Robust Message-Privacy Preserving Image Copy Detection for Cloud-based Systems

M. Diephuis, S. Voloshynovskiy*, O. Koval and F. Beekhof
Stochastic Information Processing Group
Université de Genève
1227 Carouge, Switzerland
Email: {Maurits.Diephuis, svolos, Oleksiy.Koval, Fokko.Beekhof}@unige.ch

***Abstract*—In this paper we propose an architecture for message-privacy preserving copy detection and content identification for images based on the signs of the Discrete Cosine Transform (DCT) coefficients. The architecture allows for searching in encrypted data and places the computational burden on the server. Sign components of the low frequency DCT coefficients of an image are used to generate a dual set of keys that in turn are used to encrypt the source image and serve as a robust hash that can be queried for content identification. The statistical properties of these DCT sign vectors are modelled and we investigate the influence of a set of real world image distortions on them. Finally, the trade-off between the discriminative power of such vectors, the offered security and the resilience against errors is demonstrated.**

I. Introduction

The last decade has seen a huge rise in massive centralized multimedia repositories accessible via the Internet, known in popular language as 'the cloud'. Example services include Microsoft Azure, Apple iCloud, Dropbox, and recently risen to notoriety, Megaupload. The latter illustrates the need for copyright verification, image-content identification, copy-detection, privacy and authentication perfectly. Additionally, these online repositories are increasingly accessed by so called thin clients such as mobile phones which still lack significant processing power. Finally, storing personal data on third party cloud servers gives rise to many privacy and security concerns. This necessitates the protection of data and the ability to process multimedia data in the protected domain efficiently. Given this situation, we advocate an architecture that enables searching for an image in the encrypted domain in which the computational burden lies on the server side.

Multimedia security broadly falls into three categories: systems based on classic cryptography [1], [2], those based on watermarking [3], and on content fingerprinting. Watermarking is traditionally deployed for copyright measures [4], [5]. Monitoring and indexing of content is the domain of content fingerprinting [6] while visual encryption has been used to achieve access control and to preserve user privacy within systems [16]. Notable attempts have been made to introduce signal processing into the cryptographic domain [7].

Multimedia security and third party storage distribution, or the cloud, imposes a number of requirements to multimedia encryption that differ significantly from classical cryptography. Currently it is believed that these requirements can be satisfied based on *perceptual encryption*. This type of encryption enables the addition of security measures to an architecture while simultaneously allowing for searching and copy detection.

We propose a relatively simple architecture for images based on the sign of low frequency DCT coefficients [8]. For visual scrambling based on DCT [9], [10] the sign of the coefficients is vastly more significant than their magnitudes since it contains a coarsely quantized image phase. In our architecture, the vector with the sign components from the low frequency DCT image transform coefficients is used as an input for two functions. The first function expands the vector to form a key that drives the XOR based encryption. The second function is a hash which is used as a robust identifier in the Database. Enrolled images are scrambled prior to storage with that key.

This paper is organized as follows. Section II introduces the formal problem, the statistical model that will be deployed for evaluation of the architecture and the architecture itself. Statistical evaluation of the DCT sign vectors, their discriminative power and resilience to noise is done in Section III. Security evaluation is the subject of Section IV. Finally, Section V concludes the paper.

Notation Scalar random variables are designated by capital letters X, and bold capitals $\mathbf{X}$ denote vector random variables. Corresponding small letters x and $\mathbf{x}$ denote their respective realizations, where $\mathbf{x} = \{x[1], x[2], \ldots, x[J]\}$. The binarized version of $\mathbf{x}$ is represented by $\mathbf{b}_\mathbf{x}$. $X \sim p(x)$ indicates that the random variable follows $p_X(x)$. $\mathcal{B}$ indicates the Bernoulli distribution. $E(.), \Psi_1(.), \Psi_2(.)$ and $Q(.)$ denote Encryption, Hash, Mapper, and Quantization functions. Outputs of $\Psi_1(.)$ and $\Psi_2(.)$ are $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$. The DCT transform of an image $\tilde{\mathbf{x}}$ with index m is denoted by $\mathbf{x}(m)$. Encrypted DCT coefficients from that image are $\mathbf{c}(m)$. Vectors with binarized DCT coefficients are expressed as $\mathbf{b}_{\mathbf{x}(m)}$. If multiple k such binarized DCT coefficients vectors are generated from the same image $\tilde{\mathbf{x}}(m)$, one uses: $\mathbf{b}_{\mathbf{x}^k(m)}$, $k \in \{1, 2, \ldots, K\}$.

II. Problem formulation and Methodology

A. Architecture and Problem formulation

The basic architecture for image enrollment and copy detection is shown in Figure 1. For enrollment an image $\tilde{\mathbf{x}}(m)$ is converted via the DCT into a matrix $\mathbf{x}(m)$ which contains its DCT coefficients. From this matrix the $N \times N$ top left, low

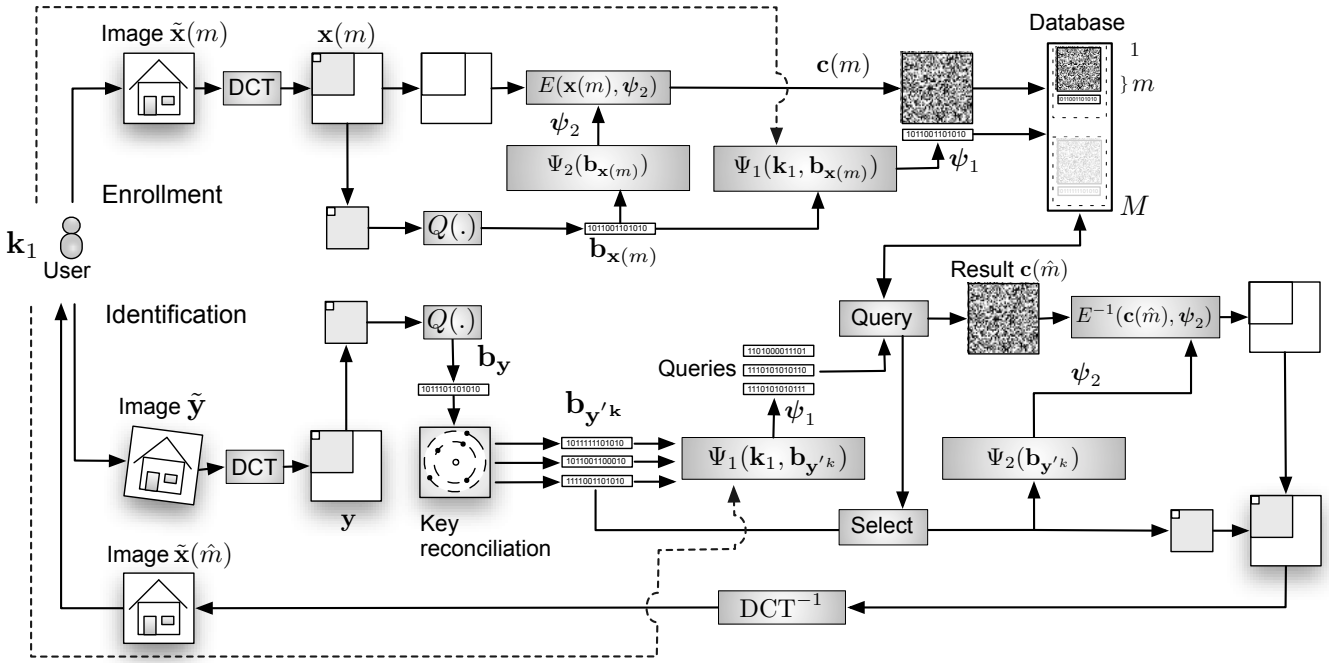


Figure 1: Enrollment and copy-detection framework.

frequency components are copied and the DC component is removed resulting in $L = (N \times N) - 1$ elements. In matrix $\mathbf{x}(m)$, these $N \times N$ elements are then given a random sign. The copied L low frequency components are binarized by the $sign(.)$ function, which is a particular case of quantizer $Q(.)$, as follows:

$$\begin{aligned}\mathbf{b}_{\mathbf{x}(m)} =& \{sign(x(m)[1]), sign(x(m)[2]),\\ & \ldots, sign(x(m)[L])\},\end{aligned} \tag{1}$$

where for $i \in \{1 \ldots L\}$, $b_{\mathbf{x}(m)}[i] \in \{0, 1\}$ and $\forall a, sign(a) = 1$ if $a \geq 0$ and 0, otherwise. The resulting binary vector $\mathbf{b}_{\mathbf{x}(m)}$ is used in two ways. First, it serves as a seed for mapper $\Psi_2(.)$ that generates the matrix needed for encryption. The encryption function $E(.)$ is a simple XOR against all DCT coefficients thus randomly flipping their signs. The mapper $\Psi_2(.)$ serves to generate a random pattern $\boldsymbol{\psi}_2$ big enough to XOR the DCT coefficient signs. The result $\mathbf{c}(m)$ is stored in the database. Furthermore, $\mathbf{b}_{\mathbf{x}(m)}$ is hashed with cryptographic hash-function $\Psi_1(.)$ and an optional secret key $\mathbf{k}_1$ to form the database key $\boldsymbol{\psi}_1$ belonging to this particular image entry m. Hash function $\Psi_1(.)$ ensures that the server can not access the L low frequency coefficients itself.

Copy detection of a query image is done along similar lines. An image $\tilde{\mathbf{y}}$ is presented to the system. It is transformed by DCT to form DCT coefficient matrix $\mathbf{y}$. Again the $N \times N$ low frequency components are extracted, and $(N \times N) - 1$ components are binarized to form $\mathbf{b}_{\mathbf{y}}$. Using the search procedure [11], [12] a number of candidate keys is generated from the Hamming sphere around $\mathbf{b}_{\mathbf{y}}$. In our case we will refer to this procedure as *key reconciliation*. The procedure simply flips bits in the vector within a certain sphere bound, starting with the least reliable bit first. It was shown in [11] that $O(L2^{LP_b})$ database lookups are required to find the right image, where $H_2(P_b) = -P_b \log_2(P_b) - (1 - P_b) \log_2(1 - P_b)$ and P_b is the probability of bit error [13]. All these combinations are hashed using hash function $\Psi_1(.)$ and presented as a query to the database. The server returns the encrypted DCT coefficients belonging to that identifier in case of a perfect match. The user then uses his own vector $\mathbf{b}_{\mathbf{y}}$ and mapper $\Psi_2(.)$ to make a bitmask to XOR the returned DCT coefficients and to reconstruct the image by inserting them back in the DCT matrix. The final image can then be obtained via the inverse DCT transform.

B. Fingerprint model

Central to the characteristics and performance of this architecture is the length of the vector $\mathbf{b}_{\mathbf{x}(m)}$ with the $(N \times N) - 1$ extracted DCT coefficient sign components. Different choices for N will be statistically modelled and discussed. We will show in Section III that extracted DCT coefficient sign vectors behave asymptomatically as i.i.d. vectors drawn from $B(p)$ where $p = 0.5$ for certain values of L. The longer the vector, the larger its potential discriminative value. As the DCT transform organizes the resulting matrix coefficients by frequency, the first elements in $\mathbf{b}_{\mathbf{x}(m)}$ represent the lower frequencies from an image $\tilde{\mathbf{x}}(m)$ and will therefore be more

resilient to distortions in the image domain, as outlined in Section III.

III. Statistical properties and simulation

In order to analyse the behaviour of the DCT sign vectors a digital communications approach is taken in which various distortions in the original image domain are reflected in the binary domain as bit flips, characterized by the probability of bit error, P_b [13]. DCT sign vectors are ascertained as outlined by Section II.

The discriminative power of the designed fingerprints will be investigated. Two principal issues will be analysed. Firstly, the relation between bits within each individual DCT sign vector will be ascertained in the distortion-free regime. Secondly, we measure P_b for DCT coefficient sign vectors that originate from an image, and a degraded version of this image, in order to demonstrate how distortions impact the fingerprints their discriminative power. From these results the upper bound for computational complexity is determined. Finally, the performance of the system as a whole is evaluated.

The used dataset originates from [14] and consists of 1338 uncompressed TIF images.

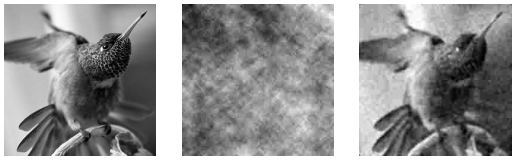
A. Descriptor statistics

In order to assess the discriminative power of the proposed fingerprint system, the statistics of the binary fingerprints are analysed. It is well known that the maximum amount of uniquely distinguishable typical sequences generated from a given binary stationary source is limited by $2^{H(\mathbf{B_x})}$, where $H(\mathbf{B_x})$ is the joint entropy of L random variables $B_{\mathbf{x}_i}$, $i \in \{1, 2, \ldots, L\}$. One needs to accurately estimate $H(\mathbf{B_x})$ to evaluate this bound. However, estimating such a function of a joint distribution is extremely complex in a high dimensional space. The other extreme case corresponds to the assumption that the bits are jointly independent, i.e. $H(\mathbf{B_x}) = \sum_{i=1}^{L} H(B_{\mathbf{x}_i}) = LH(B_{\mathbf{x}})$. This situation only requires the estimation of the marginal distributions. This approach while lacking a certain accuracy, provides an upper bound on the sought discriminative power, according to the chain rule for entropy [13]. Another information-theoretic driven approach, used in biometrics [15], consists in the estimation of $H(\mathbf{B_x})$ under the pair-wise independence assumption. Following the above mentioned justification, the two latter approaches are presented in this paper.

The estimates of $H(\mathbf{B_x})$ for joint i.i.d. assumptions and for the pair-wise bit independence are presented in Figures 2a and 2b for fingerprint lengths of $(N \times N) - 1$ where $N \in \{8, 16, 32\}$. The presented results imply that the extracted fingerprints possess asymptotic pair-wise independence, corresponding to a pair-wise entropy of 2 bits. Furthermore they confirm the accuracy of the independence bound.

B. Channel Distortion Statistics

Since distortions in the original image domain are reflected in the binary domain as bit flips, the probability of bit-error P_b will be used as performance metric. Channel distortions were tested for JPEG lossy compression with quality factor $\{10, 20, \ldots, 100\}$, Additive White Gaussian Noise (AWGN) with variance $\sigma^2 = \{0, 1, \ldots, 5\}$, blur with a disc size of $\theta = \{1, 2, \ldots, 5\}$ and finally scaling with parameter $s = \{0.25, 0.5, \ldots, 2\}$. All tests over all parameters were performed on DCT component sign vectors with a size of $N \times N$ where $N = \{8, 16, 32\}$ with the exception of the DC component. Shown in Figure 4 are the normalized histograms of the probability of error, P_b, between the original and the distorted DCT sign vectors for those distortion parameters resulting in the worst performance.



(a) Original image. (b) Encrypted and compressed. (c) Decompressed and decrypted image.

Figure 3: 3a to 3c show an example cycle in which an image was encrypted, lossy compressed with JPEG $q = 10$ in the encryption domain, decompressed and decrypted.

A direct consequence of the fact that the DCT magnitude coefficient components are untouched and only their signs are flipped by the encryption function $E(.)$, is that the encrypted coefficients of $\mathbf{c}(m)$ are reasonably resilient to distortions such as JPEG compression. An example where $\mathbf{c}(m)$ has been maliciously compressed with JPEG quality 10 can be seen in Figure 3.

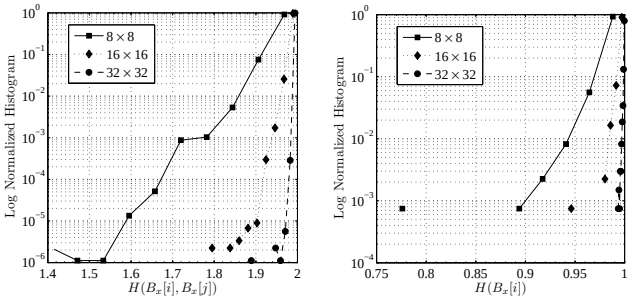
The discriminative power of the DCT sign vectors will be evaluated using the binary entropy function $H_2(.)$, since the maximal amount of identifiable sequences is upper-bounded by $2^{L(1-H_2(P_b))}$. Disregarding histogram equalisation (Figures 4m, 4n and 4o), which with a P_b value of 0.3 proved to be the most destructive, the largest P_b attained value in the real domain was of 0.09 for scaling (Figure 4j). The upper-bound for the number of distinguishable sequences therefore drops from the theoretical maximum of 2^L where $H_2(P_b) = 0$ to an upper bound of $2^{L(1-H_2(0.09))} \approx 2^{36.1}$ for $L = 64$.

C. Key reconciliation procedure

Given the worst expected bit error $P_b = 0.09$ induced by the worst case channel distortions between $\mathbf{b}_{\mathbf{x}(m)}$ and $\mathbf{b}_{\mathbf{x}(\hat{m})}$, the expected Hamming sphere radius is $LP_b = 5.8$. The complexity of the key reconciliation procedure is upper bounded by $O(L2^{LP_b}) \leq O(2^{12})$ which is computationally feasible.

D. Performance analysis

In this section we present results of the performance analysis of the proposed copy-detection system based on binary fingerprints. To this effect the *intra*-class P_b and the *inter*-class P_b are determined. The *intra*-class P_b are all differences



(a) $H(B_{x[i]}, B_{x[j]})$ for $(N\times N)-1$ DCT sign coefficients where $N \in \{8, 16, 32\}$.

(b) $H(B_{x(m)}[i])$ for $(N \times N) - 1$ DCT sign coefficients where $N \in \{8, 16, 32\}$.

Figure 2: 2a shows the pair-wise entropy $H(B_x[i], B_x[j])$ for $i \neq j$. 2b shows the marginal entropy $H(B_x[i])$.

between fingerprints of length L originating from an image and a distorted copy, as detailed in Section III. The *inter*-class P_b are the differences between DCT sign coefficient vectors originating from different images. It is determined by taking all M images from our dataset and determining the P_b between their DCT sign coefficient vectors. This gives $\binom{M}{2}$ combinations.

Figure 5 shows the *intra*-class P_b induced by channel distortions and the *inter*-class P_b for fingerprint lengths of $(N \times N) - 1$ where $N \in \{8, 16, 32\}$. Even in the worst case for $N = 8$ these estimated probability mass functions do not overlap which proves that in this configuration this architecture can identify all copies error-less. This point is further illustrated by the Receiver Operating Characteristic (ROC) graphs shown in 5d, 5e and 5f where $P_d = \Pr[\hat{m} = m]$ and $P_{fa} = \Pr[\hat{m} \neq m]$.

IV. Security

The security assessment of the developed copy detection system will be performed within the recently developed framework of so-called *message privacy* (MP) security [16]. The main motivation to deviate from the classical cryptography approach is the evident incapability with this multimedia encryption scheme to satisfy the main cryptographic requirements. Furthermore, the encryption key is a function of the plain text.

According to the developed paradigm, security is related to the fidelity with which multi-media data can be recovered. Following such an approach, a multimedia encryption system is considered insecure if a high quality reconstruction of the image, giving its scrambled version, is possible. Pending the specific application different quality reconstructions will constitute a break.

For this purpose a number of tests was devised that all aim at reconstructing an image from its scrambled version where the attacker has access to different amounts of original data. The attacker has access to all the original magnitudes of all DCT coefficients from the database. Furthermore, we assume that the attacker has obtained secret key $\mathbf{k}_1$ and $N' \times N'$ bits of the lowest frequency DCT sign coefficients, where $N' \in \{8, 16, 32\}$. Our experiments demonstrate that perceptual recognition becomes possible when $N = 32$ although the PSNR value between the original and the reconstructed image remains as low as 20.0 dB. The results of the data recovery when all the information is leaked due to a Database compromise is shown in Figure 6c and confirms the MP security potential of the proposed scheme.

V. Conclusions

In this paper, a relatively simple architecture is presented for searching for identical image copies based on the signs of DCT image coefficients. It is shown that binary vectors from DCT sign coefficients exhibit high discriminative power in the distortion free regime as well as in the case in which the query has been degraded by some form of channel distortion. Finally the MP security of the system is assessed and it is demonstrated that it is infeasible to attain a high quality reconstruction of the original, even when the attackers exploits significant information leakage.

All code, data and documentation will be published online upon acceptance on `http://sip.unige.ch/`.

VI. Acknowledgment

This work is partly funded by SNF-grant 20021-132337.

References

[1] A. Uhl and A. Pommer, *Image and Video Encryption Image and Video Encryption. From Digital Rights Management to Secured Personal Communication*. Springer-Verlag, 2005, vol. Advances in Information Security, no. 15.

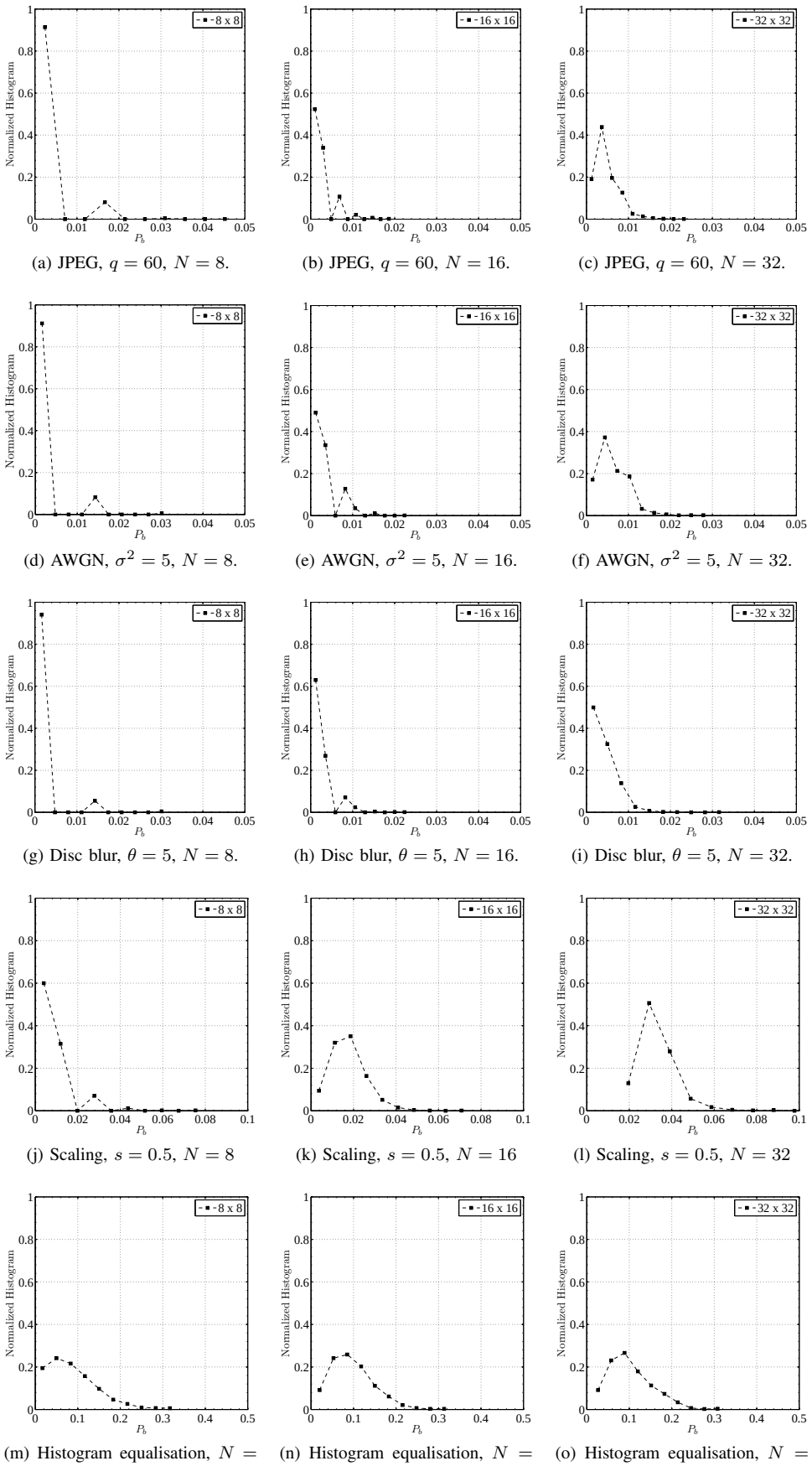
[2] B. Furht, E. Muharemagic, and D. Socek, *Multimedia Encryption and Watermarking*. Berlin, Heidelberg: Springer-Verlag, 2005, vol. Multimedia Systems and Applications, no. 28.

[3] E. B. Corrochano, Y. Mao, and G. Chen, "Chaos-based image encryption," pp. 231–265, 2005.

[4] L. Pérez-Freire and F. Pérez-González, "Spread-spectrum watermarking security," *Trans. Info. For. Sec.*, vol. 4, pp. 2–24, March 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1651156.1651158

[5] I. Cox, M. Miller, J. Bloom, and C. Honsinger, *Digital watermarking*, 2002, vol. 11.

[6] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system with an efficient search strategy," *Journal of New Music Research*, vol. 32, no. 2, pp. 211–221, 2003. [Online]. Available: http://dx.doi.org/10.1076/jnmr.32.2.211.16746



(a) JPEG, $q = 60$, $N = 8$. (b) JPEG, $q = 60$, $N = 16$. (c) JPEG, $q = 60$, $N = 32$.

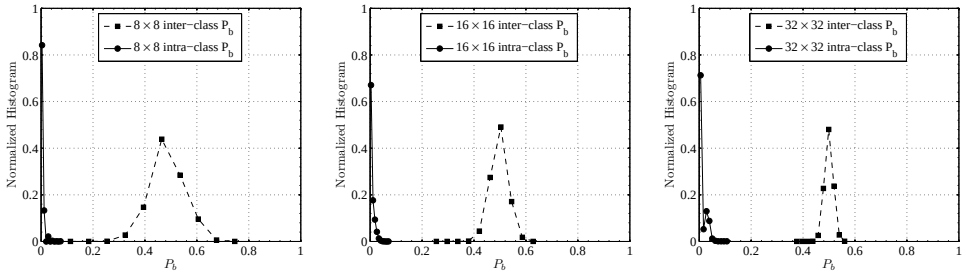
(d) AWGN, $\sigma^2 = 5$, $N = 8$. (e) AWGN, $\sigma^2 = 5$, $N = 16$. (f) AWGN, $\sigma^2 = 5$, $N = 32$.

(g) Disc blur, $\theta = 5$, $N = 8$. (h) Disc blur, $\theta = 5$, $N = 16$. (i) Disc blur, $\theta = 5$, $N = 32$.

(j) Scaling, $s = 0.5$, $N = 8$ (k) Scaling, $s = 0.5$, $N = 16$ (l) Scaling, $s = 0.5$, $N = 32$

(m) Histogram equalisation, $N = 8$. (n) Histogram equalisation, $N = 16$. (o) Histogram equalisation, $N = 32$.

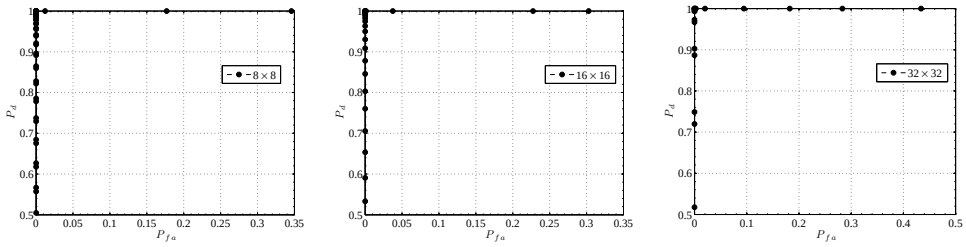
Figure 4: Channel statistics for various distortions in the real domain, where **x** and **y** denote the codebook the DCT sign vectors originating from an original image and a distorted copy. Shown are the normalized histograms of the probability of error, P_b between **x** and **y**.



(a) Performance for $(8 \times 8) - 1$ quantized DCT coefficients.

(b) Performance for $(16 \times 16) - 1$ quantized DCT coefficients.

(c) Performance for $(32 \times 32) - 1$ quantized DCT coefficients.

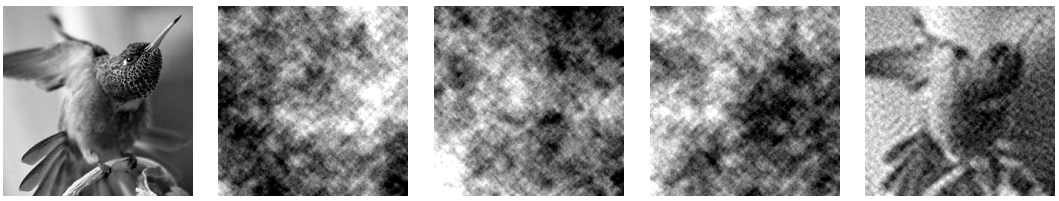


(d) Performance for $(8 \times 8) - 1$ quantized DCT coefficients.

(e) Performance for $(16 \times 16) - 1$ quantized DCT coefficients.

(f) Performance for $(32 \times 32) - 1$ quantized DCT coefficients.

Figure 5: P_b statistics for all permissible channel distortions between quantized DCT coefficients originating from identical images and their distorted copy (*intra*-class P_b) versus the *inter*-class P_b statistics from quantized DCT coefficients originating from different images.



(a) Original image. From this image the first 8×8 signs are used for encryption.

(b) Encrypted image.

(c) Restoration setting all signs randomly; PSNR = 10 dB.

(d) Restoration using the first 8×8 original signs; PSNR = 15 dB.

(e) Restoration using the first 32×32 original signs; PSNR = 20 dB.

Figure 6: Worst case example, to illustrate that it is not possible to descramble without significant original data. All reconstructions use the original magnitude of the DCT coefficients. Unknown signs are set uniformly to {-1, 1} randomly. Visually, the image starts to be human distinguishable if the first 32×32 sign components are known.

[7] Z. Erkin, "Secure signal processing: Privacy preserving cryptographic protocols for multimedia," Ph.D. dissertation, Delft University of Technology, Delft, 06/2010 2010.

[8] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *Computers, IEEE Transactions on*, vol. C-23, no. 1, pp. 90 –93, jan. 1974.

[9] C. Shi and B. Bhargava, "An efficient mpeg video encryption algorithm," pp. 381 –386, oct 1998.

[10] C. Shi, S. yih Wang, and B. Bhargava, "Mpeg video encryption in real-time using secret key cryptography," 1999.

[11] S. Voloshynovskiy, O. Koval, F. Beekhof, F. Farhadzadeh, and T. Holotyak, "Information-theoretical analysis of private content identification," in *IEEE Information Theory Workshop, ITW2010*, Dublin, Ireland, Aug.30-Sep.3 2010.

[12] F. Beekhof, S. Voloshynovskiy, O. Koval, and T. Holotyak, "Fast identification algorithms for forensic applications," December 6–9 2009.

[13] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.

[14] G. Schaefer and M. Stich, "Ucid-an uncompressed colour image database," *Proc. SPIE, Storage and Retrieval Methods and Applications for Multimedia*, vol. 5307, pp. 472–480, 2004.

[15] Y. Sutcu, S. Rane, J. Yedidia, S. Draper, and A. Vetro, "Feature extraction for a slepian-wolf biometric system using ldpc codes," Patent, 2008.

[16] D. Engel, T. Stütz, and A. Uhl, "Analysis of jpeg 2000 encryption with key-dependent wavelet packet subband structures," 2010.