten.palfner, dima.proefrock, erika.mueller}@uni-rostock.de

ABSTRACT
Many image authentication systems in the DCT domain rely on the two invariant properties of the JPEG compression algorithm, which were found out by Lin and Chang. Based on these two assumptions, the authors of these authentication systems utilize that lossy JPEG compression to a pre-defined quality factor always yields the same relationships of coefficients, which can be used to generate image content dependent signature information. In this paper, we prove that this commonly used signature generation for an authentication purpose is not secure. If someone is intended to replace this signature generation by a cryptographically secure hash-based one, then the JPEG properties really have to be invariant. We show that a considerable amount of bit fluctuations can occur caused by rounding and clipping errors due to JPEG compression, which have to be taken into account. The statement of the invariants of the JPEG compression does not hold always. We determine the distribution of coefficient fluctuations and suggest using an extended secure hash-based signature generation in conjunction with error correction coding to overcome these fluctuations.

KEY WORDS
Secure hash-based image authentication, invariant JPEG properties, coefficient alterations, error correction coding

1. Introduction

The rapid evolution of multimedia technology over the past decade has brought many advantages in the creation and distribution of image content. But beneath the ability of easy copying, transmitting and editing digital images the need for image content protection increases. Digital images can be modified or forged with a wide variety of available manipulation software and hence it is rather difficult to tell if a picture is the original one, which has been taken by a camera, or if it has been tampered with. Thus, image authentication techniques based on digital watermarking and cryptography aim to prevent illegitimate tampering and fraudulent use of modified images.

The problem of data authentication is known from the classical cryptography. To verify the exact data integrity, a signature is generated from the source signal by the use of secure hash functions (e.g., SHA-1, MD5). Afterwards, the signature message digest is encrypted with a secret key. The recipient decrypts the signature and matches it with the hash generated from the received signal [1]. If even one bit of the signal is modified, it will no longer match the signature, so any tampering can be detected. However, this property is sometimes not practical when considering distribution of images. For instance, lossy compression has to be performed to reduce the amount of data or signal processing is applied to correct gamma, to de-noise or to resample an image. These manipulations change the pixel values but not the content and hence not the authenticity.

Semi-fragile authentication methods for digital images were introduced to tolerate certain kinds of processing. For example, there are approaches quantizing the image or its transform coefficients to allow some small amount of pre-defined distortion. Other methods extract robust image features from the image such as edges, contours or zero-crossings whose correct existence is proved during the watermark verification process. The aim is to allow admissible manipulations such as JPEG compression, but to reject malicious manipulations, e.g., the addition or deletion of objects, which change the image content.

But the security has to be explicitly considered during the semi-fragile watermarking design process. As already noted by Fei et al. [2], many authentication frameworks lay to much emphasis on robustness, which brings into question security issues for authentication applications. Often, the image content is pretended to be secured by protecting only the correct existence of mean values of extensive pixel areas. An attack, intended to change the image content, can maliciously operate on these local pixel areas as long as the mean values are not changed. For example, an attacker is able to insert edges into areas of homogenous colour maintaining the mean values without raising an alarm when the authenticity is verified. The reason why most authentication systems do only protect local mean values is that taking more image content information into account results in an increased amount of data to be embedded. This, in turn, decreases the visual quality of the image. A secure alternative to strive after could be the use of cryptographically secure

hash functions mapping all important content dependent features to a small amount of bit information, which can be encrypted and embedded. To do so, invariant properties for signature generation as well as bit embedding are needed, because the output of classical hash functions alters dramatically if even one input bit is changed and hence no verification is possible.

In Section 2, we briefly review prior work on DCT-based semi-fragile watermarking and analyse the main concept of invariant JPEG-properties proposed by Lin and Chang. Experimental results show that these invariant properties are not that invariant as promised to be. Both the frequency of occurrence and the distribution of fluctuations are analysed and discussed in section 3. We suggest solutions in section 4 and present conclusions in section 5.

2. Previous Work

In [3], Lin and Chang proposed so-called invariant JPEG-properties for the first time. They suggested generating a signature for authentication purposes based on invariant relationships between DCT coefficients of the same position in two separate 8x8 blocks of an image. The relationship is alleged to be preserved when these coefficients are quantized in one or more JPEG re-compression processes. Formally, two theorems were stated:

1.) The magnitude relationship between two coefficients remains invariable through repetitive JPEG compression.

2.) A threshold is designated to protect the difference of the two coefficients, which is also proposed to remain invariant in a defined range.

For signature generation, the image is transformed to the DCT domain, resulting in M non-overlapping blocks consisting of 8x8 coefficients. A pseudo-random sequence is used to select $M/2$ non-overlapping sets of block pairs containing the coefficients $F_p(u, v)$ and $F_q(u, v)$, where $\forall\, u, v \in [0, .., 7]$ and $p, q \in [1, .., M]$. Based on the first theorem, the feature code bits $Z_1(\mu)$ are calculated as:

$$Z_1(\mu) = \begin{cases} 0 & \text{if } F_p(\mu) < F_q(\mu) \\ 1 & \text{if } F_p(\mu) \geq F_q(\mu) \end{cases} \quad (1)$$

where μ is the coefficient position in each of the two blocks. Considering the second theorem, in addition to the above equation, the feature code bits can be extended to predict the exact relationships of the DCT coefficients after compression up to a user-defined precision level N. To address the noise caused by integer rounding during the quantization process, the authors introduced an error margin τ to reduce the false alarm rate. Afterwards, the N feature code sets $Z_n(\mu)$ are concatenated, encrypted and embedded by using the following invariant properties [4]:

$$\text{If} \quad F'(\mu) = \text{Integer Round}\left(F(\mu)/Q'_0(\mu)\right)\cdot Q'_0(\mu), \quad (2)$$

$$\text{and} \quad \tilde{F}(\mu) = \text{Integer Round}\left(F'(\mu)/Q_r(\mu)\right)\cdot Q_r(\mu), \quad (3)$$

$$\text{then} \quad F'(\mu) = \text{Integer Round}\left(\tilde{F}(\mu)/Q'_0(\mu)\right)\cdot Q'_0(\mu). \quad (4)$$

Q'_0 is a pre-selected quantization table for JPEG lossy compression, whose quantization steps are larger than all quantization steps, Q_r, in subsequent JPEG compression. Equations 2-4 state that a modified coefficient $\tilde{F}_p(\mu)$, or in the same way the difference of two modified coefficients $\tilde{F}_p(\mu)$ and $\tilde{F}_q(\mu)$, can be exactly reconstructed after future JPEG compression, if $Q_r(\mu) \leq Q'_0(\mu)$. But in these equations there is no consideration to the factor that rounding and clipping is an essential part of JPEG compression. Rounding maps real values to integer ($\mathbb{R} \rightarrow \mathbb{Z}$) in the spatial domain as well as in the transform domain. Furthermore, clipping is necessary to limit the range of integer numbers according to the precision of the digital representation of the JPEG image data in the spatial domain. (8 bit/pixel: 0, 1, …, 255). These rounding and clipping errors yield big problems concerning invariant JPEG properties, as we will see in section 3.

Several attacks and improvement suggestions were proposed in [5 - 9]. Most of them use the weakness that, since watermarking schemes have a limited data embedding capacity [10], it is typical to only protect comparatively few image content information. In Lin and Chang's method, e.g., only the sign of the first ten block coefficient relationships in "zigzag"-order and one block mean value is included when the signature is generated. Attackers can delete or add objects to the image as long as they maintain this relationships below the given tolerance margin τ. An example is given in section 4 (Figure 9).

3. Rounding - JPEG's own attack

Caused by the limited data embedding capacity, only a few content features can be protected when using non-hash-based signature generation. Often, non-secure operations such as "XOR", "AND" or "averaging" concatenate several features to reduce the signature length [8, 11]. We claim that a semi-fragile watermarking system, which aims to be seriously accepted as a modern authentication solution has to incorporate cryptographic hash functions instead. But the problem of these hash functions is that any slightest change of the input data results in a totally changed output hash value. Hence, the input has to be based on invariant features, which do not change, e.g., the allegedly invariant JPEG properties proposed by Lin and Chang. Their idea is good but not optimally suited for hash-based signature generation, because different rounding and clipping processes during JPEG compression yield arbitrary bit fluctuations. Also, for the embedding of the signature bits invariant features are desirable.

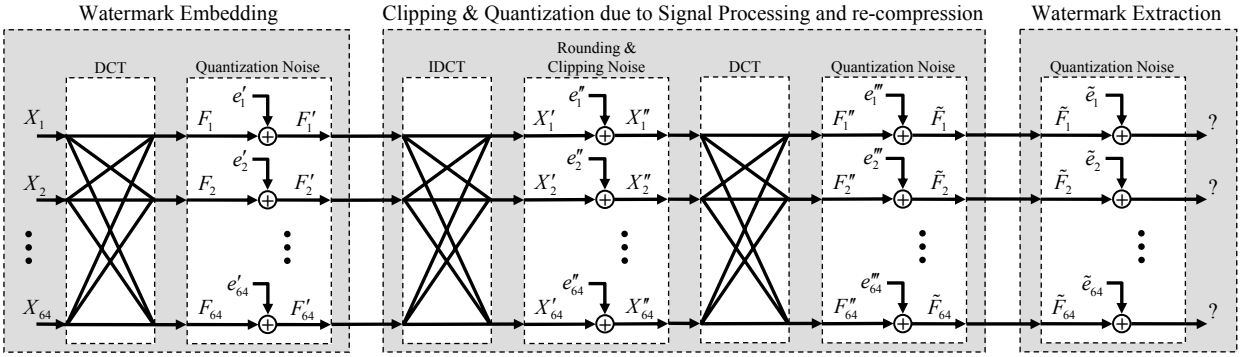

Figure 1. Clipping and quantization errors modeled as additive noise.

Rounding and clipping processes are necessary during JPEG compression or when a JPEG compressed image is simply loaded and saved in graphics software. In practice, to reduce the computational effort for JPEG compression often the DCT and IDCT transforms are calculated with finite precision. Sometimes, even the intermediate values are integers, so this kind of rounding error is difficult to gauge. Figure 1 indicates the rounding and clipping errors caused by the quantization both on the pixel values X_μ in the spatial domain and the DCT coefficients F_μ in the transform domain. Noise $|e| < 0.5$ is added to every DCT coefficient in the transform domain or to the signal samples in the spatial domain when quantization takes place.

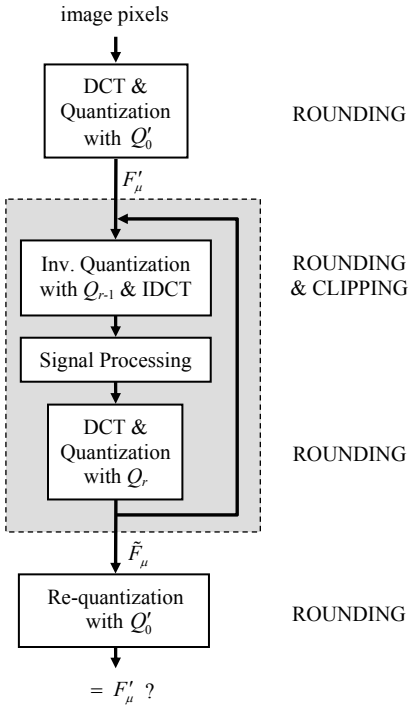

Figure 2. Multiple JPEG compression with different quantization steps $Q_r(\mu) \le Q'_0(\mu)$.

Figure 1 also shows the transformation from one domain to another through the coupling of the 64 subchannels, denoting the 8x8 block values. This coupling causes the rounding or clipping errors of one single coefficient or pixel value to spread over all other signal elements. Roughly speaking, in a worst case scenario the particular errors caused by rounding and clipping interfere. These interferences can have a significant influence on single coefficient values at the signature extraction and verification site. In their work, Lin and Chang consider the 64 coefficients in the 8x8 block as independent subchannels. But this assumption is not practical. Instead, a summation of particular error values can be expected. However, also under the hypothesis of independent subchannels, single coefficients or pairs of coefficients can change dramatically as well, when they are quantized more than one time. Figure 3 gives a numerical example of this kind of alterations, where a coefficient can be considered standalone or in differential form together with a second coefficient. The pre-quantized coefficient $F'_p = 16$ alters to the wrong value $F'_p = 32$ caused by two further quantization processes and hence the verification fails. This means that if single coefficients have disadvantageous values during signature generation using Lin and Chang's theorems, fluctuations occur, even though the condition $Q_r(\mu) \le Q'_0(\mu)$ is met.

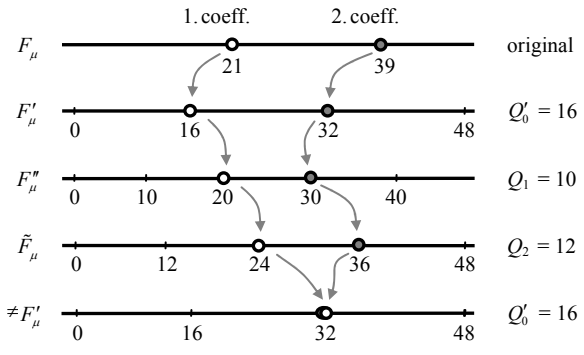

Figure 3. Numerical example of coefficient alterations caused by two further quantization processes, regardless of any kind of transformation or clipping errors.

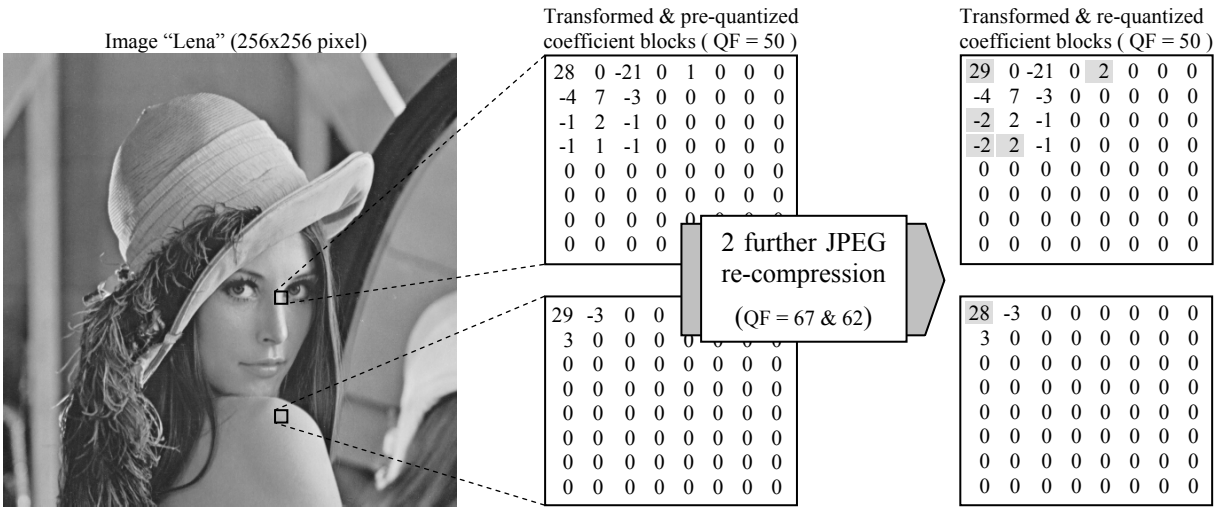

Figure 4. Practical example of DCT coefficient alterations caused by 2 further JPEG re-compression cycles.

Considering only the sign of a block pair relationship of two coefficients, or in other words, the first theorem, fluctuations occur if the coefficient differences are small enough. For example, in Figure 4, two 8x8 blocks from the 256x256 Lena test image are selected. Q'_0 for pre-quantization was chosen to be the quantization table for JPEG quality factor $QF = 50$. As can be seen, the difference between both pre-quantized DC coefficients is small. When transformed to the spatial domain, two further times re-compressed, once using $QF = 67$ and once again with $QF = 62$, afterwards, both DC coefficients do change their values. Equation 1 yields the signature bit "1" when pre-quantizing and after further compression the bit value "0". Other coefficients of the 8x8 blocks do change as well. In Figure 5, we demonstrate the alterations of both considered DC coefficients, firstly pre-quantized to $448/16 = 28$ and $464/16 = 29$, crossing over due to interfering errors. As opposed to the numerical example shown in Figure 3, here we have to proceed on the assumption that, since all 64 DCT transformation subchannel errors can interfere, "crossing coefficient alterations" are possible to occur.

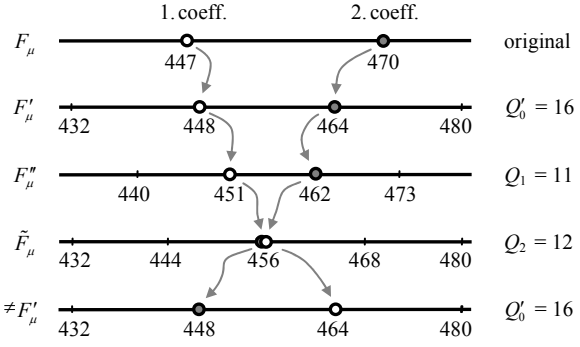

Figure 5. "Crossing coefficient alterations" caused by interfered rounding errors due to two further quantization processes according to the example shown in Figure 4.

Lin and Chang already mentioned that bit fluctuations are possible to occur. To overcome this problem they introduced the above mentioned margin τ to be used when extracting and verifying the signature bits. Also other authentication systems based on the allegedly invariant properties of JPEG compression, such as [2, 6, 11], aim to avoid detection errors by raising a tolerance margin for coefficient pairs with small differences.

When extracting and verifying the image features, Equation 1 is used in an extended version (Eqn. 5). Again, we only consider the case, protecting the sign of a block pair and not the exact difference of the relationship, because practical non-hash-based applications only use this kind of Lin and Chang's method to keep the signature length short. Assuming that the signature bits have been extracted correctly, the features can be verified as follows:

$$\begin{aligned}&\text{If } Z_1(\mu) = 0 \text{ and if } F_p(\mu) < F_q(\mu) + \tau,\\ &\quad\text{then "block pair relationship = correct",}\\ &\quad\text{else "block pair relationship = manipulated".}\\ &\text{If } Z_1(\mu) = 1 \text{ and if } F_p(\mu) > F_q(\mu) + \tau,\\ &\quad\text{then "block pair relationship = correct",}\\ &\quad\text{else "block pair relationship = manipulated".}\end{aligned} \tag{5}$$

This means that as long as the relationship of two coefficients is maintained, extended through the tolerance margin τ, no alarm is raised during verification process.

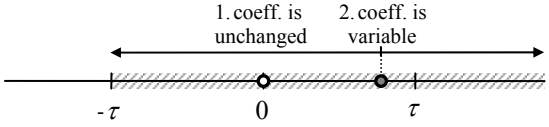

Figure 6. Range of tolerance for coefficient relationship.

The tolerance margin is suggested to be set zero, if no further JPEG re-compression is to be expected, or to be set to the value Q_0' to achieve high robustness.

By simulations on numerous test images, we have tried to determine the probability of occurrence of coefficient alterations. Firstly, the images were pre-quantized using QF=50 and transformed to the spatial domain. Afterwards, some JPEG re-compression cycles have been applied to the test images iteratively using all combinations of $QF = 50\ldots99$ for Q_1, Q_2 and Q_3. When transformed back to the DCT domain and re-quantized using $QF = 50$ again, multiple coefficients have changed their values, as can be seen in Figure 7 statistically. Additionally, we applied an exponential curve fitting to the experimental data.

The average number of coefficient errors per 8x8 block increases with the number of compression iterations. Roughly speaking, though the visual impact of any image is not even slightly changed due to multiple high quality JPEG re-compression, a lot of DCT coefficients can alter arbitrarily caused by interfered rounding and clipping errors. Figure 8 shows a log-probability scaled histogram-like distribution of the differential sizes of occurring coefficient errors. For example, ±1 means a difference of $\pm Q_0'$.

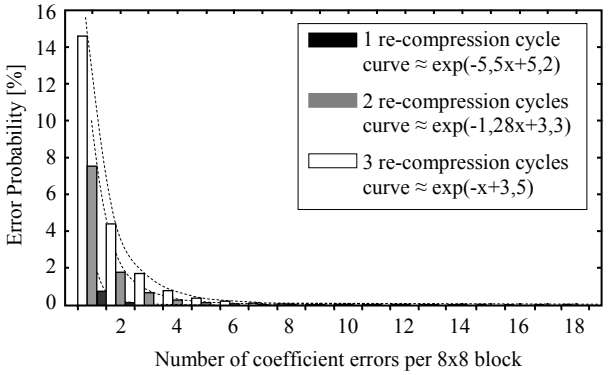

Figure 7. Probability that multiple coefficient errors occur in the same 8x8 DCT block.

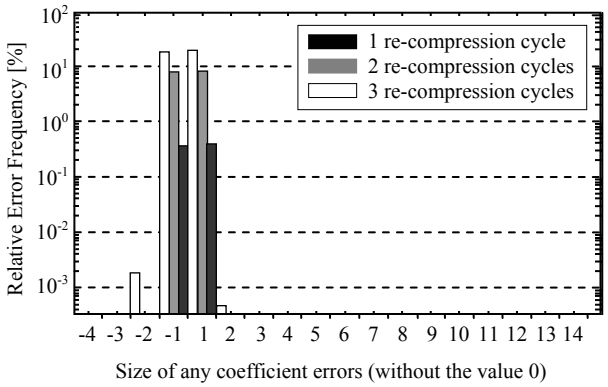

Figure 8. Histogram of the sizes of coefficient errors.

4. Hash-based signature generation

Since Lin and Chang's method only protects the sign relationships of the first few low-frequency DCT coefficients together with the mean values of the 8x8 image blocks, attacks can be applied changing the image content. For example, the following three 32x32 pixel images would yield the same generated signature bit stream and hence no alarm would be raised during signature verification. The mean values of every 8x8 block and their first 10 coefficient sign relationships remain constant. However, more sophisticated attacks are possible. For example, a specific attack could be possible maliciously adapting the coefficients to manipulated image content.

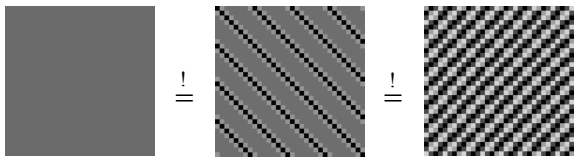

Figure 9. 3 sample images yielding the same signature.

Hence, we claim that more than only the sign relationships of the first few low-frequency coefficients and the block mean value should be protected by an authentication system. An attacker should not be able to change the image content undetected, even though the maliciously tampered image content (e.g., noise) makes no sense to an observer anymore.

Lin and Chang's method of feature code generation by using the two proposed theorems could be used in an extended manner to determine invariant image content dependent properties. Therefore, the practical use of the second theorem is indispensable to protect the real differences of coefficient relationships and hence to avoid attacks as in Figure 9. Furthermore, not only one mean value per 8x8 DCT block, or in other words the DC coefficient, should be protected. Also the numerical ranges of the 63 AC coefficients, respectively the frequency components, have to be considered. This should eliminate the possibility for an attacker to simply shift both coefficient values while maintaining the relationship difference. The huge amount of data due to this extended feature code generation must be used as input for a secure cryptographic hash function. Otherwise the image quality would be degraded too much during signature embedding. Afterwards, the small amount of hash output bits has to be asymmetrically encrypted as already suggested by Lin and Chang as well as to be embedded as robust as possible. For example in [12], the authors proposed ECC-based data hiding techniques for the DCT domain employing special criteria to guess robust embedding locations. The only problem left is by coefficient fluctuations, roughly speaking the not absolutely invariant JPEG properties, disturbing the hash functionality.

We have found out by simulations that the alterations of the 8x8 block coefficient values due to rounding and clipping processes are most often in the range [-1…+1]. Maybe, ECC could be one solution to this problem as well. For example, as we discussed in [13], ECC can work similar to a more sophisticated multidimensional vector-quantization with bit reconstruction capabilities, single errors are spread to multiple samples and hence the original bit values can be reconstructed.

In [14], we found out that the wavelet domain of JPEG-2000, because of its bit plane-oriented signal processing, is better suited for authentication watermarking purpose. Single pixel changes in the spatial domain are spread to multiple coefficients in different subbands in the wavelet domain. As a result, clipping and rounding errors in the spatial domain do not affect single transform coefficients as much as in the case of JPEG-based approaches. Further, JPEG2000 re-compression requires no re-quantization of the wavelet coefficients with a different quantization step size, since the quantization interval is always a multiple of two. Hence, no interfering rounding errors can occur when the image is JPEG2000 re-compressed.

5. Conclusion

We analyzed the authentication framework proposed by Lin and Chang, which is based on several allegedly invariant properties of the JPEG compression process. Lossy JPEG compression to a pre-defined quality factor without any false alarm regardless of the number of compression iterations was promised to be accepted. But we proved that the DCT coefficients as well as relationships between pairs of DCT coefficients used for signature generation and embedding alter dramatically due to further JPEG re-compression. We showed that the used signature generation is not secure. If someone is intended to use secure hash-based signature generation instead, problems occur due to coefficient alterations. By simulations on numerous standard test images, we determined the probability that coefficient errors occur at the signature verification site caused by rounding and clipping processes as an essential part of commonly used JPEG compression. Finally, from the results of our examination we concluded possible solution suggestions.

References:

[1] Technical Institute of Standards and Technology (NIST), Digital signature standard (DSS), *Tech. Rep.*, *FIPS PUB 186-2*, 2000.

[2] C. Fei, D. Kundur and R. Kwong, Analysis and Design of Secure Watermark-based Authentication Systems, *IEEE Trans. Signal Processing Supplement on Secure Media*, accepted for publication 2004, to appear.

[3] C. Y. Lin and S.-F. Chang, A Robust Image Authentication Method Surviving JPEG Lossy Compression, *Proc. of SPIE, 3312*, San Jose, CA, 1998, 296-307.

[4] C. Y. Lin and S.-F. Chang, Semi-fragile watermarking for authenticating JPEG visual content, *Proc. of SPIE, 3971*, San Jose, CA, 2000, 140-151.

[5] R. Radhakrishnan and N. Memon, On the security of digest function in the SARI image authentication system, *IEEE Trans. Circuits Systems Video Technology, 12*, 2002, 1030-1033.

[6] K. Maeno, Q. Sun, S.-F. Chang and M. Suto, New Semi-Fragile Image Authentication Watermarking Techniques Using Random Bias And Non-Uniform Quantization, *Proc. of SPIE, 4675*, 2002, 659-670.

[7] T. Uehara and R. Safavi-Naini, On (In) security of 'A Robust Image Authentication Method', *Proc. 3rd IEEE Pacific-Rim Conf. on Multimedia*, Hsinchu, Taiwan, 2002, 1025-1032.

[8] J. Wu, B. B. Zhu, S. Li and F. Lin, New Attacks on SARI Image Authentication System, *Proc. of SPIE, 5306*, San Jose, CA, 2004, 602-609.

[9] C.-K. Ho and C.-T. Li, Semi-fragile Watermarking Scheme for Authentication of JPEG Images, *Proc. ITCC, 1*, Las Vegas, USA, 2004, 7-11.

[10] P. H. W. Wong and O. C. Au, A capacity estimation technique for JPEG-to-JPEG image watermarking, *IEEE Trans. Circuits Syst. Video Techn. 13*(8), 2003, 746-752.

[11] Y. Zhao, P. Campisi and D. Kundur, Dual Domain Watermarking for Authentication and Compression of Cultural Heritage Images, *IEEE Trans. Image Processing, 13*(3), 2004, 430-448.

[12] K. Solanki, N. Jacobsen, S. Chandrasekaran, U. Madhow and B. S. Manjunath, Robust Image-Adaptive Data Hiding Using Erasure and Error Correction, *IEEE Trans. Image Processing 13*(12), 2004, 1627-1639.

[13] M. Schlauweg, D. Pröfrock, T. Palfner and E. Müller, Quantization-based semi-fragile public-key watermarking for secure image authentication, *Proc. of SPIE, 5915,* San Diego, CA, 2005.

[14] T. Palfner, M. Schlauweg and E. Müller, A Secure Semi-fragile Watermarking Algorithm for Image Authentication in the Wavelet Domain of JPEG2000, *Proc. 2nd Intern. Conf. on Innovations in Information Technology*, Dubai, UAE, 2005.