TWO-LAYER FRAGILE WATERMARKING METHOD FOR ENHANCED TAMPERING LOCALISATION

Sergio Bravo-Solorio[*†] *and Asoke K. Nandi*

Department of Electrical Engg. & Electronics. The University of Liverpool Brownlow Hill, Liverpool, L69 3GJ, UK; a.nandi@liverpool.ac.uk, S.Bravo-Solorio@warwick.ac.uk

ABSTRACT

This paper presents a new fragile watermarking method, whereby a secure block-wise and an interlaced watermarking mechanisms are hierarchically structured to provide higher tampering localisation capabilities. The block-wise method provides security against conventional distortions, cropping, as well as sophisticated attacks, whereas the interlaced scheme is aimed at enhancing the localisation accuracy achieved by the block-wise method. Results are presented to illustrate the improved localisation capabilities of the proposed method in comparison with some five existing watermarking schemes.

***Index Terms*—** Fragile watermarking, tampering localisation, authentication.

1. INTRODUCTION

The availability of image editing software equipped with sophisticated tools has sparked serious concerns about the integrity of digital images, especially in application domains where sensitive information is handled – e.g. law enforcement applications. Fragile watermarking describe techniques to embed information imperceptibly – i.e. a *watermark* – in digital images, so that detected changes in the watermark indirectly expose manipulations in the *host* image [1].

Fragile watermarking systems can produce binary (yes/no) answers to state whether or not a host image has been tampered with. Nonetheless, the fact that a watermark undergoes the same transformation as the host image opens up the possibility of identifying the tampered region, while verifying the remainder of the image. This important feature is commonly referred to as *tampering localisation* (or simply *localisation*). Identifying the altered pixels correctly could be useful to evaluate whether the semantic meaning of an image has been affected significantly. Furthermore, depending on the nature of the distortion, it may be possible to restore the altered content by means of denoising or inpainting techniques, e.g. [2].

Wong [3] proposed a scheme, whereby a message authentication code (MAC) is independently generated from the seven most significant bit-planes (MSBPs) of every non-overlapping pixel-block in the image. Then, the least significant bit-plane (LSBP) of each block is replaced with the MAC derived from the block itself. This method provides an effective localisation, but is susceptible to vector quantisation (VQ) attacks [4]. That is, the blocks of a host image can be swapped without being noticed or, even worse, blocks copied from a set of images watermarked with the same key can be assembled together to form a completely new image that would go undetected.

[*]Sergio Bravo-Solorio is supported by the National Council of Science and Technology (CONACyT) of Mexico.

[†]Sergio Bravo-Solorio is now with the Department of Computer Sciences at the University of Warwick.

To localise distortions, while effectively thwarting VQ attacks, Fridrich [5] proposed to make the MAC dependent on a block-index and a unique image index. Lin *et al.* [6] used the six MSBPs of every pixel-block to encode a MAC, which is embedded in the two LSBPs of the subsequent block in a block-mapping sequence generated pseudo-randomly. A hierarchical mechanism is adopted to localise and even recover the altered blocks. However, the correlation between the blocks can be estimated to perform successful attacks [7]. In [8], a sparse set of wavelet coefficients are watermarked in accordance with a contextual non-deterministic dependence mechanism, which involves all the wavelet coefficients across the wavelet decomposition levels. The aim is to protect all the pixels without altering all the wavelet coefficients. He *et al.* [9] proposed to map every wavelet coefficient, in the coarser sub-band, to a 4-bit code, which is then embedded in the 2×2 pixel-block corresponding to another wavelet coefficient selected pseudo-randomly. At the receiver side, possibly distorted codes are localised in a bitmap, and then a post-processing mechanism is adopted to remove isolated blocks. In [10], the MAC, derived from the five MSBs of every single pixel and a pseudo-random code, is embedded into the three LSBPs of the image. At the receiver end, two distributions, corresponding to the altered and genuine pixels, are employed to localise pixels corrupted in their five MSBs. This method manages to localise altered pixels accurately only if the altered area is not too extensive. In [11], a secret key is used to generate a circular block-mapping sequence. The MAC computed for a block is embedded in the subsequent block in the sequence. One of the four least significant bits (LSBs) in every pixel is pseudo-randomly selected to allocate a bit of the MAC.

In this paper, a new method is presented to provide enhanced tampering localisation. The proposed method relies on two watermark layers. One is embedded by means of a secure block-wise method resilient to cropping. A second watermark, which is spread over the image following an interlaced arrangement, is used to refine the localisation achieved at the first layer. Results demonstrate that the proposed two-layer (TL) fragile watermarking method outperforms five state-of-the-art methods, in terms of localisation performance. The TL method is detailed in Section 2, while Section 3 elaborates on the embedding distortion and the security aspects of the scheme. Some results are presented in Section 4 and the paper is concluded in Section 5.

2. PROPOSED METHOD

2.1. Embedding process

Consider an $n_1 \times n_2$ grey-scale image X divided into non-overlapping blocks of $m \times m$ pixels, where $m = 2c_1$ for some $c_1 \in \mathbb{N}$; additionally, it will be assumed that $n_1 = c_2 m$ and $n_2 = c_3 m$, for any $c_2, c_3 \in \mathbb{N}$. Let X_p be the p-th block in X, for $p = 1, \ldots, m_{\text{bw}} =$

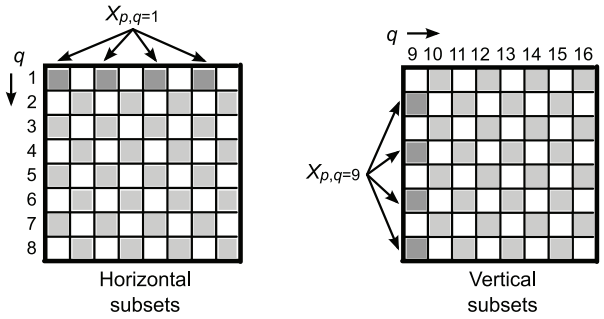


Fig. 1: Arrangement of the subsets of pixels in the block X_p.

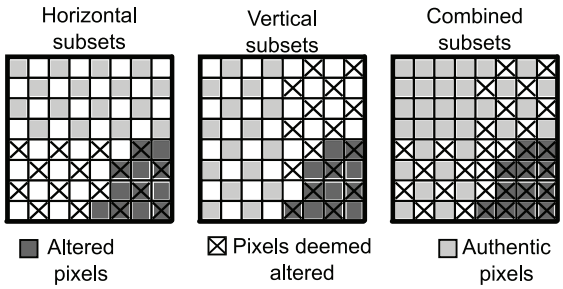


Fig. 2: Refinement mechanism in a partially altered pixel-block.

n_X/m^2, where $m_{\text{bw}} = (n_X/m^2)$ is the total number of blocks, and $n_X = n_1 n_2$ is the total number of pixels in X. In this process, both an interlaced watermark and a secure watermark are embedded, as detailed below.

2.1.1. Interlaced watermarking method (Layer 1)

The central idea of this method is to embed interlaced context-dependent codes, which will be used at the receiver side to refine the localisation bitmap encoded using the secure watermark.

Divide every block X_p into subsets of $n_{\text{iw}} (= m/2)$ pixels each, following the chessboard-like arrangement illustrated in Figure 1. The grey squares in every row of the left figure, as well as the grey squares in every column of the right figure, will form a different subset. Let $X_{p,q}$ denote the q-th subset in X_p, for $q = 1, \ldots, 2m$. The figure shows the arrangement of both horizontal and vertical subsets in a block of 8×8 pixels (i.e. $m = 8$, and $n_{\text{iw}} = 4$), which is the block-size adopted in the experiments reported in Section 4. Observe that the location of the pixels in the horizontal subsets are alternated with the location of the pixels in the vertical subsets.

For the sake of simplicity, in our notation, every subset will be treated as a $1 \times n_{\text{iw}}$ array of pixels. For every subset $X_{p,q}$, compute a n_{iw}-bit code as,

$$h_{p,q} = \mathcal{H}(\hat{X}_{p,q}, p, q, k) \ , \tag{1}$$

where $\mathcal{H}(\cdot)$ is a cryptographic hash function, $k \in \mathbb{N}$ is a secret key and $\hat{X}_{p,q}$ is the block $X_{p,q}$ with its two LSBPs removed – i.e. $\hat{X}_{p,q} = \delta_2 \lfloor X_{p,q}/\delta_2 \rfloor$, where $\delta_i = 2^i$, and $\lfloor x \rfloor$ returns the largest integer not greater than x.

Let $H_{p,q}$ be a $1 \times n_{\text{iw}}$ binary array encoded with the first n_{iw} bits in $h_{p,q}$. Every subset $X_{p,q}$ is then watermarked as,

$$X'_{p,q} = \hat{X}_{p,q} + \delta_1 \, H_{p,q} \tag{2}$$

The watermarked subsets $X'_{p,1}, \ldots, X'_{p,n_{\text{iw}}}$ are assembled together, as illustrated in Figure 1, to form the p-th watermarked block X'_p in the watermarked image X'.

2.1.2. Secure block-wise watermarking method (Layer 2)

The block-wise method, proposed in [12], is adopted to embed a second watermark. This method is secure against VQ attacks and is capable of restoring the original shape, and correct any possible displacements resulting from cropping the watermarked image.

For each block X'_p, an authentication bit string, of length m^2, is encoded as, $w_p = \mathcal{I}_X \,||\, n_1 \,||\, n_2 \,||\, p$, where $\mathcal{I}_X$ is an image index exclusively associated to X, and $||$ denotes concatenation of bits. Note that all the authentication bits share a *common prefix* (i.e. $\mathcal{I}_X \,||\, n_1 \,||\, n_2$), whose length in bits will be denoted by μ.

Then, a bit string, of length m^2, is computed as,

$$h_p = \mathcal{H}(\bar{X}'_p, k) \ \oplus \ w_p \ , \tag{3}$$

where $\bar{X}'_p = \delta_1 \lfloor X'_p/\delta_1 \rfloor$. With the bits in h_p, encode an $m \times m$ binary array, denoted as H_p, which is subsequently embedded to produce the p-th watermarked block by $X^w_p = \hat{X}'_p + H_p$.

2.2. Detection process

Consider an $\tilde{n}_1 \times \tilde{n}_2$ image Y, divided into non-overlapping blocks of $m \times m$ pixels.

2.2.1. Secure block-wise watermarking method (Layer 2)

For each pixel-block Y_p, encode a bit string h'_p with the LSB of the n_{iw} pixels in the block, and extract the authentication bit string as,

$$w'_p = \mathcal{H}(\bar{Y}_p, k') \ \oplus \ h'_p \ , \tag{4}$$

where $\bar{Y}_p = \delta_1 \lfloor Y_p/\delta_1 \rfloor$ and k' is the key provided by the user. Let $\mathbf{A} = \{w'_{a_1}, \ldots, w'_{a_f}\}$ be the set of authentication strings, whose μ left-most bits are identical to each other. If the cardinality of $\mathbf{A}$ will be greater than a predefined threshold τ, the image Y is deemed watermarked. Otherwise, m^2 shifted versions of Y are generated and analysed as described above. In every shifted version, all the pixels in Y are displaced λ_1 rows and λ_2 columns, where $-m < \lambda_1, \lambda_1 \leq 0$. If none of the shifted versions were regarded as watermarked, the detection algorithm is terminated altogether.

If Y was deemed watermarked, retrieve its original dimensions – i.e. n_1 and n_2 – from the common prefix. In case of cropping ($\tilde{n}_1 \neq n_1$ or $\tilde{n}_2 \neq n_2$), the dimensions of Y are restored by adding rows/columns of zeros. Let $\mathcal{B}(w'_p)$ be a function that returns the block index from the $(m^2 - \mu)$ right-most bits in w'_p. Find a set of authentication bit strings $\mathbf{B} = \{w'_{b_1}, \ldots, w'_{b_g}\}$, $\mathbf{B} \subseteq \mathbf{A}$, such that $\mathcal{B}(w'_{b_1}) - b_1 = \ldots = \mathcal{B}(w'_{b_g}) - b_g = \lambda$. The value λ is a *common displacement*, which indicates the number of block slots the content in Y has to be shifted to correct a possible displacement that results from cropping the left or the upper borders of the host image. Finally, a bitmap M is encoded for tampering localisation; note that only the pixel-blocks associated to the bit strings in $\mathbf{B}$ are genuine.

2.2.2. Interlaced watermarking method (Layer 1)

The following mechanism is executed only in pixel-blocks regarded as tampered by the secure block-wise method above.

Without loss of generality, let Y_p be a detected tampered block partitioned into subsets $\{Y_{p,q}|q = 1, \ldots, n_{\text{iw}}\}$, as shown in Section 2.1.1. Let $\breve{h}_{p,q}$ be an n_{iw}-bit code formed with the second LSB of every pixel in the subset. Then, compute another n_{iw}-bit code as,

$$h'_{p,q} = \mathcal{H}(\hat{Y}_{p,q}, p, q, k') \ , \tag{5}$$

where $\hat{Y}_{p,q} = \delta_2 \lfloor Y_{p,q}/\delta_2 \rfloor$. In the corresponding block M_p of the bitmap M, localise both the authentic subsets (where $h'_{p,q} = \breve{h}_{p,q}$) and the tampered subsets (where $h'_{p,q} \neq \breve{h}_{p,q}$). This procedure is illustrated in Fig. 2. Observe that a significant portion of the authentic pixels were verified (light-grey squares), thereby increasing the accuracy of the method.

2.2.3. Post-processing

The procedure detailed below is conducted to enhance the localisation results further.

Let $M(i, j)$ denote the binary pixel at the coordinates (i, j) in the bitmap M, where $M(i, j) = 0$ represents a pixel deemed authentic and $M(i, j) = 1$ represents a pixel deemed altered. Additionally, let $s_{i,j}$ be the sum of the closer neighbours of $M(i, j)$ in the four cardinal directions – i.e. $s_{i,j} = M(i - 1, j) + M(i + 1, j) + M(i, j - 1) + M(i, j + 1)$. The following condition is tested for every $M(i, j) = 1$ to refine the localisation bitmap.

$$M(i,j) = \begin{cases} 1 & \text{if } s_{i,j} \geq 1 \\ 0 & \text{if } s_{i,j} = 0 \end{cases} \ , \tag{6}$$

Finally, the condition below is tested for every $M(i, j) = 0$, to fill some possible gaps in the localisation bitmap whenever a pixel is surrounded by at least five pixels flagged as tampered.

$$M(i,j) = \begin{cases} 1 & \text{if } s_{i,j} \geq 5 \\ 0 & \text{if } s_{i,j} = 0 \end{cases} \ . \tag{7}$$

3. ANALYSIS

3.1. Distortion

To estimate the average embedding distortion, it will be assumed that the every bit in the two LSBPs of a host image will be changed with the same probability. Thus, the mean square error (MSE) is,

$$\text{MSE} = \frac{1}{16} \sum_{i=0}^{3} \sum_{j=0}^{3} (i - j)^2 = \frac{5}{2} \ , \tag{8}$$

So, the approximate average peak signal-to-noise ratio (PSNR) is,

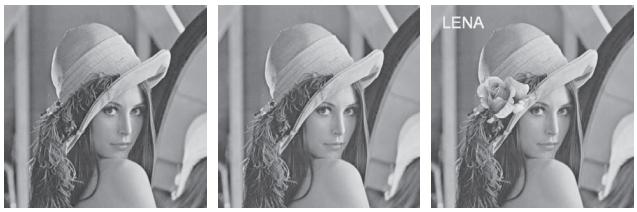
$$\text{PSNR} \approx 10 \log_{10} \left(\frac{2 \ \max_X^2}{5} \right) = 44.2 \text{ dB} \ , \tag{9}$$

where $\max_X$, typically set to 255, is the maximum pixel value in X.

3.2. Security

In following analysis, it is assumed that the output of the hash function $H(\cdot)$ is drawn from a uniform distribution, as in the case of cryptographic hash functions (e.g. the standard SHA).

The likelihood that a non-watermarked image will be deemed watermarked will be determined as follows. Let E_1 be the event that the authentication bit strings retrieved from two different blocks, say w'_u and w'_v, share the same common prefix. According to the well-known *birthday paradox* [13], the probability of E_1 is given by,



(a) **(b)** **(c)**

Fig. 3: Example tampering. (a) Original 512×512 image. (b) Watermarked image. (c) Tampered image.

$$\mathcal{P}_{E_1}(2^\mu, m_{\text{bw}}) = 1 - \frac{2^\mu}{(2^\mu - m_{\text{bw}})! \ 2^{\mu \ m_{\text{bw}}}} \ . \tag{10}$$

Now, let E_2 be the event that the prefix of another block matches the one retrieved from the two blocks above. The probability that E_2 will occur is $\mathcal{P}_{E_2} = 2^{-\mu}$. Let $\mathcal{X}_1$ be a random variable that indicates the number of occurrences of the event E_2. Since an image will be deemed watermarked only if E_2 occurs, at least, τ times, this can be modelled as the remainder of a cumulative binomial distribution [13] given by,

$$\mathcal{P}_{\mathcal{X}_1}(\mathcal{X}_1 \geq \tau) = 1 - \sum_{i=0}^{\tau-1} \binom{m_{\text{bw}}}{i} \mathcal{P}_{E_2}^i (1 - \mathcal{P}_{E_2})^{m_{\text{bw}} - i} \ , \tag{11}$$

where $\binom{m_{\text{bw}}}{i}$ is the binomial coefficient.

Observe that once the block has been deemed authentic by the block-wise detector, there is no way one of its subsets can be deemed otherwise. That is, even in the worst case scenario, the interlaced watermarking mechanism will not increase the false positive rate. This comes to the expense of increasing the chances that an altered pixel will be mistakenly deemed valid to one in every $2^{n_{\text{iw}}}$.

4. RESULTS

The following are the parameter settings employed for the proposed TL method: $m = 8$, $n_{\text{wi}} = 4$, and τ was set to 1% of the total number of blocks. For comparison purposes, the localisation performance of the proposed TL method has been compared with the following existing methods: Fridrich [5], Lin *et al.* [6], He *et al.* [9], Li and Si [8], and He *et al.* [11].

Two standard metrics were employed to compare the localisation performance of the six methods: the Accuracy (ACC = M ∩ G / G) and the False Positive Rate (FPR = (M ∪ G - G) / G), where M is the encoded localisation bitmap and G is the ground truth. Note that an ideal detection would result in ACC=1 and FPR=0.

4.1. Example tampering

The Lena image, in Figure 3(a), was the test image. The distortion induced by the TL method was estimated by means of the peak signal-to-noise ratio (PSNR) between the watermarked and the original image, which was assessed to be 44.2 dB. The PSNR values obtained from the images watermarked with the other methods were: Fridrich 51.1 dB, Lin *et al.* 44.2 dB, He *et al.* [9] 51.1 dB, Li and Si 54.7 dB, and He *et al.* [11] 39.5 dB. Figure 3(c) shows the doctored image. Results show that the localisation performance achieved

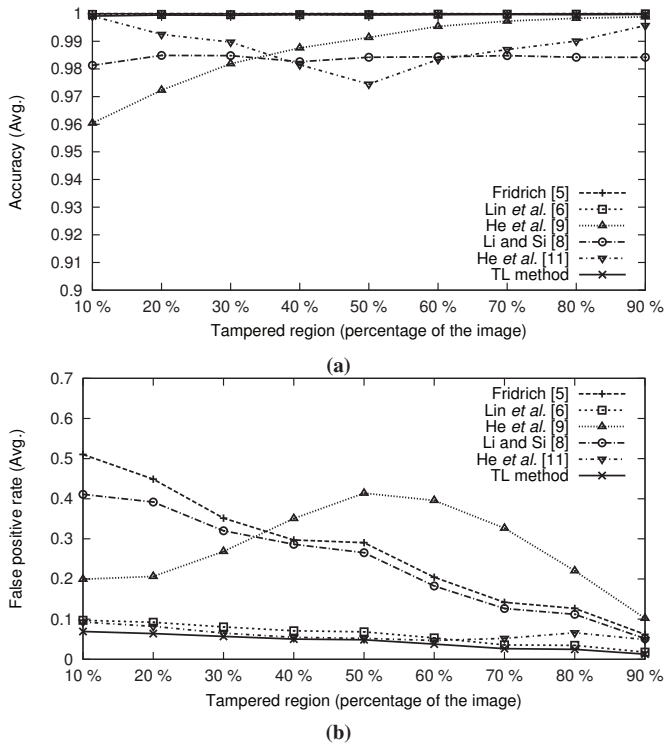


Fig. 4: Localisation vs size of tampered region. (a) Accuracy. (b) False positive rate.

by the TL method (ACC=1 and FPR=0.12) compares favourably to the rest of the methods; Fridrich, ACC=1 and FPR=0.6; Lin *et al.*, ACC=1 and FPR= 0.2; He *et al.* [9], ACC=0.95 and FPR=0.04; Li and Si, ACC=0.98 and FPR=0.46; He *et al.* [11], ACC=1 and FPR=0.2.

4.2. Localisation performance in varying tampering areas

The watermarked versions of the 220 images of size 800×600, in the Caltech-256 data set [14], were subjected to distortions in areas covering from 10% to 100% (with increments of 10%) of the number of pixels in the image. The averaged results were compared as follows. Figure 4(a) shows that all the methods correctly identified over 95% of the tampered pixels. In fact, in all the cases, the accuracy was assessed to be 1 when using Fridrich's, He *et al.*'s [11], Lin *et al.'s* or the TL method. Figure 4(b) shows that, only when the tampered region covered 10% of the image, He *et al.*'s [9] method managed to validate a larger proportion of unaltered pixels than the rest of the schemes. For the remainder of the experiments, however, the TL method compared favourably to the rest of the schemes in terms of FPR. Observe that the TL method even outperforms He *et al.*'s [11] scheme, which induces a higher distortion into the host images (around 39.5 dB).

5. CONCLUSIONS

The proposed TL method hierarchically structures two mechanisms to provide higher localisation capabilities. The security of the scheme is provided by a secure block-wise method, whilst an interlaced mechanism is employed to enhance the localisation achieved by the block-wise method. Results show that the TL method outperforms five state-of-the-art schemes in terms of localisation. Furthermore, this is achieved without inducing an excessive amount of distortion into host images as in other methods found in recent literature. The localisation of scattered altered pixels (e.g. salt-and-pepper noise) poses challenges to the TL method. Further work has been planned to cope with this shortcoming.

6. REFERENCES

[1] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, *Digital watermarking and steganography*, Morgan Kauffman, 2nd edition, 2008.

[2] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[3] P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. of ICIP – IEEE International Conference on Image Processing*, Chicago, IL, USA, 1998, vol. 1, pp. 455 – 459.

[4] M. Holliman and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 432 – 441, 2000.

[5] J. Fridrich, "Security of fragile authentication watermarks with localization," in *Proc. of Security and Watermarking of Multimedia Contents*, CA, USA, 2002, vol. 4675, pp. 691 – 700, SPIE.

[6] P.-L. Lin, C.-K. Hsie, and P.-W. Huang, "A hierarchical digital watermarking method for image tamper detection and recovery," *Pattern Recognition*, vol. 38, no. 12, pp. 2519 – 2529, 2005.

[7] C.-C. Chang, Y.-H. Fan, and W.-L. Tai, "Four-scanning attack on hierarchical digital watermarking method for image tamper detection and recovery," *Pattern Recognition*, vol. 41, no. 2, pp. 654–661, 2008.

[8] C.-T. Li and H. Si, "Wavelet-based fragile watermarking scheme for image authentication," *Journal of Electronic Imaging*, vol. 16, no. 1, pp. 1–9, 2007.

[9] H. He, J. Zhang, and H-M Tai, "A wavelet-based fragile watermarking scheme for secure image authentication," in *Proc. of IWDW – International Workshop on Digital Watermarking*, 2006, vol. 4283, pp. 422 – 432.

[10] X. Zhang and S. Wang, "Statistical fragile watermarking capable of locating individual tampered pixels," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 727–730, 2007.

[11] H.-J. He, J.-S. Zhang, and H.-M. Tai, "Block-chain based fragile watermarking scheme with superior localization," in *Proc. of IH – Information Hiding*, 2008, pp. 147–160.

[12] S. Bravo-Solorio and A. K. Nandi, "Secure fragile watermarking method for image authentication with improved tampering localisation and self-recovery capabilities," *Signal Processing*, vol. 91, no. 4, pp. 728–739, 2011.

[13] Morris H. Degroot and Mark J. Schervish, *Probability and Statistics*, Addison Wesley, 1st edition, 1975.

[14] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep. 7694, California Institute of Technology, 2007.