Fuzzy-ART Based Adaptive Digital Watermarking Scheme

Chip-Hong Chang, *Senior Member, IEEE*, Zhi Ye, and Mingyan Zhang

***Abstract*—In this paper, a novel transform domain digital watermarking scheme that uses visually meaningful binary image as watermark has been developed. The method embeds the watermark information adaptively with localized embedding strength according to the noise sensitivity level of the host image. Fuzzy adaptive resonance theory (Fuzzy-ART) classification is used to identify appropriate locations for watermark insertion and its control parameters add agility to the clustering results to thwart counterfeiting attacks. The scalability of visually recognizable watermark is exploited to devise a robust weighted recovery method with composite watermark. The proposed watermarking schemes can also be employed for oblivious detection. Unlike most oblivious watermarking schemes, our methods allow the use of visually meaningful image as watermark. For automation friendly verification, a normalized correlation metric that suits well with the statistical property of our methods is used. The experimental results show that the proposed techniques can survive several kinds of image processing attacks and the JPEG lossy compression.**

***Index Terms*—Digital watermarking, Fuzzy adaptive resonance theory (Fuzzy-ART), data concealment.**

I. Introduction

INVISIBLE digital watermarking is a process of permanently embedding multimedia content into digital signal carrying information about the ownership and identification of the intellectual property so that the existence of the watermark is virtually unperceivable by human sensory system. Similar to any intellectual property on paper, information published or distributed on a networked environment needs to be safeguarded against piracy and malicious manipulation. Although encryption is possible to provide secured delivery of valuable information by deterring counterfeiters from hijacking the copyrighted information, it fails to control the distribution of the illegal copies of the original work upon decryption by the authorized recipients. The use of invisible digital watermarking schemes to certify the image legitimacy and enable the tracking of the distribution of its permitted copies is a viable solution to this problem.

There are four main types of digital watermarking schemes for digital still images: copyright protection, traitor tracking, copy protection, and image authentication [4], [5], [8], [14], [17], [19], [21]. Despite the wide variant of usages and applications, a good watermark should possess some common characteristics such as imperceptibility, robustness, security, reliability and low computation cost.

Most watermark embedding processes are performed in either spatial domain or transform domain. In spatial domain watermarking schemes [2], [8], [19], the watermark image is directly embedded into the host image by changing its pixel values. Both the insertion and extraction processes are relatively simple compared to transform domain watermarking schemes. However, it is more difficult for spatial domain watermarks to achieve imperceptibility because of the need to embed high intensity digital watermark for robustness tends to degrade the image visual quality.

In transform domain watermarking schemes [1], [4], [5], [7], [9], [12], [13], [16], [21], transform domain coefficients of the host image are modulated by the watermark information. They have several advantages over spatial domain schemes. First, it is more difficult for attackers to extract the marked information and hence to alter the watermarks since the watermark is irregularly distributed all over the host images. Second, one can select certain bands that possess perceptually significant features to embed the watermark. Third, it is the transform domain coefficients that are modified rather than the pixel values of the host image, making it plausible to reduce the visual artifacts of the marked image even though the watermark is introduced into the selected coefficients that contribute significantly to the host image intelligibility. However, frequency-based watermarking schemes are generally susceptible to geometrical transformation attacks. This lost of synchronization can be efficiently detected and the transformation parameters can be recovered using an elegant method proposed by Izquierdo [13] lately for spread-spectrum like transform domain watermarking schemes. As the host images are often divided into smaller equal size blocks in transform domain watermarking schemes, they can also be vulnerable to counterfeiting attack [11]. Wong and Memon [20] use the security of a proven cryptographic hash function for a block-wise independent digital watermarking scheme to work against the counterfeiting attack. Unfortunately, their watermarking scheme belongs to the class of a fragile digital watermarking scheme which is not designed to be resilient against even a mild image processing attack. One main weakness of their watermarking scheme is that the risk of rejecting an authentic watermarked image is high even though it has not been maliciously manipulated as noises in the communication channel are likely to corrupt more than the least significant bits of the pixels in transmitting the marked image.

Manuscript received December 4, 2002; revised July 7, 2003. This paper was recommended by Associate Editor E. Izquierdo.

C. H. Chang and Z. Ye are with the Center for High Performance Embedded Systems, Nanyang Technological University, 639798 Singapore (e-mail: echchang@ntu.edu.sg).

M. Zhang is with Tech Semiconductor Singapore Pte. Ltd., 738799 Singapore (e-mail: mingyan@techsemicon.com.sg).

Digital Object Identifier 10.1109/TCSVT.2004.839992

Zhao and Koch [21] developed a watermarking scheme based on discrete cosine transform (DCT). In their approach, images are segmented into 8×8 blocks, and each block is transformed into the DCT domain. Out of the 64 DCT coefficients, three from a predefined set of eight coefficients covering the low to medium frequency bands of the image block are selected. Since standard image compression techniques normally truncate the higher frequency components by applying the quantization matrix, leaving the lower frequency components relatively unchanged; altering only the low and medium frequency components makes the watermark more resilient to attacks. However, as only 3/64 of the coefficients are used to embed the watermark, the amount of watermark information that can be embedded is limited. In most previous works on watermarking [7], [13], [17], [19], the watermark is usually a random sequence of bits or binary codewords. A visually meaningful watermark, like a company logo, or a copyright date stamp is more intuitive as an identifier and can convey more information than an uncorrelated sequence of binary numbers. Hsu and Wu [12] proposed another block-wise DCT domain watermarking scheme where a visually recognizable binary pattern is used as watermark. The watermark is permutated and shuffled into every image block according to the variances of the image block. Recently, Luo *et al.* [15] also proposed a watermarking scheme that applies to JPEG images and uses visually recognizable patterns as watermark. Their scheme suffers from a very low embedding capacity. Only a single bit is embedded in every four 8×8 bit macroblocks. We envisage that the amount of information that can be embedded is very image dependent and a good balance between imperceptibility and robustness is hard to strive by compelling to embed watermark information globally. In our proposed watermarking scheme, the DCT coefficients in selected regions of the host image are adaptively modulated by the watermark information with the help of Fuzzy adaptive resonance theory (Fuzzy-ART) [3], [5], [10], [18]. Thus, the strength of the watermark can be increased to enhance the robustness without introducing noticeable degradation to the visual quality of the host image. Besides, our method can accommodate more DCT coefficients covering the low to medium frequency bands than Zhao's [21] and Luo's [15] methods to embed the watermark.

Fuzzy-ART proposed by S. Grossberg is one of the more recent and popular members in the family of neural networks. According to Carpenter *et al.*[3], Fuzzy-ART is "capable of rapid stable learning of recognition categories in response to arbitrary sequences of *analog* or *binary* input patterns." In statistical terminology, "recognition categories" are clusters and "arbitrary sequences of input patterns" are data. Hence, Fuzzy-ART is a method for clustering data. Fig. 1 shows the architecture of the Fuzzy-ART network [10]. It consists of two layers of computing cells or neurons, and a vigilance subsystem controlled by an adjustable vigilance parameter $\rho \in [0, 1]$. The input vectors are applied to the Fuzzy-ART network one by one. The network seeks for the "nearest" cluster that "resonates" with the input pattern according to a "winner-take-all" strategy and updates the cluster to become "closer" to the input vector [3], [10], [18]. In the process, the vigilance parameter determines the similarity of the inputs belonging to the same cluster. For the same set of inputs, the similarity of elements in one cluster grows as the vigilance parameter increases, leading to a larger number of trained clusters. The choice parameter and the learning rate are two other factors that influence the quality of the clustering results. Our proposed watermarking scheme uses Fuzzy-ART to cluster the host image. Only those images blocks akin to the selected clusters are used to embed the watermark information. We found from experiments that the blocks having fewer large-magnitude selected coefficients require stronger embedding strength to survive quantization. However, the embedding strength is limited by the ability to sustain imperceptibility. Judging from the nature of the selected clusters, every cluster is associated with a distinctive embedding strength: clusters that are smooth, i.e., having more low frequency components are given stronger embedding strength and vice versa. In our approach, Fuzzy-ART trains each host image independently and identifies good locations for watermark insertion according to the user specified parameters. Thus, the clustering result would be very different from image to image, which leads to a diversity of embedding strengths. Therefore, the attackers cannot counterfeit watermark by simply replacing a similar block from a marked image [11]. A formal analysis of its resistance against the counterfeiting attack is provided toward the end of the paper.

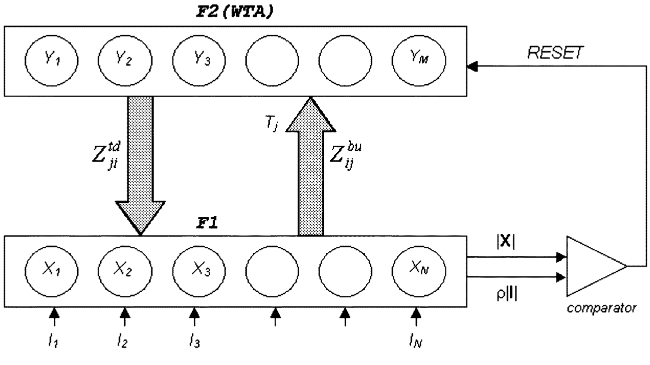


Fig. 1. Topological structure of the Fuzzy-ART architecture.

The paper is organized as follows. In Section II, the method of adaptively embedding the watermark into the host image is described. The use of a polarity mask for enhancing the detectability of the watermark object pixels and its statistical basis are explained. Section III describes the watermark extraction method and introduces a novel robustness enhancement recovery method by superimposing several weighted copies of reduced size watermark. Another important contribution, which is the oblivious detection of visually meaningful watermark, is also described in this section. The mathematical derivation of the principle behind our proposed oblivious detection method is also provided. Several experiments have been performed in Section IV to demonstrate the robustness of our proposed watermarking schemes against common image processing attacks. The results are analyzed and compared with recent published results of several DCT-based watermarking schemes with similar features. A formal analysis of the robustness of our proposed schemes against counterfeiting attacks targeted on block-wise transform domain watermarking schemes is given. The paper is concluded in Section V.

II. Adaptive Watermark Embedding Scheme

The proposed adaptive digital watermarking scheme operates in the DCT domain. For ease of exposition, the method is explained with the gray-scale images. Greater embedding agility can be achieved with color images by exploiting the additional dimensions of the pixel values. The skeleton of the proposed watermarking scheme resembles the basic algorithms of Hsu and Wu [12] with several distinct differences, in which the use of the Fuzzy-ART is the most important one.

A. Preprocessing by Fuzzy-ART Clustering

The host gray-scale image of size $M \times N$, is preprocessed to prepare for Fuzzy-ART classification. First, the host image is divided into nonoverlapping blocks of size 8 × 8

$$I = I_1 \| I_2 \| \cdots \| I_p \tag{1}$$

where $I_i, 1 \le i \le p$, is a macroblock of 8 × 8 pixels. If p is not a whole number, extra pixels are truncated, i.e., $p = \lfloor M/8 \rfloor \times \lfloor N/8 \rfloor$.

Each block of the host image is then DCT transformed independently

$$\begin{aligned} J &= \mathrm{DCT}(I) \\ &= \mathrm{DCT}(I_1) \| \mathrm{DCT}(I_2) \| \cdots \| \mathrm{DCT}(I_p) \\ &= J_1 \| J_2 \| \cdots \| J_p \end{aligned} \tag{2}$$

where DCT denotes the block-wise DCT. J_i is the DCT of $I_i, 1 \le i \le p$. It is also a macroblock consisting of 8 × 8 DCT coefficients.

Each block J_i is converted to a 1 × 64 array, following a raster scanning order. All arrays are normalized by dividing each input vector by its norm. An input vector J is said to be normalized when there exists a constant $\gamma > 0$ such that $|J| = \gamma$ for all input vectors [10]. Complement-coding rule [3] is applied for the normalization of the input vectors. Each array J_i is expanded to array J_{ai} of dimension 1 × 128 with the element values lie between 0 and 1 to form the input training patterns for the Fuzzy-ART operation.

$$J_a = J_{a1} \| J_{a2} \| \cdots \| J_{ap} \tag{3}$$

$$J_{ai} = (a_i, a_i^c) = (a_{i1}, a_{i2}, \cdots, a_{i64}, a_{i1}^c, a_{i2}^c, \ldots, a_{i64}^c) \tag{4}$$

where $1 \le i \le p$ and $a_{ik}^c = 1 - a_{ik}$, for $k = 1, 2, \ldots, 64$.

Finally, all input arrays are applied to the Fuzzy-ART network for classification. The Fuzzy-ART function produces two outputs, a weight matrix (WM), and a codebook (CB)

$$(\mathrm{WM}, \mathrm{CB}) = \text{Fuzzy-ART}(J_a, \alpha_F, \beta_F, \rho_F) \tag{5}$$

where α_F is the choice parameter, β_F is the learning rate, and ρ_F is the vigilance parameter.

Each row of the WM contains the detailed information about one cluster and the CB collects the cluster index of each block. The number of rows of the WM represents the number of clusters generated by the Fuzzy-ART network. Fig. 2 shows the flow diagram of the Fuzzy-ART operation. Interested reader can refer to [3], [10], [18] for more details.

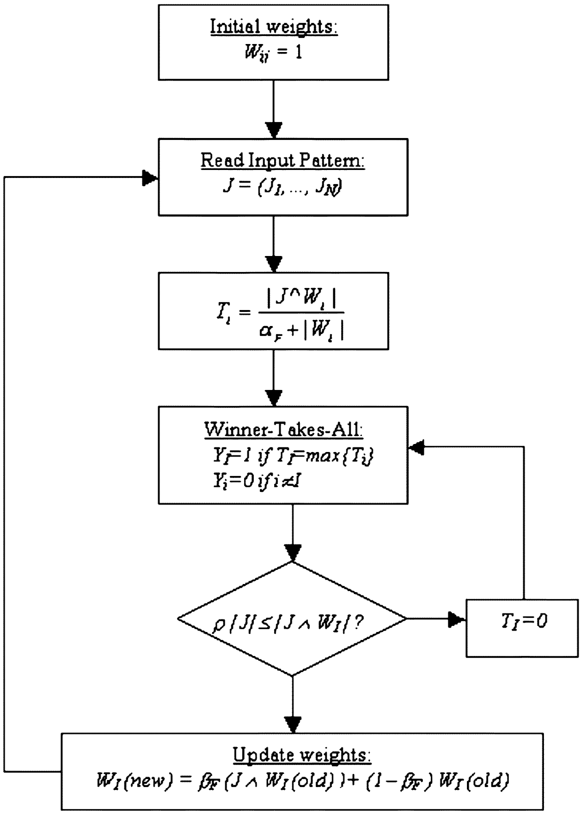

Initial weights:
$W_{ij} = 1$
Read Input Pattern:
$J = (J_1, \ldots, J_N)$
$T_i = \frac{|J \wedge W_i|}{\alpha_F + |W_i|}$
Winner-Takes-All:
$Y_I = 1$ if $T_I = \max\{T_i\}$
$Y_i = 0$ if $i \ne I$
$\rho|J| \le |J \wedge W_I|$?
$T_I = 0$
Update weights:
$W_I(new) = \beta_F (J \wedge W_I(old)) + (1 - \beta_F) W_I(old)$

Fig. 2. Flow diagram of the Fuzzy-ART operation.

B. Watermark Encryption

The watermark we used is a visually meaningful binary image of size $m \times n$. To circumvent the explicit block-wise spatial correspondence between the host and watermark images, a stream cipher is used to diffuse the watermark across the block boundaries over the entire image. In our implementation, we use a cryptographically strong randomizer, software-optimized encryption algorithm (SEAL) [6], to achieve the information diffusion. This dispersion of the spatial relationship of the watermark enhances the watermark robustness against image cropping attack. Because of the properties of the cryptographic function, the watermark bitmap resembles random noise with balanced number of ones and zeros. Thus, the desirable statistical characteristics are sustained for its reliable detection under a spread-spectrum-like embedding scheme

$$\begin{aligned} W_e &= \mathrm{Encrypt}(W, K_s) \\ &= \{w_e(i,j) \in W_e | w_e(i,j) \\ &= w(i,j) \oplus b_{Ks}(in + j) \quad \forall 0 \le i < m, 0 \le j < n\} \end{aligned} \tag{6}$$

where $w(i,j) \in \{0,1\}$ is the binary value of the watermark pixel at coordinate (i,j). $b_{Ks}(r) \in \{0,1\}$ is the rth bit generated by a stream cipher with the secret key, K_s.

The encrypted watermark, W_e is also divided into q nonoverlapping macroblocks of size $m_w \times n_w$ so that each macroblock can be embedded into a distinct host image block.

$$W_e = W_{e1} \| W_{e2} \| \cdots \| W_{eq} \tag{7}$$

where $W_{ei}, 1 \le i \le q$, $q = mn/m_w \times n_w$ is a macroblock of $m_w \times n_w$ binary pixels.

C. Adaptive Embedding

In general, the embedding process can be modeled mathematically as

$$J^* = \varepsilon_{Ke}(I, W) \tag{8}$$

where I is original image, W is the watermark information, J^* is the watermarked image, ε is the insertion function, and K_e is the encryption key.

Two selections are essential for making our proposed watermarking scheme robust against counterfeiting attacks. First, a subset of clusters is selected to make up sufficient number of host image blocks for watermark insertion according to the results of Fuzzy-ART training. Altogether q out of p blocks are selected and each selected block is denoted as $S_i, 1 \leq i \leq q$. Second, out of the 64 DCT coefficients of each selected host image block, some are selected to embed the watermark pixels. In fact, the candidate clusters are chosen based on the number of zero valued coefficients left after quantization. The priority of the selection is in the order of increasing number of zero valued coefficients.

To invisibly embed the watermark (i.e., embed into the higher frequency components) and to survive the lossy data compression (i.e., embed into the low-frequency components) is contradictory. Therefore, a reasonable trade-off is to embed the watermark in the middle-frequency band of the image [12]. In our proposed watermarking scheme, we only embed the watermark information into the coefficients selected from the low-median, median, and median-high frequency bands according to the cluster features. $64\gamma(mn/MN)$ DCT coefficients are selected from each candidate block S_i where γ is the fraction of selected blocks. The selected coefficients form a reduced macroblock R_i, which is of the same size as that of encrypted watermark block.

$$\begin{aligned} R &= \text{Reduce}(S) \\ &= \text{Reduce}(S_1)||\text{Reduce }(S_2)||\cdots||\text{Reduce}(S_q) \\ &= R_1||R_2||\cdots||R_q \end{aligned} \tag{9}$$

With the encrypted digital watermark W_e and the reduced image R, which contains only the selected frequency components of the original image, we are ready to insert the watermark. The information of each encrypted watermark block will be cast into the reduced image block at the corresponding spatial location. Instead of modifying the selected coefficients with a uniform gain factor globally, a localized gain factor adaptive to the intra block DCT coefficients sensitivity is used. This image dependent embedding will strengthen its security against counterfeiting [11] and increase its robustness to DCT coefficient difference in scale. Equation (8) can be rewritten to describe our embedding scheme as follows:

$$\begin{aligned} R^* &= \varepsilon_{Ke}(J, W_e) \\ &= \varepsilon_{Ke1}(R_1, W_{e1})||\varepsilon_{Ke2}(R_2, W_{e2})||\cdots \\ &\quad ||\varepsilon_{Keq} \quad (R_q, W_{eq}) \\ &= R_1^*||R_2^*||\cdots||R_q^* \end{aligned} \tag{10}$$

$$\begin{aligned} R_i^*(k,l) &= R_i(k,l) \pm \alpha(i)|R_i(k,l)| \\ &\qquad \forall 1 \leq i \leq q, 0 \leq k < m_w, 0 \leq l < n_w \end{aligned} \tag{11}$$

where $R_i(k,l)$ is the DCT coefficient value at coordinate (k,l) of the reduced macroblock R_i. $\alpha(i)$ is the gain factor controlling the embedding strength of the watermark in the reduced block R_i. The value of $\alpha(i)$ is determined by the intensities of the high and low frequency components in that block. A block having many large magnitude selected coefficients can afford a lower embedding strength and a block with fewer large magnitude coefficients requires stronger embedding strength to survive quantization but the embedding strength is limited by the ability to sustain imperceptibility. The watermark information modulates the values of the selected DCT coefficients by changing the sign of the second term in (11).

The statistical characteristics of (11) have been analyzed by [1] where the direction of scaling [± sign of (11)] of the DCT coefficient is determined by a pseudo-random binary bit sequence (the watermark) of length L. For brevity and without loss of accuracy, (11) can be rewritten as

$$t_i^* = t_i + \alpha|t_i|x_i \tag{12}$$

where t_i^* and t_i are the marked and unmarked DCT coefficients, respectively. $x_i \in \{1, -1\}$ is the polarity bit controlled by the encrypted watermark pixel, with $i = 1, 2, \ldots, L$.

The normalized correlation, z is proposed and studied in [1]

$$z = \frac{1}{L}\sum_{i=1}^{L}(t_i y_i + \alpha|t_i|x_i y_i) \tag{13}$$

where y_i is the ith bit of a generic test watermark generated by a pseudo-random binary sequence.

Under the hypothesis that both t_is and x_is are zero mean, independent, and equally distributed random variables, the following mean and variance of z have been theoretically computed in [1]

$$\sigma_z^2 = \begin{cases} \frac{1+2\alpha^2}{L}\sigma_t^2 + \frac{\alpha^2}{L}\sigma_{|t|}^2, & \text{if } X = Y \\ \frac{1+2\alpha^2}{L}\sigma_t^2, & \text{if } X \neq Y \\ \frac{1}{L}\sigma_t^2, & \text{if no watermark} \end{cases} \tag{14}$$

$$\mu_z = \begin{cases} \alpha\mu_{|t|}, & \text{if } X = Y \\ 0, & \text{if } X \neq Y \\ 0, & \text{if no watermark} \end{cases} \tag{15}$$

where X and Y are the actual and test watermarks, respectively.

It was found that μ_z does not depend on the sequence length L, and that it increases with the gain factor α. To achieve a low-error detection probability for watermark verification based on z, the distance, $k = \mu_z/\sigma_z$ between the Gaussian curves of z for the cases of matched and unmatched watermark, must be sufficiently large.

In our watermarking scheme, the watermark used is a visually recognizable pattern. This deterministic watermark is cryptographically randomized to approximate a zero mean Gaussian distribution. The findings of [1] have been exploited here to increase the factor k by raising the mean μ_z and lowering the standard deviation σ_z of the marked images at where the essential watermark information (the foreground object with pixel values "0") is embedded. This is accomplished by using a polarity mask to scale the coefficients to be embedded with the watermark object pixels in the direction toward the block mean value. The enhancement of the object pixels is accompanied by a spreading of the background pixels in the recovery process. Making use of the fact that the binary object and background pixels are mutually exclusive, the normalized correlation metric is applied only to the more reliably detected object pixels for watermark detection.

A two-dimensional (2-D) pattern mask is used to compute the polarity of each pixel within each reduced image block. That is,

$$\begin{aligned} P &= \text{Polarity}(R) \\ &= \text{Polarity}(R_1)\|\text{Polarity}(R_2)\|\cdots\|\text{Polarity}(R_q) \\ &= P_1\|P_2\|\cdots\|P_q \end{aligned} \tag{16}$$

$$P_i(k,l) = \begin{cases} 1, & \text{if } R_i(k,l) > \beta_i \\ 0, & \text{otherwise} \end{cases} \quad 1 \le i \le q, 0 \le k < m_w, 0 \le l < n_w \tag{17}$$

where the average value of the selected frequency coefficients, β_i is given by

$$\beta_i = \frac{\sum_{k=0}^{m_w-1}\sum_{l=0}^{n_w-1} R_i(k,l)}{m_w \times n_w}. \tag{18}$$

With the help of this polarity mask P, the watermark pixels are embedded into the reduced DCT block R according to (10) and (11). The ± sign of (11) is governed by the polarity mask (P) and the encrypted watermark image (W_e). It takes the plus sign when the corresponding bit in P XOR W_e is "1" and the minus sign otherwise. In other words, the polarity is enhanced if the embedding pixel value is 1, and the polarity is reversed if the embedding pixel value is 0. The use of polarity is to ensure that the DCT coefficients are scaled in the direction toward the block mean value in order to achieve a low-error detection probability [1].

The modified frequency coefficients R^* is re-mapped into J to obtain J^*.

$$J^* = R^*\|J_u = R_1^*\|R_2^*\|\cdots\|R_q^*\|J_u = J_1^*\|J_2^*\|\cdots\|J_p^* \tag{19}$$

where J_u denotes those DCT coefficients that has not been selected for modulation.

Finally, an inverse block DCT (IDCT) is carried out on the resulting blocks to obtain the watermarked image

$$\begin{aligned} I^* &= \text{IDCT}(J^*) \\ &= \text{IDCT}(J_1^*)\|\text{IDCT}(J_2^*)\|\cdots\|\text{IDCT}(J_p^*) \\ &= I_1^*\|I_2^*\|\cdots\|I_p^*. \end{aligned} \tag{20}$$

The overall adaptive watermark embedding process is illustrated in Fig. 3.

III. Watermark Extraction

A. Single Watermark Recovery

The extraction of watermark requires the original image I, the possibly corrupted watermarked image I', the watermark W, the key used in the watermark encryption, and the location map of the selected blocks and the indexes used for the mapping of J_i to R_i. Similar to the embedding function, the extraction function can be represented as

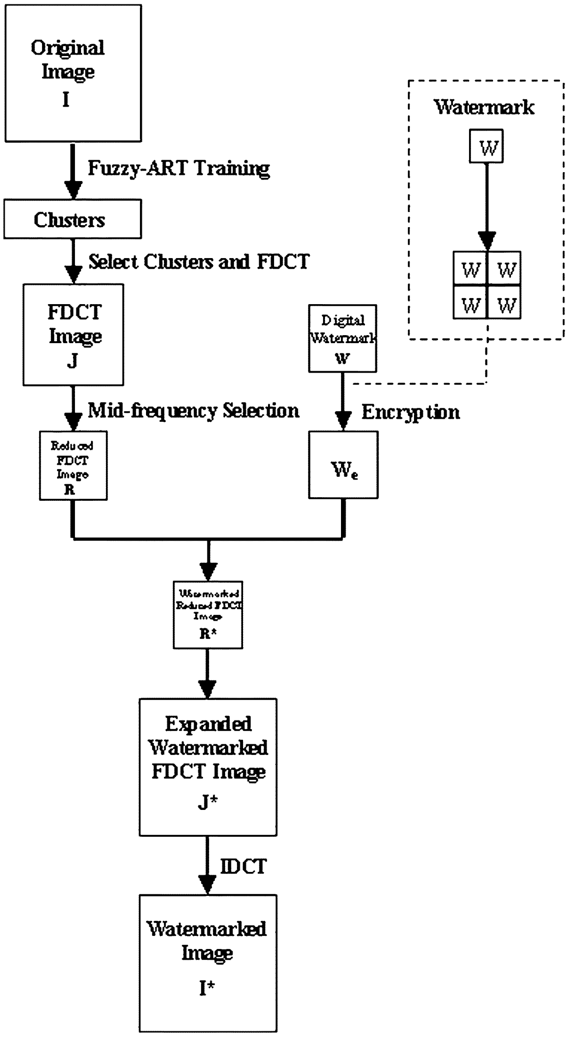


Fig. 3. Fuzzy-ART adaptive watermark embedding process.

$$W' = D_{K_d}(I, I', W) \tag{21}$$

where D is the watermark extraction function, K_d is the decryption key, and W' is the recovered watermark.

In the watermark recovery process, both the original image I and the image in question I' are block DCT transformed independently.

$$\begin{aligned} J &= \text{DCT}(I) \\ &= \text{DCT}(I_1)\|\text{DCT}(I_2)\|\cdots\|\text{DCT}(I_p) \\ &= J_1\|J_2\|\cdots\|J_p \end{aligned} \tag{22}$$

$$\begin{aligned} J' &= \text{DCT}(I') = \text{DCT}(I_1')\|\text{DCT}(I_2')\|\cdots\|\text{DCT}(I_p') \\ &= J_1'\|J_2'\|\cdots\|J_p'. \end{aligned} \tag{23}$$

According to the location map of the selected blocks, the candidate blocks, S and S' are extracted and from which the reduced macroblock R and R' are formed.

$$R = \text{Reduce}(S) \tag{24}$$

$$R' = \text{Reduce}(S'). \tag{25}$$

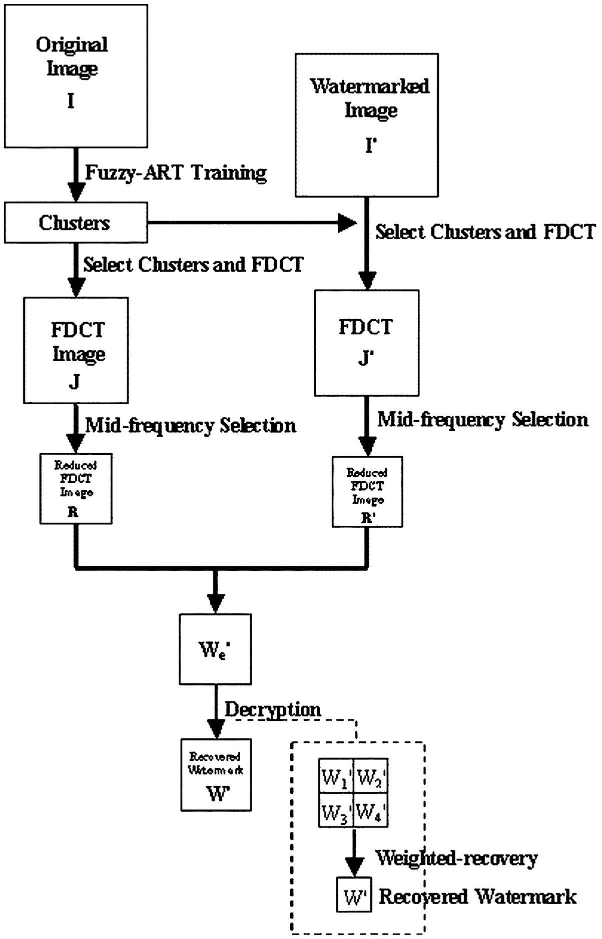


Fig. 4. Watermark retrieval process.

The polarity pattern mask P' is generated by subtracting the original reduced image R from the possibly corrupted watermarked reduced image R'. The possibly corrupted encrypted watermark can be extracted by simply logically XORing P' and the polarity mask P generated by the original image, i.e.,

$$P_i'(k,l) = \begin{cases} 1, & \text{if } R_i'(k,l) - R_i(k,l) > 0 \\ 0, & \text{otherwise} \end{cases}$$
$$1 \le i \le q, 0 \le k < m_w, 0 \le l < n_w \quad (26)$$
$$W_e' = \text{XOR}(P, P'). \quad (27)$$

Finally, the secret key K_s is used to decipher W_e' to recover the watermark W'

$$W' = \text{Decrypt}(W_e', K_s). \quad (28)$$

Fig. 4 shows the watermark retrieval process.

B. Weighted Recovery of Multiple Watermarks

One advantage of using visually recognizable binary watermark is the scalability. By reducing the resolution of the watermark, the perceptibility can be improved as fewer coefficients are altered but the robustness may not necessarily be enhanced. Depending on the coefficients chosen for embedding the fewer watermark pixels, certain attacks like cropping might thin the watermark further. If the watermark can be scaled down, we suggest the use of multiple identical watermarks with selective weightage of each copy in the recovery process to enhance the robustness of the watermark. The watermark is reduced and duplicated t times to form a composite watermark of the original size. This composite watermark is encrypted and embedded into the host image following the similar steps as presented in Section 2. As shown in Fig. 3, the watermark can be replaced by a composite watermark (enclosed in the dotted rectangular box) constructed from four identical quarter-sized watermarks.

In the retrieval process, multiple duplicates of the watermark are obtained after the decryption. These t retrieved watermark copies, $W_1', W_2', \ldots, W_t'$, where W_i' are $\{1, -1\}$ encoded from $\{1, 0\}$, are weighted according to their embedding strength or gain factor α and superimposed to form a single reduced resolution watermark. The recovered watermark is expressed in (29).

$$W'(m_wk+i, n_wl+j) = \begin{cases} 1, & \text{if} \frac{\sum_{s=1}^{t} W_s'(m_wk+i, n_wl+j)/\alpha_k(k,l)}{t} > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (29)$$

For ease of reference in comparing the experimental results, we abbreviate the Fuzzy-ART adaptive watermarking schemes with single and composite watermarks as FAS and FAM, respectively.

C. Oblivious Detection

The disadvantage of the scheme presented in the earlier sections is the requirement of the original image during the watermark retrieval process. Access to the original image must be granted to anyone who needs to verify the existence of the watermark. The presence of the original image allows some form of preprocessing like reorientation and patching of cropped parts by comparing it with the possibly corrupted marked images, thus lowering the probability of erroneous detection. However, in some cases, methods that do not require the original image are preferred. For instance, in broadcast monitoring and electronic commerce applications where the transmission or downloading of a piece of commercial may be authenticated by the automatic extraction of watermark at a remote terminal. A printer may grant the permission to print an image only if the correct watermark is detected. Under those scenarios, the demand of a database of large size to securely store all the original images may not be practical [9], [14], [16].

We have exploited the advantage of visually meaningful watermark in weighted recovery. However, the use of visually recognizable pattern has also known to be obscure to oblivious detection. In this section, we propose an oblivious detection method that does not require the original image to be transmitted or stored in the database for the retrieval. Thanks to the Fuzzy-ART clustering, the amount of information needed in the retrieval process has been reduced significantly. Instead of the original image, smaller amount of data, which are generated from the original image, are transmitted to the authenticator. Unlike the oblivious methods of [9] and [16], a visually recognizable watermark can still be used.

The watermark embedding process is similar to the method described in Section II-C up to the point where the reduced image of (9) is formed. However, the mask generation process is different from the nonoblivious case. This time, the mask is

TABLE I
POLARITIES OF MARKED AND UNMARKED DCT COEFFICIENTS

x_i	t_i	t_i*		
		$\alpha < 1$	$\alpha > 1$	$\alpha = 1$
+	+	+	+	+
+	–	–	+	0
–	+	+	–	0
–	–	–	–	–

generated to allow the direction of scaling to be recovered by the receiver in order to authenticate the watermark. According to the embedding equation, (12), the marked DCT coefficient, t_i^* can fall on either side of the corresponding unmarked coefficient, t_i. If the decision threshold is set to t_i, then the corresponding encrypted watermark bit, x_i encoded in $\{1, -1\}$ format can in principle, be recovered reliably if any malicious manipulation on the marked image attempts to modified t_i^* in the opposite direction by an amount not exceeding the embedding factor, $\alpha|t_i|$. What is required is a method to estimate the unmarked DCT coefficient value, t_i from the marked DCT coefficient, t_i^* and used it as a decision threshold for the recovery of the encrypted watermark pixel, $x_i \in \{1, -1\}$. A binary polarity mask indicating the relative position of t_i and t_i^* is generated and transmitted to the authenticator for this purpose.

It can be shown from the watermark embedding equation, (12), that the signs of x_i, t_i, and t_i^* are related and the relationship is given in Table I.

The results of Table I can be proven as follows. When $x_i = 1, t_i^* = t_i + \alpha|t_i|$

$$\begin{aligned}&\text{if } t_i > 0, t_i^* = (1+\alpha)|t_i| \Rightarrow t_i^* \geq 0 \text{ if } \alpha \geq -1\\ &\text{if } t_i < 0, t_i^* = -|t_i| + \alpha|t_i| = (\alpha - 1)|t_i| \Rightarrow t_i^* \geq 0 \text{ if} \alpha \geq 1\end{aligned}$$

When $x_i = -1, t_i^* = t_i - \alpha|t_i|$.

$$\begin{aligned}&\text{if } t_i > 0, t_i^* = (1-\alpha)|t_i| \Rightarrow t_i^* \geq 0 \text{ if} \alpha \leq 1\\ &\text{if } t_i < 0, t_i^* = -|t_i| - \alpha|t_i|\\ &\qquad = -(1+\alpha)|t_i| \Rightarrow t_i^* \geq 0 \text{ if} \alpha \leq -1.\end{aligned}$$

Two important and useful properties are noted from Table I, if the use of $\alpha = 1$ is avoided.

Property 1: If $\alpha < 1, \text{sign}(t_i^*) = \text{sign}(t_i)$

Property 2: If $\alpha > 1, \text{sign}(t_i^*) = \text{sign}(x_i)$

It should be noted that $\text{sign}(t_i)$ are image dependent and $\text{sign}(x_i)$ have been randomized, and they should have no correlation. According to the above properties, a polarity mask $\boldsymbol{P}^b = [p_i^b]$ for each reduced image block, R_i is generated where each bit p_i^b is generated based on the signs of t_i^* and x_i for $\alpha < 1$ and the signs of t_i^* and t_i for $\alpha > 1$. The value of p_i^b is determined as follows:

$$\text{For } \alpha < 1, \text{if } \text{sign}(t_i^*) = \text{sign}(x_i), p_i^b = 1. \text{ Otherwise}, p_i^b = 0. \tag{30}$$

$$\text{For } \alpha > 1, \text{if } \text{sign}(t_i^*) = \text{sign}(t_i), p_i^b = 1. \text{ Otherwise}, p_i^b = 0. \tag{31}$$

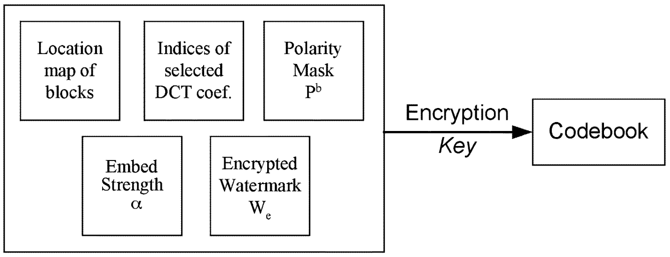


Fig. 5. Codebook content for oblivious detection.

Given the polarity mask, $\boldsymbol{P}^b$, each unmarked DCT coefficient, t_i can be determined from the corresponding marked DCT coefficient by Properties 1 and 2 as follows If $p_i^b = 1, \text{sign}(t_i^*) = \text{sign}(x_i) = \text{sign}(t_i)$ from (30), (31), Properties 1 and 2. If $t_i^* > 0$, then $t_i^* = t_i + \alpha|t_i|x_i = t_i + \alpha|t_i| = t_i + \alpha t_i = (1+\alpha)t_i$. Otherwise, if $t_i^* < 0, t_i^* = t_i + \alpha|t_i|x_i = t_i - \alpha|t_i| = t_i + \alpha t_i = (1+\alpha)t_i$. Hence

$$t_i = \frac{1}{1+\alpha} t_i * . \tag{32}$$

If $p_i^b = 0, \text{sign}(t_i^*) = -\text{sign}(x_i) = \text{sign}(t_i)$ for $\alpha < 1$ from (30) and Property 1. If $t_i^* > 0$, then $t_i^* = t_i + \alpha|t_i|x_i = t_i - \alpha|t_i| = t_i - \alpha t_i = (1-\alpha)t_i$. Otherwise, if $t_i^* < 0, t_i^* = t_i + \alpha|t_i|x_i = t_i + \alpha|t_i| = t_i - \alpha t_i = (1-\alpha)t_i$. For $\alpha > 1, \text{sign}(t_i^*) = -\text{sign}(t_i) = \text{sign}(x_i)$ from (31) and Property 2. If $t_i^* > 0$, then $t_i^* = t_i + \alpha|t_i|x_i = t_i + \alpha|t_i| = t_i - \alpha t_i = (1-\alpha)t_i$. Otherwise, if $t_i^* < 0, t_i^* = t_i + \alpha|t_i|x_i = t_i - \alpha|t_i| = t_i - \alpha t_i = (1-\alpha)t_i$. Hence

$$t_i = \frac{1}{1-\alpha} t_i * . \tag{33}$$

Based on the above analysis, an additional encrypted codebook is generated in the embedding process for oblivious detection. As shown in Fig. 5, this codebook is made up of the following five components:

1) polarity mask $\boldsymbol{P}^b$;
2) location map of the blocks selected for watermark insertion;
3) indexes of selected DCT coefficients for block reduction;
4) embedding strength α;
5) encrypted watermark W_e.

Since most of these information needed are binary data (polarity mask and encrypted watermark), the total size of these codebook information to be sent to the watermark verifier is limited. To reduce the amount of repetitive information to be transmitted, different selections of DCT coefficients and their corresponding embedding strengths can be categorized and each type of reduced block mapping is represented by a unique identifier. The overhead is calculated to be merely 2.3% of the size of the original image.

For the watermark retrieval, we need the watermarked image, the codebook, and the encryption key. The codebook is first decrypted to recover the polarity mask $\boldsymbol{P}^b$. The DCT coefficients, $\tilde{t}_i$ of the original image are estimated from the DCT coefficients, $\tilde{t}_i^*$ of the possibly corrupted received watermarked image according to (32) if $p_i^b = 1$ and (33) if $p_i^b = 0$. To recover the

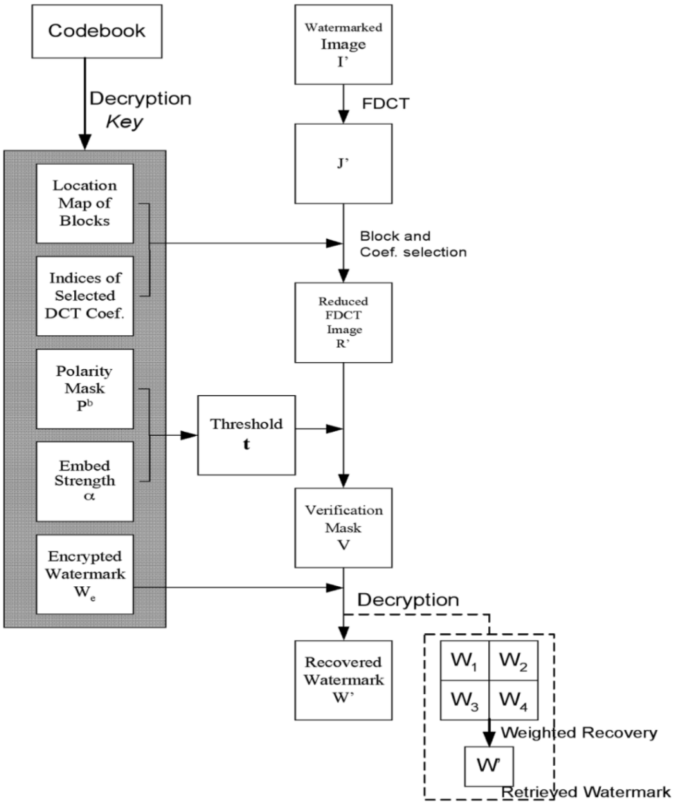


Fig. 6. Oblivious watermark recovery process.

encrypted watermark bit x_i, a verification mask, $\boldsymbol{V} = [v_i]$ is generated as follows:

$$\begin{aligned}&\text{if } \tilde{t}_i^* \geq \tilde{t}_i, \text{then } v_i = 0\\ &\text{Otherwise}, v_i = 1.\end{aligned} \tag{34}$$

Without attack, $\tilde{t}_i^* = t_i^*$ and $\tilde{t}_i = t_i$, a genuinely marked image will generate a verification mask $(\boldsymbol{V})$. The recovered $\boldsymbol{V}$ is XORed with the encrypted watermark $\boldsymbol{X}$ and then decrypts it to obtain the correct visual watermark. A wrong watermark will recover an incorrect verification mask. When this erroneous verification mask is XORed with the encrypted watermark $\boldsymbol{X}$, the extracted watermark bitmap will be illegible even if the output is decrypted with the correct key.

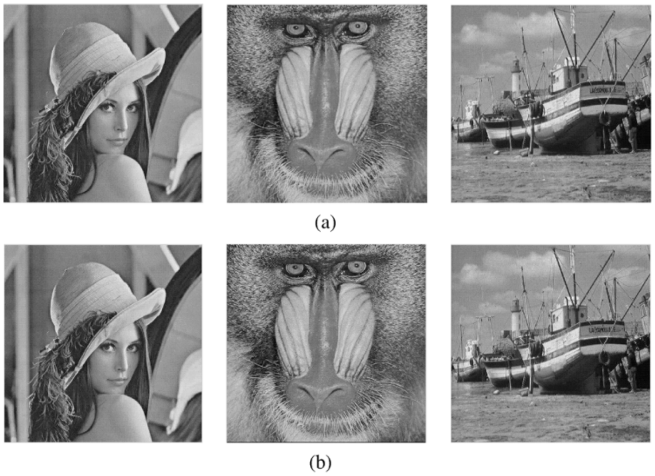
Fig. 6 shows the oblivious watermark recovery process. Since the oblivious detection method can also be used for weighted recovery of composite watermark, for ease of reference in presenting the experimental results, we abbreviate the proposed oblivious detection scheme as FAOS when it is applied to simple watermark, to distinguish it from FAOM when the same method is applied to the composite watermark.

IV. Experimental Results and Discussions

To evaluate the performance of the proposed watermarking schemes, three different types of images are chosen for the tests. The selected images are of size 512 × 512 pixels, and consist of 4096 image blocks of 8 × 8 pixels. The clustering statistics of the three test images marked by FAS scheme are given in Table II. The column "Ratio(%)" represents the fraction of cluster blocks to the total number of image blocks.

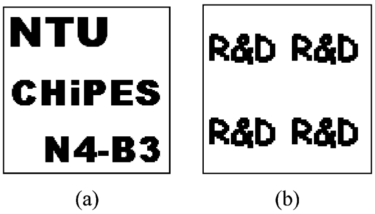
TABLE II
Fuzzy-ART Clustering Results

Name of Image	Image Type	No. of trained clusters	Ratio (%)
Lena	Smooth natural	708	17.3
Baboon	Rich in texture	1884	46.0
Boat	High contrast natural	1006	24.6


(a)

(b)

Fig. 7. Test images and their corresponding watermarked images. (a) Original images. (b) Watermarked images.



(a)

(b)

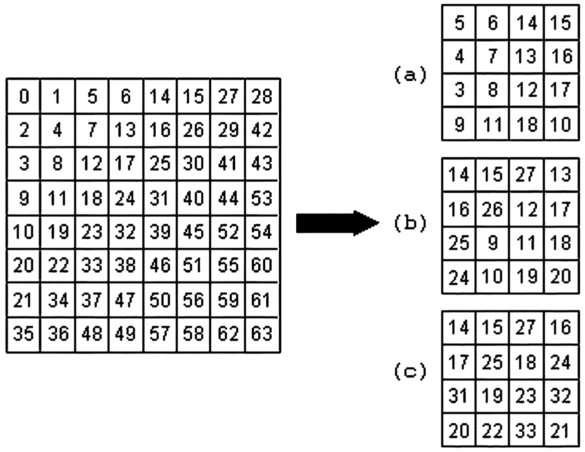
Fig. 8. Watermark images. (a) Simple watermark. (b) Composite watermark.

The three test images, namely: "Lena," "Baboon," and "Boat," with their respective watermarked versions (using the FAS scheme) are shown in Fig. 7. There is no visually distinguishable difference between the original images and their corresponding watermarked versions.

Fig. 8 shows the binary images for the simple and composite watermarks, both are of size 128 × 128 pixels, which is 1/16 the size of the host image.

The watermark image size, i.e., the amount of information to be embedded can be made variable by varying the number of selected blocks and the number of DCT coefficients to be modified. In this paper, a quarter of the host image (1024 blocks) is used to embed the watermark information. The selection of blocks is based on the Fuzzy-ART training process presented in Section II. We selected 16 DCT coefficients from each of the 8 × 8 DCT coefficient block. Depending on the noise sensitivity level of its associated cluster, three different sets of DCT coefficients can be selected from an image block to form a reduced image block of size 4 × 4. The mapping process is illustrated in Fig. 9. The coefficients are numbered from 0 to 63 in zigzag order.

Most perceptually based image compression algorithms use frequency sensitivity and spatial masking as their visual models [13]. Since JPEG is a widely supported standard image format,



0	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

(a)

5	6	14	15
4	7	13	16
3	8	12	17
9	11	18	10

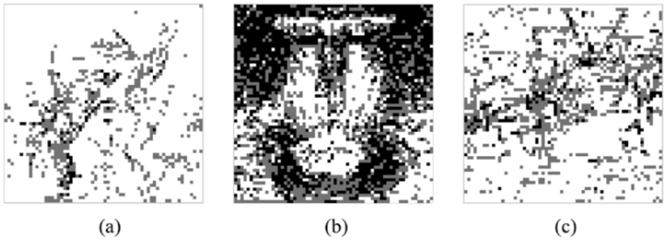
(b)

14	15	27	13
16	26	12	17
25	9	11	18
24	10	19	20

(c)

14	15	27	16
17	25	18	24
31	19	23	32
20	22	33	21

Fig. 9. Mapping of an 8 × 8 macroblock into a 4 × 4 reduced block.



(a) (b) (c)

Fig. 10. Distribution of low-, median-, and high-frequency clusters. (a) Lena. (b) Baboon. (c) Boat.

we use the number of zero-valued DCT coefficients after JPEG quantization as a simple distortion visibility measure to categorize each unique Fuzzy-ART clustered block into either low frequency ($\geq$52 zeros), median frequency (47–52 zeros) or high-frequency ($\leq$47 zeros) clusters. Fig. 10 shows the distribution of the three types of cluster in the test images. The low, median, and high-frequency clusters are marked in white, gray, and black colors, respectively. It can be seen that the low frequency blocks occupy most of the smooth image areas, the median frequency blocks correspond more to regions with high-edge features, and the high-frequency blocks correspond to highly textured areas. Since the distortion visibility in highly textured areas and edges is very low, we select as much of the high-frequency clusters, followed by the median and low frequency clusters to make up sufficient number of blocks to insert the watermark. To make the embedded information more resilient to JPEG compression attack, we modify the DCT coefficients toward the low-frequency region [Fig. 9(a)] for image blocks selected from the high-frequency clusters and toward the high-frequency region [Fig. 9(c)] for image blocks selected from the low frequency clusters. Fig. 9(b) is used for the image blocks of medium frequency clusters. To sustain the imperceptibility criterion, coefficients of block types Fig. 9(a)–(c) are modified with low, moderate, and strong embedding strength (α), respectively.

If the watermarked image has not been subjected to malicious manipulation, the extracted watermark would be a visually recognizable pattern. The viewer can compare the retrieved watermark with the original watermark visually. However, the subjective measurement is dependent on factors such as the experience of the viewers, the experimental conditions, etc. Besides, it is not suitable for automatic verification. Therefore, a quantitative measurement is needed to provide an objective judgment of the extracted watermark authenticity and for automation friendly verification. We use the normalized correlation (NC) [1] as an objective measure. NC is defined as follows

$$\mathrm{NC} = \frac{\sum_i \sum_j W(i,j)W'(i,j)}{\sum_i \sum_j |W(i,j)|^2} \tag{35}$$

Our watermarking scheme is deliberately designed to enhance the detectability of the object pixels of the watermark by virtue of the statistical property of the scaling direction of (12). Hence, instead of the summation over the entire watermark pixels in (35), the normalized correlation can be performed at only the coordinates of the object pixels. We should note that NC does not always correlate well with human vision, as with other difference distortion measures. However, it is still quite a fair benchmark for watermark methods applying on the gray-scale images [14]. Fig. 11 shows the correlation coefficients obtained from authenticating the watermarked "Boat" with 1000 random watermarks. In Fig. 11(a), "Boat" is watermarked by FAOS and subjected to a lossy JPEG compression with quality factors of 90 and 50. The quality factor provides an insight into the level of perceptual distortion suffered by an image upon compression and its values range from 0–100, where 100 being the highest perceptual quality, and 0 being the poorest. Its relation with the absolute compression ratio is image dependent and implicit. For "Boat," quality factors of 100, 90, and 50 are equivalent to absolute compression ratios of 1.63, 4.44, and 12.34, respectively. In Fig. 11(b), the same test is performed with FAOM scheme. In the simulation, 1000 random watermarks are generated, with the 500th watermark being the genuine one. Statistically, watermark with $\mathrm{NC} = 0.5$ represents the least likely case. From Fig. 11, the NC values of all but the no. 500 watermark are fluctuating around 0.5 (i.e., 0.5 ± 0.04 for FAOS and 0.55 ± 0.07 for FAOM) and the 500 sample shows an exceptionally high value, enough to distinguish itself clearly from the rest. The quality factor has very little effect on the NC value of the incorrect watermarks but the NC value of the genuine watermark drops more rapidly with reducing quality factor. A detection threshold is usually set to verify the authenticity of the watermark based on the desire to minimize both false positives and false rejections [13]. Through the detector responses of statistical tests, a detection threshold of 0.6 for FAOS and 0.65 for FAOM shall provide sufficient guard against the false positives and false rejections to use the NC value for verifying the existence of the watermark.

We have simulated a series of attacks on the watermarked images. In what follows, we will present the simulation results on FAS and FAM, followed by the simulation results on FAOS and FAOM. Lastly, the issue on counterfeiting watermark will be addressed.

A. Simulation Results on FAS Scheme and FAM Scheme

1) Cropping Tests: Cropping is the easiest operation among popular image manipulations. In this set of tests, each watermarked image is subjected to a wide range of cropping ratios. Fig. 12 and Fig. 13 show the cropped images with their corresponding extracted watermarks after the two groups of cropping

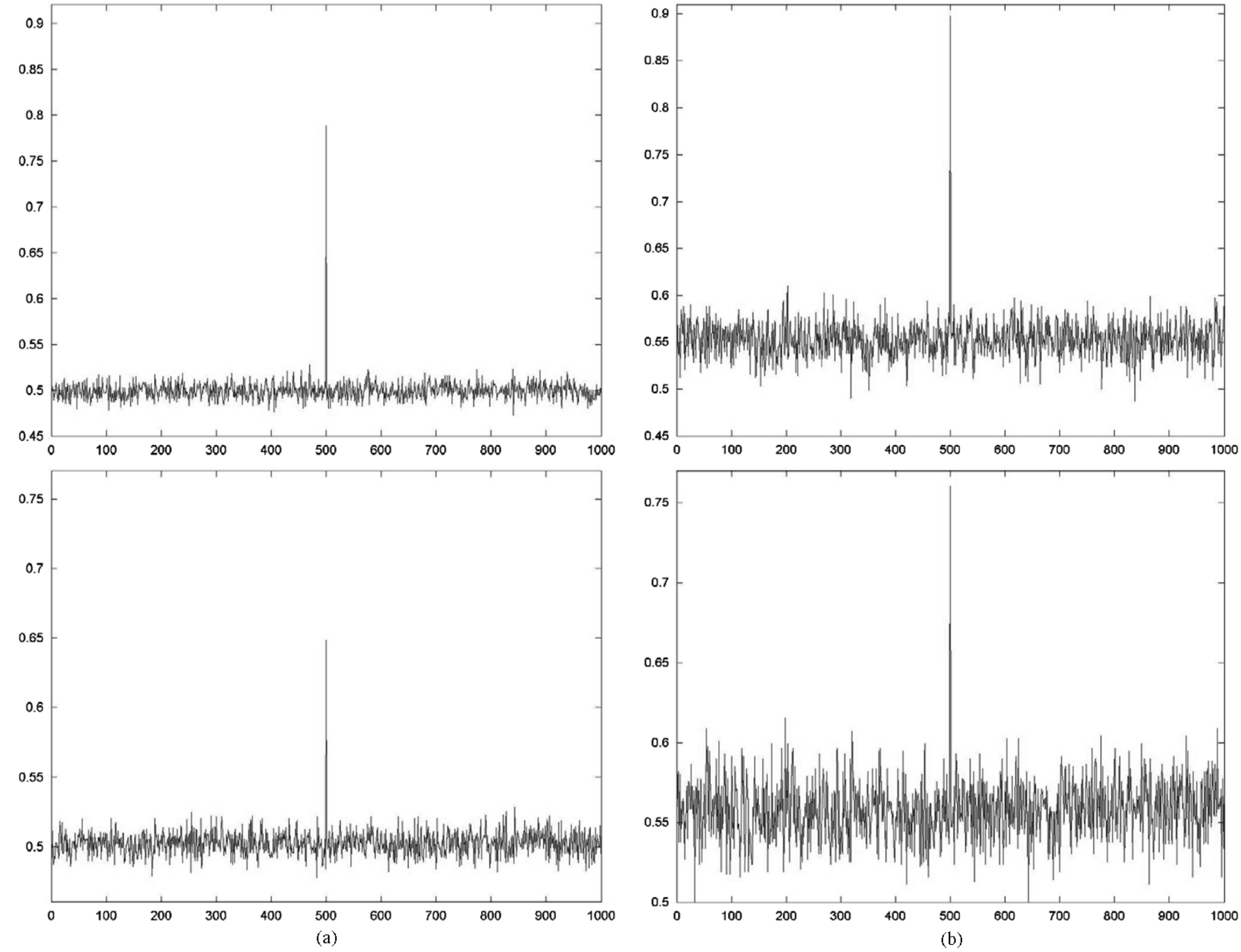


Fig. 11. Correlation results for watermarked "Boat" subjected to JPEG compression ratio with quality factors of 90 (top) and 50 (bottom) using (a) FAOS and (b) FAOM.

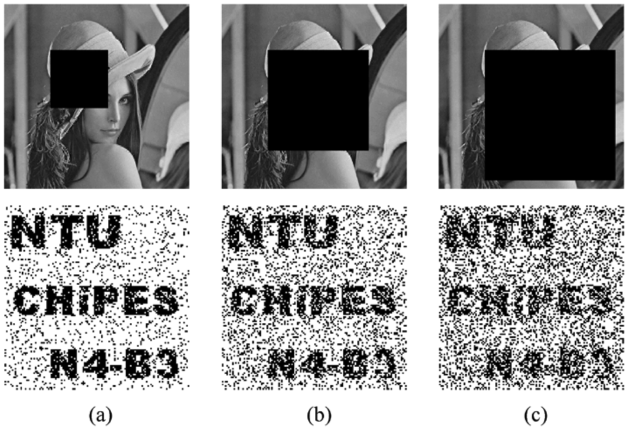


Fig. 12. Extracted watermarks from cropped FAS marked image with different cropping ratios. (a) 10%. (b) 30%. (c) 50%.

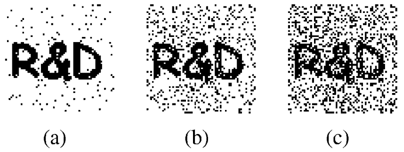


Fig. 13. Extracted watermarks from cropped FAM marked image with different cropping ratios. (a) 10%. (b) 30%.(c) 50%.

attacks on FAS and FAM watermarked images. From Fig. 12 and 13, the embedded watermark information is still legibly recovered even after 50% of the marked image has been cropped.

The objective similarity metrics NC are also computed and charted for comparison in Fig. 14, where the dotted and solid curves represent the NC values of FAS and FAM schemes, respectively. The NC values are plotted against the percentage of cropped area in Fig. 14.

For cropping attacks, it is found that the watermark can survive large cropping ratio in all three types of test images. This can be explained by the fact that the selected blocks are distributed all over the image and the cropped area may contain only a fraction of embedded information. The risk of losing large amount of information will come into picture only if the selected blocks are concentrated in some area. This can be avoided by selecting only the nonneighboring blocks in addition to the mentioned selection criteria. Reducing the number of neighboring blocks may imply a relaxation

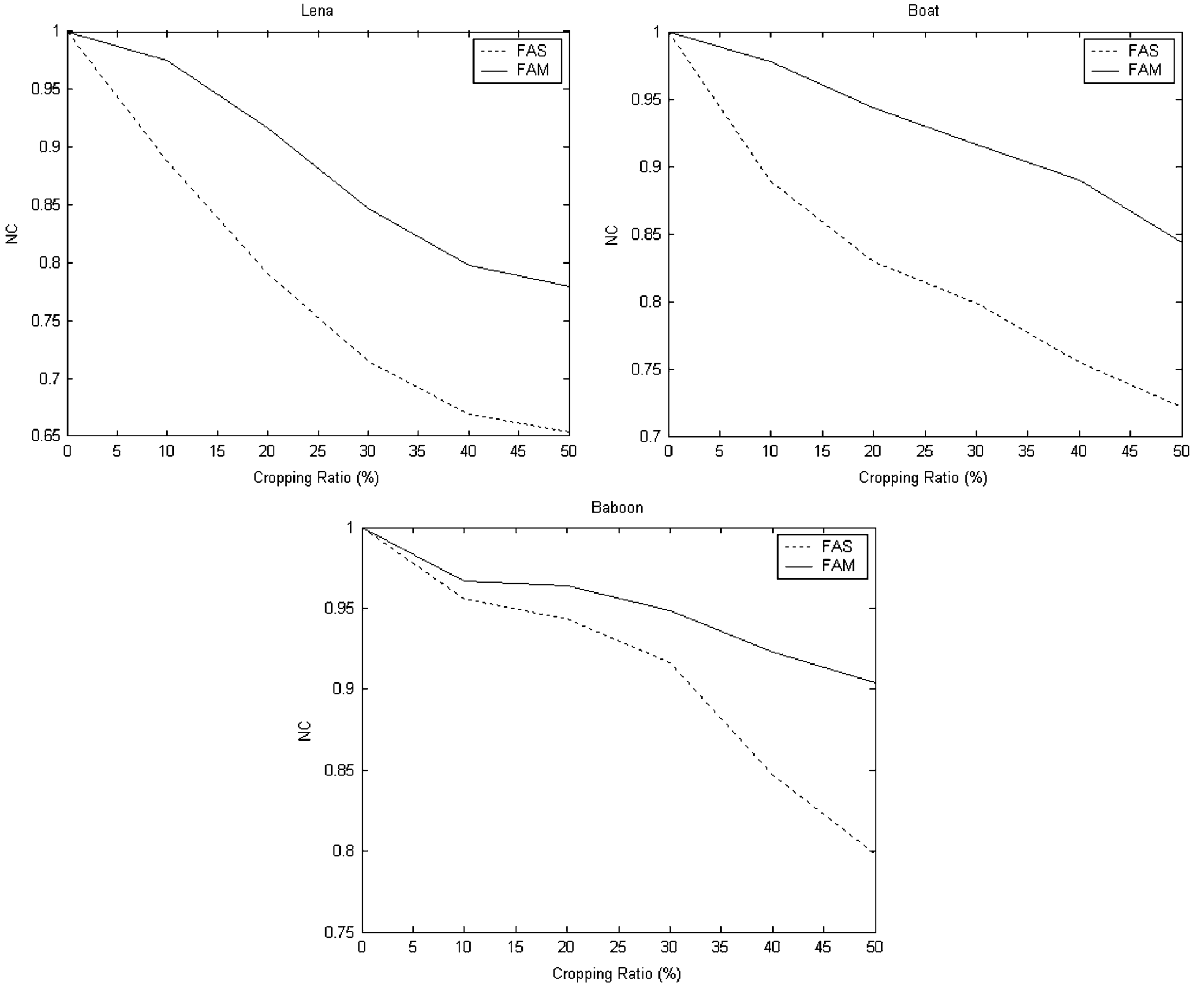
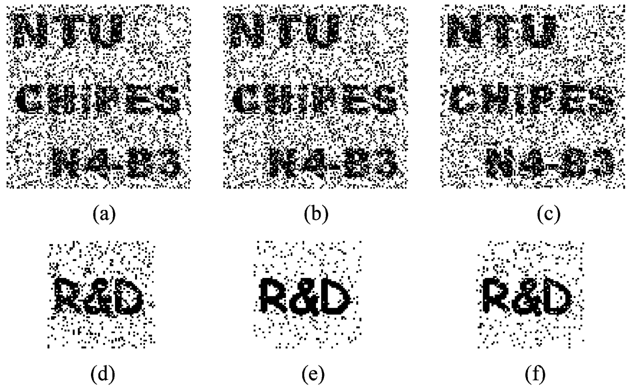


Fig. 14. Cropping tests for FAS and FAM schemes.

in the criteria for selecting the clusters in order to locate sufficient number of blocks for embedding. From the results of Fig. 14, the NC curves for the FAM scheme always lie above their respective FAS counterparts, suggesting positively the robustness enhancement provided by the weighted recovery method.

2) Compression Tests: Being the most classical and ubiquitous image processing attack, JPEG compression of various compression ratios is applied to the watermarked images. Fig. 15 shows the extracted watermark from FAS marked images [Fig. 15(a)–(c)] and FAM marked images [Fig. 15(d)–(f)] after the compression attack. The number in the bracket denotes the compression ratio.

The extracted watermarks are visually recognizable, despite being contaminated by noisy spots. Nevertheless, the objective measurements detect the existence of watermark unambiguously. The similarity metrics plotted against the quality factor of JPEG compression for the test images are shown in Fig. 16, where the dotted and solid curves are used to represent the NC values of FAS and FAM schemes, respectively.



(a) (b) (c)

(d) (e) (f)

Fig. 15. Extracted watermarks from compressed watermarked images. (a) Lena (8.89). (b) Baboon (8.40). (c) Boat (8.34). (d) Lena (8.89). (e) Baboon (8.40). (f) Boat (8.43).

From Fig. 16, the NC value is relatively high over a wide range of quality factors, which suggests that the proposed watermarking scheme is robust to the compression attack. The

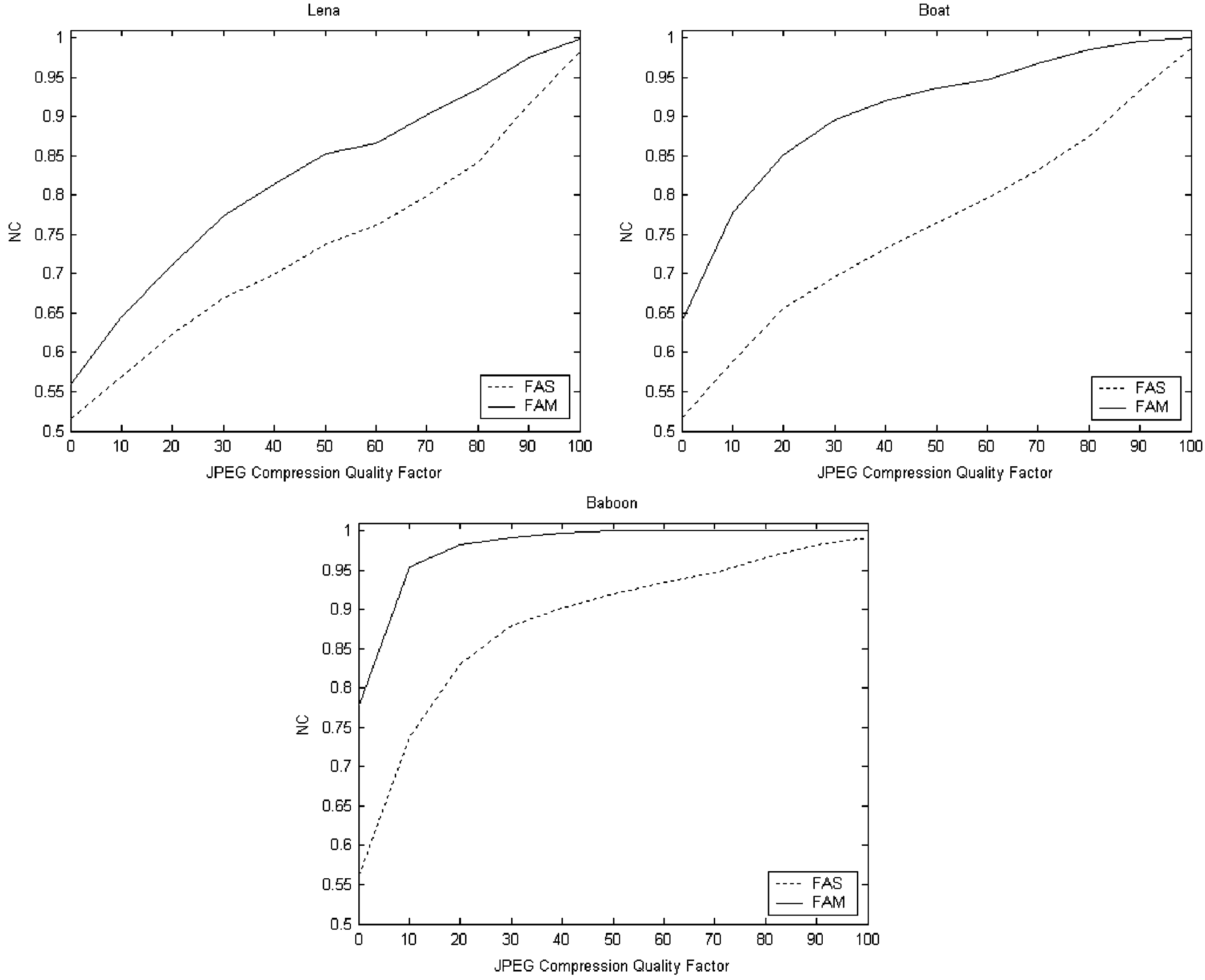


Fig. 16. Compression test for FAS and FAM schemes.

weighted recovery technique has better performance, as evident from the FAM curve lying consistently well above the FAS curve. It is noticed that both curves experience slower degradation in the similarity metrics with reduced quality factor for high texture images like "Baboon". Hsu and Wu [12] and Luo *et al.* [15] conducted the same set of tests on "Lena". Fig. 17 shows that both FAS and FAM outperforms Hsu and Wu's scheme at compression ratio higher than 7, while Luo's results have slightly better performance than FAS, it loses out to FAM when the compression ratio is higher than 12. The differences in embedding capacity should be noted in the above comparison. For a 512 × 512 gray level image, the resolutions of the binary watermark images of Hsu and Wu [12], Luo *et al.* [15] and ours are 256 × 256, 32 × 32, and 128 × 128, respectively. Based on our adaptive embedding scheme, if we chose to reduce the amount of information to be embedded by not modifying some of the high-frequency coefficients, our FAS can also outperform Luo's method. Hsu and Wu achieve a high embedding capacity by embedding the watermark pixels globally across the host image. Although we can increase the embedding capacity to match their method, we foresee a similar steep degradation of performance at some point despite our performance can still be better than theirs at high compression ratio. The superior performance is owing to the appropriate choice of watermark insertion regions coupled with the adaptive embedding strength.

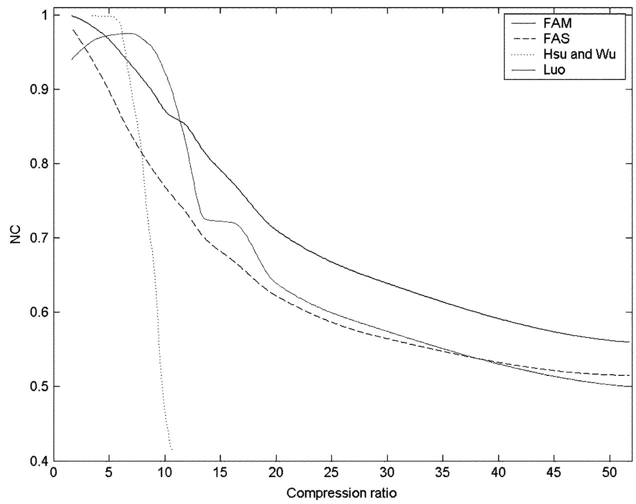


Fig. 17. Results comparison for JPEG compression attacks.

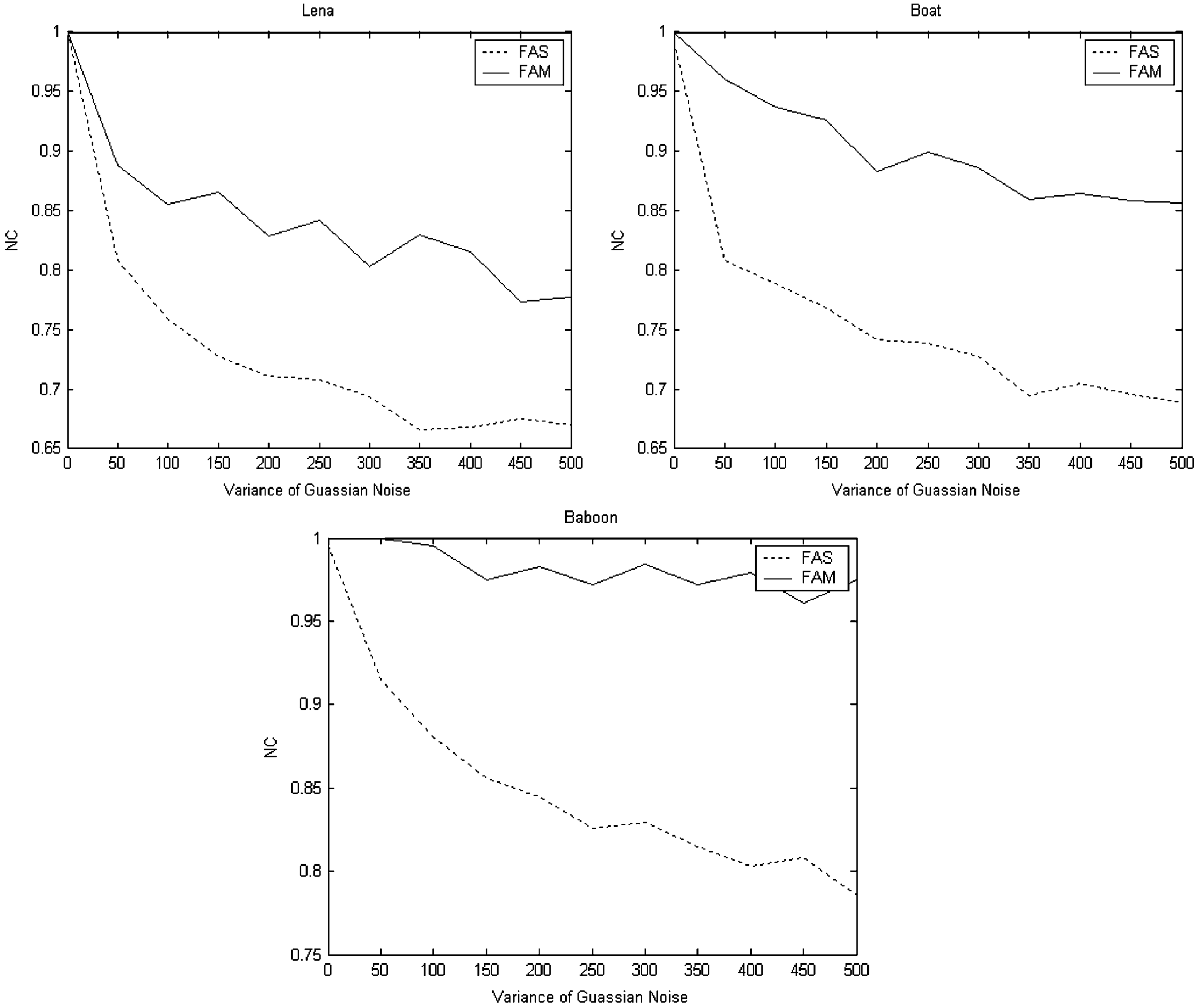


Fig. 18. Additive noise attacks for FAS and FAM schemes.

3) Additive Guassian Noise Attacks: In this set of test, Guassian noise of different energy levels is applied to the watermarked images, so as to study the robustness to additive noise attacks. The results are summarized in Fig. 18.

From Fig. 18, we can see that the proposed watermarking scheme outperforms other schemes in most cases [12], [15], [21]. Again, FAM proves to be more robust against the additive noise attacks. In all the three types of images, the NC value for FAM is well above 0.75.

Similarly, results from Luo's method [15] is charted in Fig. 19 for comparison. As shown in Fig. 19, the FAM scheme yields better performance when the noise energy level is above 120, though the FAS scheme is less robust in this type of attack.

4) Other Image Processing Attacks: The watermarked images are also subjected to several standard image processing attacks including: 1) filtering; 2) scaling (scale down to half size and recover); 3) sharpening; and 4) blurring. Fig. 20 shows the attacked FAS images and their corresponding extracted watermarks.

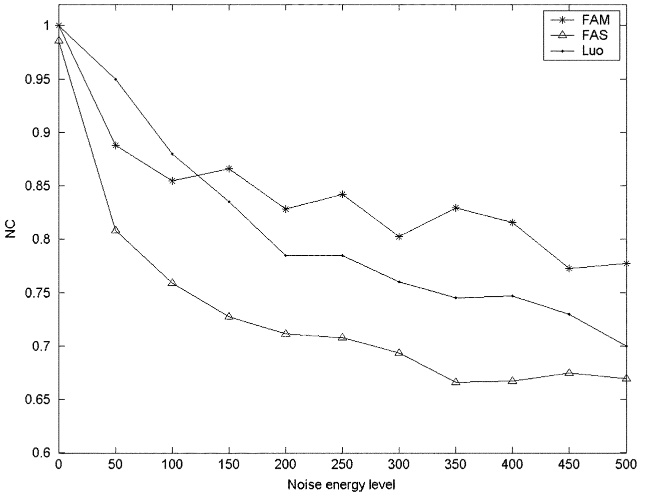


Fig. 19. Results comparison for additive Gaussian noise attacks.

All extracted watermarks are recognizable to different extent, and their corresponding NC values are also high enough to

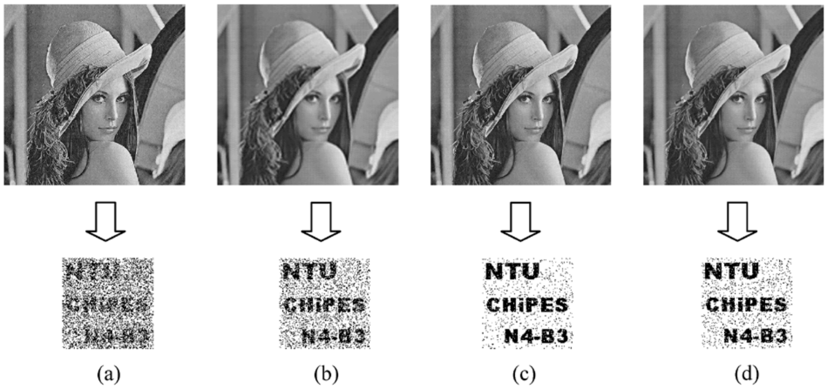


Fig. 20. Various attacked FAS images and their retrieved watermarks.

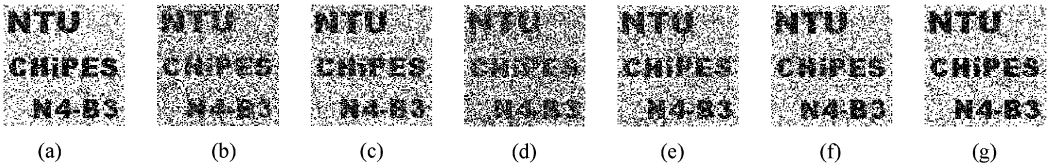


Fig. 21. Decoded watermarks for FAOS.

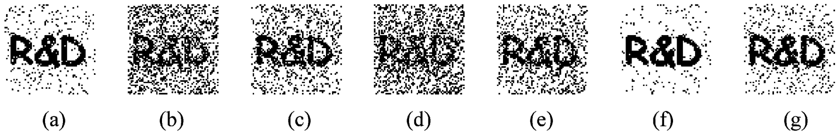


Fig. 22. Decoded watermarks for FAOM.

TABLE III
SIMULATION RESULTS ON VARIOUS ATTACKS FOR FAS AND FAM

Image / Operation	Lena		Boat		Baboon	
	FAS	FAM	FAS	FAM	FAS	FAM
Filtering	0.6753	0.7792	0.6945	0.7950	0.6492	0.8754
Scaling	0.8132	0.9290	0.8001	0.9574	0.8333	0.9779
Sharpening	0.9416	0.9968	0.9259	0.9921	0.9143	0.9968
Blurring	0.9164	0.9905	0.9060	0.9890	0.9069	0.9937

indicate the presence of the watermark. The similarity metric NC for the above mentioned attacks are summarized in Table III.

Experimental results show that the proposed watermarking scheme is less robust against filtering attacks, but relatively resilient to scaling, sharpening and blurring attacks.

B. Simulation Results on FAOS and FAOM

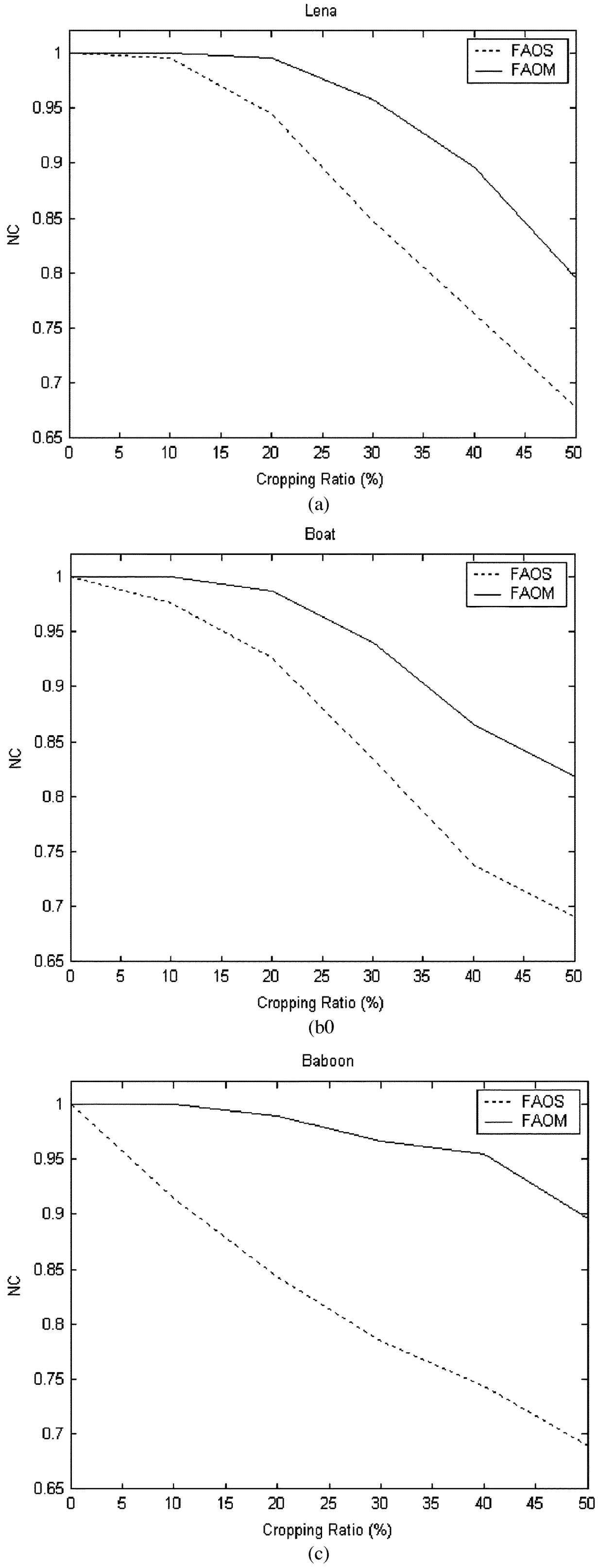
Similarly, for FAOS and FAOM schemes, the watermarked images are subjected to cropping, JPEG compression and other image manipulation attacks. Some of the decoded watermarks are shown below in Figs. 21 and 22, where the marked image is manipulated by the following operations: 1) cropped by 25% of the image area; 2) JPEG compression with quality factor 70; 3) Gaussian noise with energy level 50; 4) median filtered; 5) scaled down to half sized then resized back to original size; 6) sharpened; and 7) blurred. The simulation results are shown in Figs. 23–25, and Table IV.

It is common for the oblivious watermarking scheme to be less robust than its coherent counterparts [14]. Nevertheless, the results show that for most of the image processing attacks, the proposed schemes still perform satisfactorily.

C. Counterfeiting Attacks

In counterfeiting attacks, the forged watermark image can be embedded into the host image without the consent of the rightful image owner. Holliman and Memon [11] proved that block-wise independent watermarking schemes are vulnerable to counterfeiting attacks. In block-wise independent watermarking schemes, the process of watermark insertion and extraction are carried out block by block and independently. Holliman and Memon claim that the attackers can forge watermark by replacing blocks from the target images with K-equivalent blocks in the known watermarked images. Two blocks are said to be K-equivalent if the same partial watermark information can be extracted with the same key. In this attack, it is assumed that the attackers do not possess the key to decrypt the watermark information. However, the attackers know the watermark bitmap and have a collection of watermarked blocks to make K-equivalent block substitutions that minimize the distortion between the original and the forged watermarked blocks.

One unique feature of the proposed watermarking schemes is the subtle block-wise dependency. Besides the nature of the image blocks, there are several parameters that will result in different classifications of the image blocks, for example, the vigilance parameter and learning rate chosen for the Fuzzy-ART

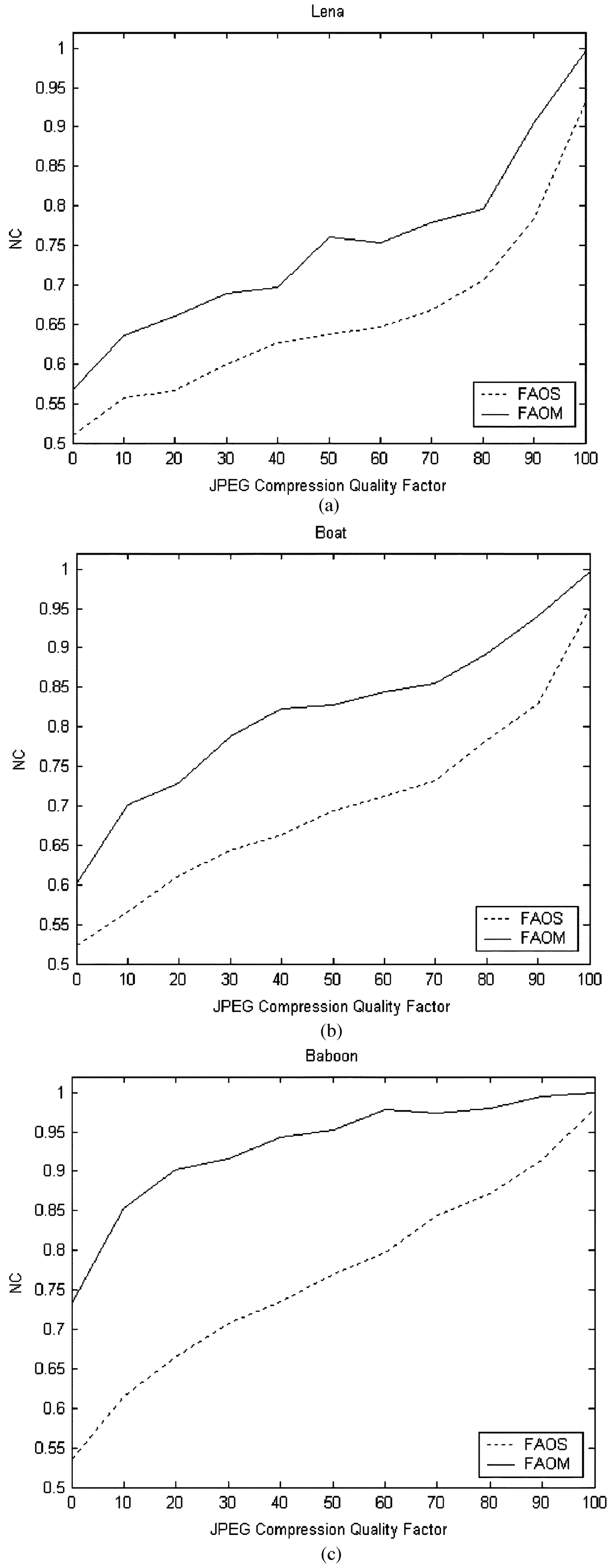


(a)

(b0

(c)

Fig. 23. Cropping attacks on FAOS and FAOM marked images.



(a)

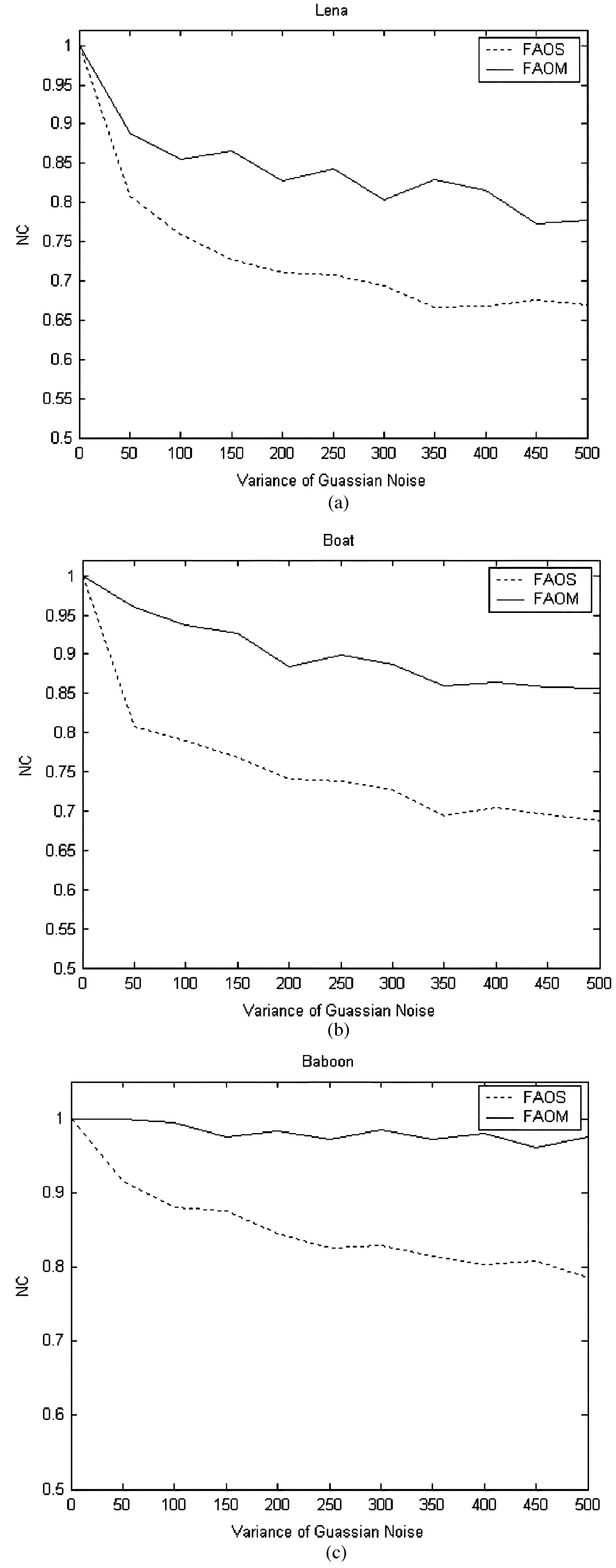
(b)

(c)

Fig. 24. Compression tests on FAOS and FAOM marked images.

training, which in term leads to different embedding strength and different sets of blocks selected for embedding. Without the knowledge of clustering and their corresponding embedding strength, the attackers will not be able to forge the watermark.

Supposed the attackers already have a watermarked image A in their database, and they want to forge the watermark contained in image A into image B. In the proposed schemes, host image A is classified into different classes. Let C_A denote the



(a)

(b)

(c)

Fig. 25. Additive noise attacks for FAOS and FAOM schemes.

set of chosen clusters in A, and $C_i, i \in [1, \mathrm{m}]$, denote a selected cluster

$$C_A = C_1 || C_2 || \ldots || C_m.$$

TABLE IV
OTHER IMAGE MANIPULATIONS ON FAOS AND FAOM MARKED IMAGES

Image / Operation	Lena		Boat		Baboon	
	FAOS	**FAOM**	**FAOS**	**FAOM**	**FAOS**	**FAOM**
Filtering	0.6367	0.7666	0.7103	0.8297	0.7596	0.9543
Scaling	0.7281	0.8297	0.7782	0.8943	0.8381	0.9858
Sharpening	0.8763	0.9842	0.8063	0.9448	0.8407	0.9921
Blurring	0.8108	0.9196	0.8455	0.9558	0.8947	0.9968

Assume that the image block $J_p \in C_i$ and it resembles image block $J'_q \in C'_i$ in image B. Under the counterfeiting attacks [11], this candidate block J_p is selected from a collection of watermarked image blocks, base on its similarity in frequency content or spatial features to the image block J'_q. The attackers will forge the watermark by replacing J'_q with J_p while in fact image B has a different set of chosen clusters

$$C_B = C'_1 || C'_2 || \ldots || C'_n, \quad \text{and } C_i \neq C'_i.$$

In addition, it is very likely that C_i and C'_i correspond to two clusters of different embedding strength. In other words, it is highly probable that they are not K-equivalent. The probability of getting sufficient number of matching K-equivalent pairs reduces exponentially as n increases. Therefore, the attackers cannot forge watermark by simply replacing visually similar blocks from the database of known watermarked images. Hence, the proposed adaptive watermarking schemes have presented significant barriers for the counterfeiting attack.

In summary, the clustering parameters, adaptive embedding strength and the location map act like a special serial number unique to each host image which is not known to the attackers without the secret key. Since the possibility of having the watermarks of two different images embedded using the same set of parameters is very low, even if the attackers can find sufficient number of exact K-equivalent blocks to carry out the forging, it is almost certain that the resulting watermark decrypted by the legitimate recipient will be illegible.

V. CONCLUSION

A new robust transform domain watermarking scheme has been presented which uses a visually meaningful binary image as watermark for image authentication. In the proposed scheme, Fuzzy-ART is used to locate the spectral positions and the modulating factors according to the perceptibility of the image, where the cryptographically randomized watermark information is embedded. Thus, the embedding strength is adaptive to both the modulated coefficient values as well as their local block features, making the watermark more resistant to various standard image processing and the counterfeiting attacks. The adaptive property can also be exploited in the watermark verification process to enhance its detectability. Compared to many robust watermarking schemes, the proposed watermarking schemes have relatively large embedding capacity. The dimension of the binary watermark image is about four times smaller than that of the host image. When the amount of information to be embedded is small or the watermark image can be shrunk without affecting its legibility, a weighted recovery technique can be applied to enhance the detectability of the composite watermark composing of multiple duplicates of the original reduced watermark. With the aid of a set of binary polarity masks that indicate the relative magnitude of the marked and unmarked DCT coefficients, oblivious detection is made possible where the watermark can be extracted without the need to transmit or store the original image for verification. The overhead incurred by the additional encrypted information for watermark detection is less than 3% of what is required to transmit the original image. Finally, we would like to note the potential of extending our watermarking schemes to safeguard against geometrical transformations by the resynchronization technique proposed by Izquierdo in [13], which involves the use of low complexity first-order differential invariants to extract the salient image features for affine mappings.

REFERENCES

[1] M. Barni, F. Bartolini, V. Cappellini, and A. Piva, "A DCT-domain system for robust image watermarking," *Signal Process.*, vol. 66, no. 3, pp. 357–372, 1998.

[2] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Syst. J.*, vol. 35, no. 3&4, pp. 313–336, 1996.

[3] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "Fuzzy ART: An adaptive resonance algorithm for rapid, stable classification of analog patterns," in *Proc. Int. Joint Conf. Neural Networks*, 1991, pp. 411–416.

[4] C. H. Chang, M. Zhang, and Z. Ye, "A content-dependent robust and fragile watermarking scheme," in *Proc. 2nd IASTED Int. Conf. Visualization, Imaging and Image Processing*, Sep. 2002, pp. 201–206.

[5] C. H. Chang, Z. Ye, and M. Zhang, "Fuzzy-ART based digital watermarking scheme," in *Proc. IEEE Asia Pacific Conf. Circuits and Systems*, vol. 1, APCCAS-2002, Dec. 2002, pp. 425–426.

[6] D. Coppersmith and P. Rogaway, "Software-Efficient Pseudorandom Function and the Use Thereof for Encryption," U.S. Patent 5 454 039 , Sep. 26, 1995.

[7] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Process.*, vol. 6, no. 12, pp. 1673–1687, Dec. 1997.

[8] C. Dautzenberg and F. M. Boland, "Watermarking Images," Dept. Electron. and Elect.Eng., Trinity College Dublin, Tech. Rep., 1994.

[9] J. J. Eggers and J. K. Su, "A blind watermarking scheme based on structured codebooks," in *Proc. IEE Secure Images and Image Authentication Colloq.*, London, U.K., 2000, pp. 4/1–4/21.

[10] T. S. Gotarredona, B. L. Barranco, and A. G. Andreou, *Adaptive Resonance Theory Microchips Circuit Design Techniques*. Norwell, MA: Kluwer, 1998, pp. 23–30.

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

[12] C. T. Hsu and J. L. Wu, "Hidden digital watermarks in images," *IEEE Trans. Image Process.*, vol. 8, no. 1, pp. 58–68, Jan. 1999.

[13] E. Izquierdo, "Using invariant image features for synchronization in spread spectrum image watermarking," in *EURASIP J. Appl. Signal Process.*, vol. 4, 2002, pp. 410–417.

[14] S. Katzenbeisser and F. Petitcolas, Eds., *Information Hiding Techniques for Steganography and Digital Watermarking*. Boston, MA: Artech House, 2000.

[15] W. Luo, G. L. Heileman, and C. E. Pizano, "Fast and robust watermarking of JPEG files," in *Proc. IEEE 5th Southwest Symp. Image Analysis and Interpretation*, Apr. 2002, pp. 158–162.

[16] C. Rey and J. L. Dujelay, "Blind detection of malicious alterations on still images using robust watermarks," in *Proc. IEE Secure Images and Image Authentication Colloquium*, London, U.K., Apr. 2000, pp. 7/1–7/6.

[17] J. J. K. O. Ruanaidh, W. J. Dowling, and F. M. Boland, "Watermarking digital images for copyright protection," *IEE Proc. Vis. Image Signal Process.*, vol. 143, no. 4, pp. 250–256, Apr. 1996.

[18] W. S. Sarle, "Neural networks and statistical models," in *Proc 19th Annu. SAS User Group Int. Conf.*, Apr. 1995, pp. 1538–1550.

[19] R. G. Schyndel, A. Z. Tirkel, and C. F. Osborne, "A digital watermark," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, Nov. 1994, pp. 86–90.

[20] P. W. Wong and N. Memon, "Secret and public key image watermarking schemes of image authentication and ownership verification," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1593–1601, Oct. 2001.

[21] J. Zhao and E. Koch, "Embedding robust labels into images for copyright protection," in *Proc. Int. Congress IPR for Specialized Information, Knowledge and New Technologies*, Vienna, Austria, Aug. 1995, pp. 242–251.

Chip-Hong Chang (S'92–M'98–SM'03) received the B.Eng. (hons.) degree in electrical engineering from National University of Singapore, Singapore, in 1989, and the M.Eng and Ph.D. degrees from the School of Electrical and Electronic Engineering of Nanyang Technological University (NTU), Singapore, in 1993 and 1998, respectively.

His industrial experience includes being a Component Engineer of General Motors, Singapore, and a Technical Consultant of Flextech Electronics Pte. Ltd. He joined the Electronics Design Centre of Nanyang Polytechnic as a Lecturer in 1993. Since 1999, he has been with the School of Electrical and Electronic Engineering, NTU, Singapore, where he is currently an Assistant Professor. He has served a number of administrative roles during his academic career. He holds concurrent appointments at the university as the Deputy Director of the Centre for High Performance Embedded Systems and the Program Director of VLSI Design and Embedded Systems research group of the Centre for Integrated Circuits and Systems. His current research interests include low-power arithmetic circuits, design automation and synthesis, and algorithms and architectures for digital image processing. He has published about 90 refereed international journal and conference papers and book chapters.

Zhi Ye received the B.Eng. and M.Eng degrees (hons.) from Nanyang Technological University, Singapore, in 2002 and 2004, respectively.

She is currently working in the School of Electrical and Electronic Engineering of Nanyang Technological University as a Project Officer. Her research interests include digital watermarking and digital signal processing.

Mingyan Zhang received the B.Eng. and M.Eng. degrees (first class hons.) from Nanyang Technological University, Singapore, in 2002 and 2004, respectively.

She is currently working as a Failure Analysis Engineer with Tech Semiconductor Singapore Pte. Ltd. Her research interests include digital watermarking and low-power arithmetic circuits.