Digital-Invisible-Ink Data Hiding Based on Spread-Spectrum and Quantization Techniques

Chun-Hsiang Huang, Shang-Chih Chuang, and Ja-Ling Wu, *Fellow, IEEE*

***Abstract*—A novel data-hiding methodology, denoted as digital invisible ink (DII), is proposed to implement secure steganography systems. Like the real-world invisible ink, secret messages will be correctly revealed only after the marked works undergo certain prenegotiated manipulations, such as lossy compression and processing. Different from conventional data-hiding schemes where content processing or compression operations are undesirable, distortions caused by prenegotiated manipulations in DII-based schemes are indispensable steps for revealing genuine secrets. The proposed scheme is carried out based on two important data-hiding schemes: spread-spectrum watermarking and frequency-domain quantization watermarking. In some application scenarios, the DII-based steganography system can provide plausible deniability and enhance the secrecy by taking cover with other messages. We show that DII-based schemes are indeed superior to existing plausibly deniable steganography approaches in many aspects. Moreover, potential security holes caused by deniable steganography systems are discussed.**

***Index Terms*—Digital invisible ink (DII), plausible deniability, steganalysis, steganography.**

I. Introduction

HIDING secrets and only revealing them to certain users may be instincts of humankind. Looking back on history, events involving the art of secret hiding can be found in many aspects of human civilization. Before the digital era, messages are often hidden with steganographic skills. Steganographic techniques in the real world fall into two branches, linguistic steganography and technical steganography, as introduced in [1]. Linguistic steganography consists of two classes of methods: delivering secret messages via an open code where prior agreements about the true meaning of seemingly harmless phrases, gestures or expressions must be negotiated in advance, as well as semagrams that secrets are expressed in the form of visible but minute graphic details in a drawing or script.

A. Invisible Ink in the Real World

As for technical steganography, writing with invisible ink is the most renowned skill. Certain liquids like lemon juice or milk have proved popular and effective since ancient times. In general, invisible ink is a substance used in steganographic schemes so that secret messages can be invisibly written on papers. The ink is invisible during writing or soon thereafter. Later on, the hidden message may be developed (made visible) by different methods according to the type of adopted invisible ink. For example, messages written with diluted acid liquids can be developed by heating the paper. Development methods for other types of invisible inks include applying chemical liquids or vapors upon the paper, viewing the paper under ultraviolet light, and so on.

Fig. 1 shows a conventional espionage scenario in which invisible ink is used. When a sender wants to deliver some intelligence to a certain receiver over a supervised channel, he can write secret messages on the paper using some acid liquid. Note that, usually, the paper also carries some cover messages written with normal ink because sending a blank sheet of paper might arouse suspicion. The supervisor cannot find any anomaly in the paper under common viewing conditions. When the intended receiver gets the paper, some prenegotiated manipulations, e.g., the heating operation in Fig. 1, should be performed first so that the secret messages can be revealed. An introduction to invisible ink used by secret operation agents during World-War II can be found in [2].

Some characteristics of invisible-ink based steganography in the real world are summarized as follows:

1) Prenegotiated manipulations are indispensable steps for the correct extraction of genuine secrets. When certain types of invisible inks are used, corresponding development methods must be performed on the received paper to reveal the hidden message.
2) The received paper may be seriously deformed due to prenegotiated manipulations, e.g., the paper may become lumpy after heating. Since extraction of the genuine message is the real goal, visual quality of the paper after manipulation is often out of concern.
3) Cover messages play an important role as camouflage during delivery of genuine messages. However, spaces where genuine secrets can be invisibly written are always available since hand-written or printed documents always leave blank spaces for the ease of reading.

B. Steganography versus Steganalysis

After entering the digital era, digital media like images or audio clips serve as good cover objects for carrying secret messages. Therefore, digital data-hiding techniques are adopted to implement steganography systems. In a data-hiding system, the sender embeds some messages, also denoted as watermarks for some other applications, into the cover work for generating a

Manuscript received January 23, 2007; revised August 3, 2007. This work was supported in part by the Ministry of Education of Taiwan, R.O.C., under the Grant of Excellent Research Projects of National Taiwan University, 95R0062-AE00-02. The associate editor coordinating the review of this paper and approving it for publication was Dr. Mohan Kankanhalli.

The authors are with the National Taiwan University, Taipei, Taiwan, R.O.C. (e-mail: bh@cmlab.csie.ntu.edu.tw, peiz@cmlab.csie.ntu.eud.tw, wjl@cmlab.csie.ntu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2008.921733

1520-9210/$25.00 © 2008 IEEE

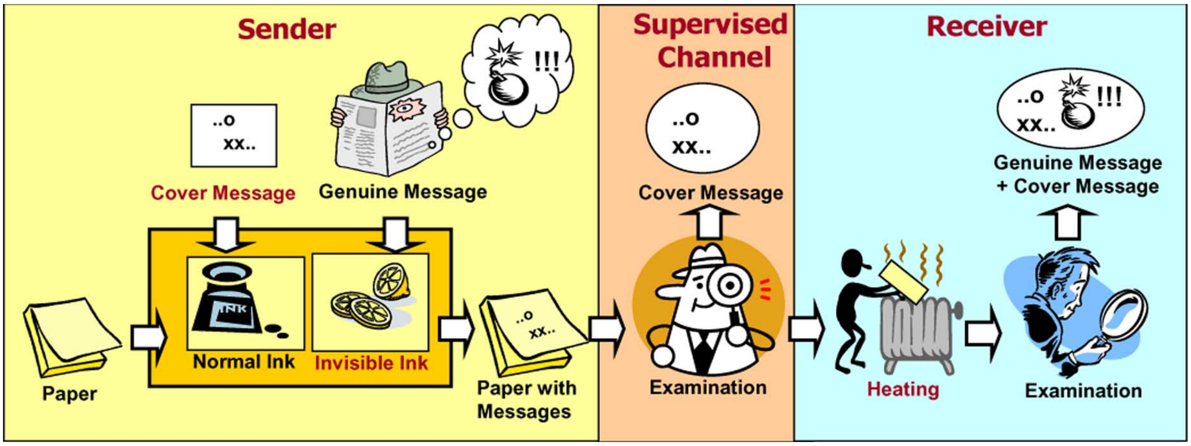


Fig. 1. Real-world espionage scenario using the invisible ink.

perceptually acceptable stego work. Afterwards, the receiver extracts the message from the received stego work. In-depth discussions about data-hiding systems can be found in [3]–[7]. In the literature of digital steganography, fidelity (the visual quality of the stego work) and capacity (the maximally allowable message length) are most important performance measures.

On the contrary, steganalysis is the practice of attacking steganographic schemes by detecting, destroying or extracting the hidden message, as introduced in [8]–[11]. Supervisors of communication channels may adopt adequate steganalysis tools to prevent unexpected communications via transmitted content. Generally speaking, deciding whether a cover work carries hidden messages is a difficult task. When the channel supervisor has the right to cease any doubtful communication, an accurate steganography detection module suffices for all his needs. Alternatively, a channel supervisor may simply introduce imperceptible or acceptable distortions to all incoming contents in expectation of hindering the extraction of potentially hidden messages. In this case, the supervisor should devise versatile and effective noises applicable to all delivered contents, while designers of steganography systems should consider robustness against potential distortions, as well as prescribed fidelity and capacity requirements. Besides, the supervisor may try to eavesdrop and interpret the secret messages. Therefore, cipher modules are often applied to the messages delivering by the sender to prevent the unauthorized receivers from reading the secret message.

Note that, in the literature, covert communications are often modeled as "the prisoners' problems" given different assumptions and the channel supervisor is therefore denoted as a warden, as firstly introduced by [12]. According to definitions specified in [13], a channel supervisor who can do nothing but only spy on the communication channels is named as a "passive" warden. On the other hand, the supervisor who is allowed to slightly modify the data being sent from the sender to the receiver is thereby an "active" warden. In this paper, most of our discussions focus on the passive-warden scenario. But for the completeness of our works, issues about the proposed methodology against the more difficult "active-warden" configuration is discussed in Section III-E.

C. *Plausible Deniability*

Plausible deniability is originally a term used in politics. It means the creation of loose and informal chains of commands in government, which allow controversial instructions given by high-ranking officials to be denied if these instructions become public. In the field of cryptography, deniable encryption allows an encrypted message to be decrypted to different meaningful plaintexts, depending on the key used. This allows the sender to have plausible deniability if he is compelled to give up his encryption key. But in strictly-defined modern cryptography, it is almost impossible to design a ciphertext that can be decrypted to several different meaningful plaintexts. In the literature of steganography, plausible deniability means the capability to deliver some genuine message under the cover of other innocuous messages. When the existence of hidden information is detected and the sender is forced to reveal the secret message, he can simply turn in one innocuous message and claim that no other information is hidden. As an example, the aforementioned real-world steganography system illustrates such a behavior. Plausible deniability has been proposed to enhance the security of steganography systems and defend current steganalysis, as described in [14]. In this paper, instead of diving into details of various plausibly deniable schemes, some high-level discussions about implementing plausibly deniable steganographic systems based on generic watermarking techniques, and the comparisons with the proposed system are provided in Section IV-A.

D. *The "Digital" Invisible Ink*

In this paper, steganography systems based on a "digital" version of invisible ink, denoted as digital invisible ink (DII), are proposed. Since we try to implement a digital version of such invisible-ink system based on existing watermarking schemes, corresponding characteristics of invisible ink listed in Section I-A shall be adequately implemented.

1) Only when the stego work undergoes certain prenegotiated manipulations, hidden messages will be correctly extracted. In subsequent discussions, the prenegotiated manipulations are media processing procedures that always

cause distortions to the stego work. Note that in our digital implementations, the types and degrees of manipulations are carefully controlled and viewed as keys to achieve better security.

2) To extract the genuine secrets, the intended receiver will deliberately and seriously distort the marked work. But for the channel supervisor or non-intended users, the marked work is still perceptually similar to the original cover work.

3) In the case of plausibly deniable steganography, the payloads extracted by the intended receiver will consist of both a cover message and a genuine message. The intended receiver can easily distinguish between the cover message and the genuine message because he can also extract the cover message solely. In some interesting cases, we will show that the cover message can be devised to help interpretation of the genuine messages.

Note that the idea of digital invisible ink data hiding is firstly revealed in [15] and then briefly exploited in [16] by the authors. In this paper, we thoroughly describe the motivation, implementations and experimental results of digital-invisible-ink data hiding. Two major data-hiding schemes, the spread-spectrum watermarking and the quantization-based watermarking, are adopted to implement the digital invisible ink and respectively discussed in Sections II and III. In addition to implementation details and experimental results, applications and inherent limitations of each scheme are also discussed. Section IV gives extensive discussions about plausible deniability and illustrates the superiority of the digital-invisible-ink methodology. Section V concludes this paper and states our future work.

II. Digital Invisible Ink Based on Spread-Spectrum Watermarking (SS-DII)

Spread-spectrum watermarking techniques [17], [18] are well-known data-hiding schemes for all kinds of media. In this section, basic spread-spectrum approaches will be modified to simulate the invisible-ink steganography in the real-world.

A. *Conventional Spread-Spectrum Watermarking*

Spread-spectrum watermarking techniques are correlation-based schemes. The process of embedding a single message bit using spread-spectrum watermarking schemes and then applying some manipulations to the stego work can be formulated as follows:

$$\hat{c} = c' + n = c + a \cdot b \cdot w + n \tag{1}$$

where c is a vector consisting of components in the cover work, c' is the corresponding vector in the stego work, and $\hat{c}$ is the vector in the distorted stego work. Moreover, a is the weighting factor deciding the embedding energy of watermark signals (which is often determined according to perceptual models or heuristic rules). b is the message bit represented as 1 or -1, and w is the predefined watermark vector, often a pseudo-random chip sequence in common spread-spectrum schemes. Finally, n is the additive noise vector caused by malicious attacks or media processing on the stego work.

In order to identify whether a suspected work $\tilde{c}$ has been marked, the correlation value between $\tilde{c}$ and w is calculated.

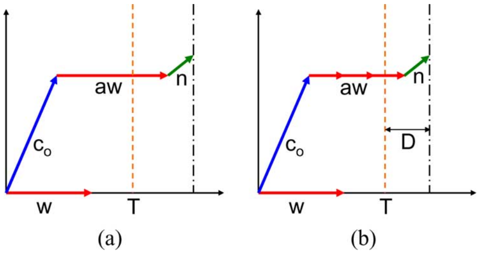


Fig. 2. Geometric models of spread-spectrum watermarking: (a) general case and (b) informed-embedding case.

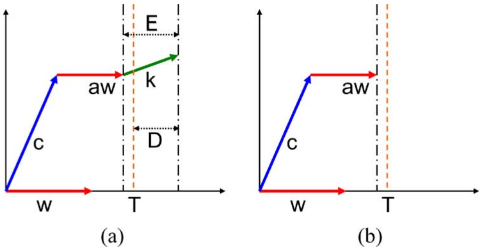


Fig. 3. In a DII data-hiding scheme, the detection result depends on whether the prenegotiate manipulations exist (a) or not (b).

Assume that $\tilde{c}$ is marked and distorted (i.e., $\tilde{c}$ is equal to), the extraction process can be described by

$$w \cdot \tilde{c} = w \cdot c + a \cdot b \cdot w^2 + w \cdot n. \tag{2}$$

Both $w \cdot c$ and $w \cdot n$ are close to zero due to the noise-like characteristic of w. If the correlation value $(w \cdot \tilde{c})$ is larger than a positive threshold value T, $\tilde{c}$ can be regarded as hidden with a message bit of 1 (i.e., $b = 1$). On the contrary, if the correlation value is less than a negative threshold value $-T$, it means that $\tilde{c}$ is carrying a message bit of -1. If we simply choose the value of T to be 0, the message bit can be determined according to whether the correlation value is positive or negative.

Fig. 2(a) shows the geometric model illustrating the prescribed embedding and detection processes. c, n and w are often regarded as vectors in a multidimensional hyperspace. With an adequately normalized w, the obtained correlation value is in fact the projection of $\hat{c}$ along the direction of w.

In an informed-embedding case, i.e., assume the effects of the cover work c and n are known, the weighting factor a can be adjusted to guarantee a successful detection such that:

$$w \cdot c + a \cdot b \cdot w^2 + w \cdot n > T + D \tag{3}$$

where D is a predefined value over the threshold value T. Fig. 2(b) illustrates this scenario.

In general-purpose watermarking applications, exactly grasping all possibilities of n is far from reality. However, it is a totally different story in passive-warden steganography applications where the channel supervisor will not introduce any additional distortions. If both the host-interference caused by c and the distortions (simulated by n) due to the sender-imposed lossy compression are predictable, detection results can be fully controlled. In fact, the proposed DII data-hiding schemes can be regarded as extensions of such an informed-embedding methodology. In the following discussions, an informed-embedding model incorporating with an additional noise vector k representing the effects of prenegotiated manipulations will be introduced. And some constraints on this model will be exploited to facilitate the invisible-ink behavior. Without loss of generality, the noise denoted by n will be omitted in following discussions.

B. Model of SS-DII

In a DII data-hiding scheme, the most essential principle is that the existence of a noise n caused by certain prenegotiated manipulations is indispensable for the successful detection of the message bit, as illustrated in Fig. 3. In Fig. 3(a), the detection result of $b = 1$ is guaranteed by employing the informed-embedding approach similar to (3). The only difference is that, now, the effect of k is considered instead of that of n. If k does not occur in the marked content, as the case shown in Fig. 3(b), the result of $b = 1$ cannot be successfully extracted.

To enable the invisible-ink liked behavior shown in Fig. 3, two conditions must be satisfied. First, the included angle between the noise vector k and the watermark vector w must be within the range of $[90^\circ, -90^\circ]$. In other words, the noise vector k must contribute positively to the correct extraction result. Second, the magnitude that the vector k projects on the direction of w, denoted as E in Fig. 3(a), must be larger than the amount D over the detection threshold T. That is, k must contribute significantly to the correct detection result. DII data-hiding "fails" when the two conditions cannot be satisfied simultaneously.

C. The SS-DII System

To implement a digital-invisible-ink data-hiding system using spread-spectrum watermarking, an iterative informed-embedding approach is proposed. Note that, in our implementation, some prenegotiated manipulations, such as lossy compression or content processing, are incorporated to distort the stego works. Furthermore, a watermark extractor is also included to estimate the effect of prenegotiated manipulations. Fig. 4 shows the flowchart of the proposed system.

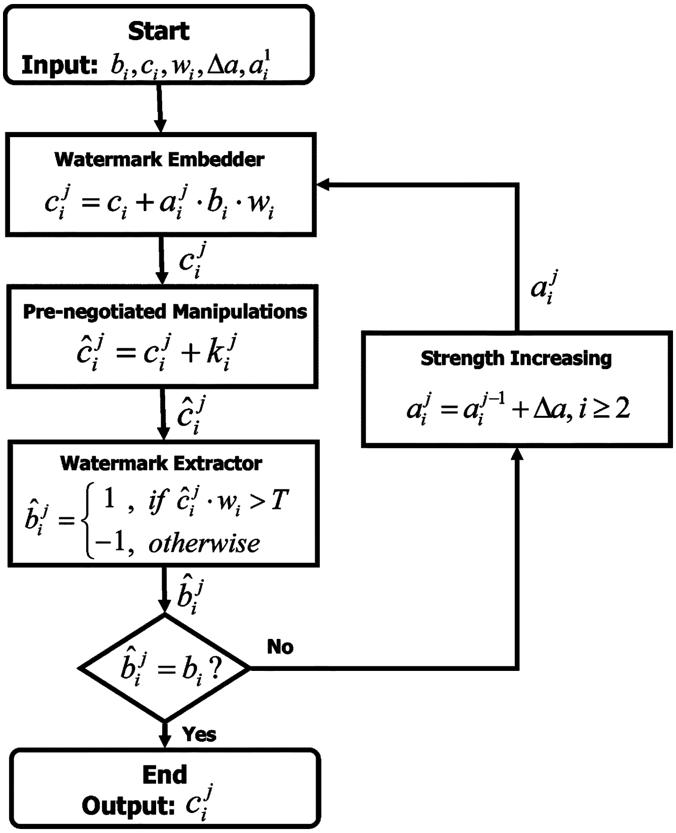

Start
Input: $b_i, c_i, w_i, \Delta a, a_i^1$
Watermark Embedder
$c_i^j = c_i + a_i^j \cdot b_i \cdot w_i$
c_i^j
Pre-negotiated Manipulations
$\hat{c}_i^j = c_i^j + k_i^j$
$\hat{c}_i^j$
Watermark Extractor
$\hat{b}_i^j = \begin{cases} 1, & \text{if } \hat{c}_i^j \cdot w_i > T \\ -1, & \text{otherwise} \end{cases}$
$\hat{b}_i^j$
$\hat{b}_i^j = b_i$?
No
Yes
Strength Increasing
$a_i^j = a_i^{j-1} + \Delta a, i \geq 2$
a_i^j
End
Output: c_i^j

Fig. 4. Flowchart of the iterative informed-embedding system that enables the invisible-ink liked behavior. The flowchart shows the jth iteration for embedding the ith watermark bit.

More specifically, assume we would like to embed a message sequence $\vec{b} = \{b_i | i = 1, \ldots, L\}$ into the cover work. Each message bit b_i is embedded into a vector c_i consisting of N components selected out of the cover work. In the jth iteration of the embedding process, each b_i will be modulated with an N-component chip sequence w_i, scaled according to the weighting factor a_i^j, and then added to c_i to produce the stego work vector c_i^j. After message embedding, prenegotiated manipulations will be applied to the stego work to introduce some distortion k_i^j, and a spread-spectrum watermark extractor is adopted to determine whether the embedded b_i can be successfully extracted when k_i^j has been applied to the stego work. Note that a_i^j is iteratively increased until b_i can successfully resist k_i^j. Though the same prenegotiated manipulations are used throughout the whole iterative process, differences between k_i^p and k_i^q, given $p \neq q$, are nature results due to the difference between c_i^p and c_i^q.

Does this iterative informed-embedding approach meet the two requirements of SS-DII data hiding? For the positive-contribution requirement, since all $w_i s$ are pseudo-randomly distributed, there is about half of the chance that the included angle between k_i and w_i lies within the range of $[90^\circ, -90^\circ]$. This fact implies that, inherently, about half $b_i s$ will never show the intended invisible-ink behavior.

As for the second condition, since this iterative watermarking approach produces weakly embedded works by immediately ceasing the strength increase after successful embedding, the magnitude of correlation value over the detection threshold [denoted as D in Fig. 3(a)] is consequently small. Therefore, as long as the prenegotiated manipulations cause significant distortions along the direction of the watermark vector (i.e., E in Fig. 3(a) is large enough), the second condition can be satisfied. However, since the noise vector k_i^j tends to be near-orthogonal to w_i, the distortions caused by prenegotiated manipulations shall be large to satisfy the significant condition of DII data hiding.

D. Securing Messages With Digital Invisible Ink

To evaluate the effectiveness and feasibility of the SS-DII scheme, we devise an application scenario where the security of hidden messages is protected using SS-DII techniques only, without involving additional security modules like ciphers or passwords. Fig. 5 depicts the flowchart of such a DII-based steganography system. In the receiving end, if the hidden message is extracted directly with the corresponding watermark extractor, only incorrect messages can be obtained. But when prenegotiated manipulations are performed and then extracted

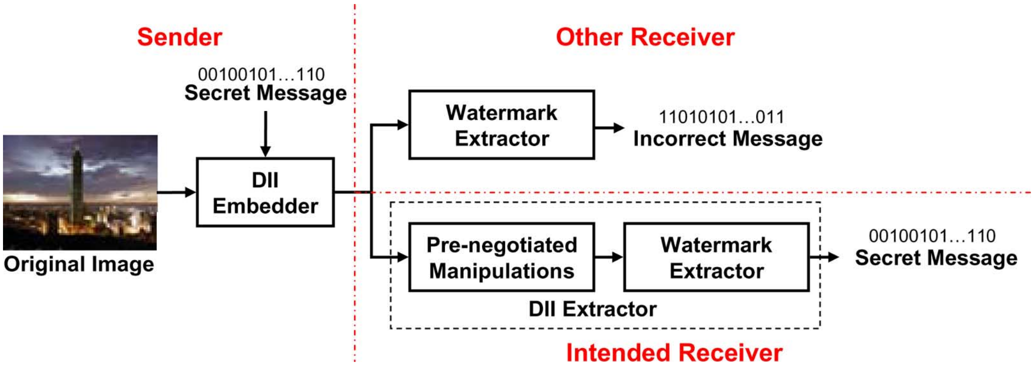


Fig. 5. Securing hidden message using the proposed DII scheme.

TABLE I
PSNR VALUES FOR SS-DII EMBEDDED IMAGES

Images	Lena	Peppers	Fruits	Baboon
PSNR (dB)	36.81	35.21	39.07	37.20

with the corresponding watermark extractor, the hidden secret message can be exactly extracted.

E. Experimental Results

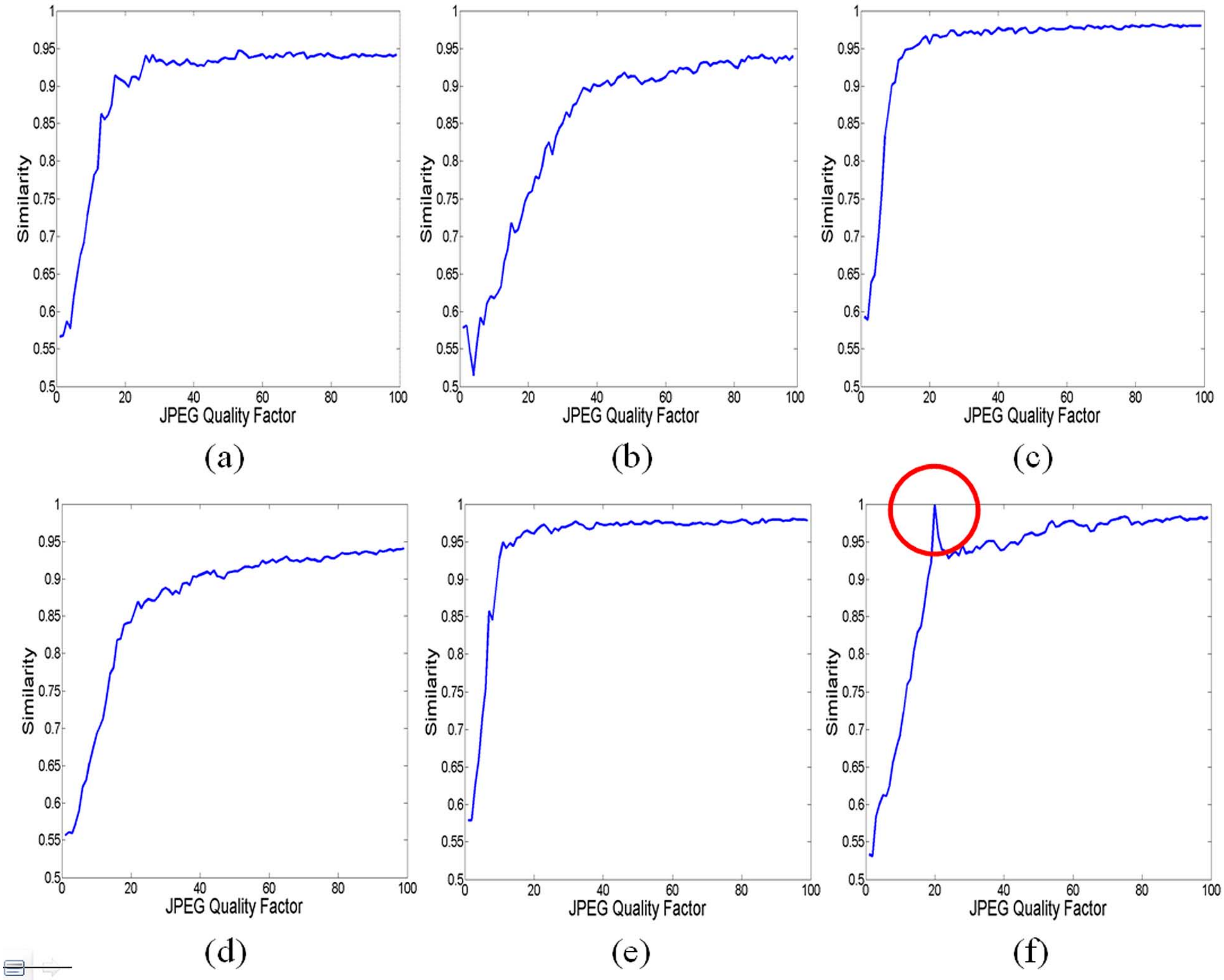
We evaluate the effectiveness of the proposed scheme using a set of common 512 × 512 test images, including Lena, Baboon, Peppers and Fruits. The employed parameter settings are:$L = 1024$, $N = 20$, $\Delta a = 0.5$ and $a^1 = 0.3$. Message bits are sequentially embedded into the global DCT coefficients of images according to the zig-zag order and the starting position is chosen to be the 2000th coefficient to avoid serious quality degradation. Note that in common application scenarios the embedding positions may be decided in a more secure manner. Here we simply adopt the zig-zag permutation for the ease of system implementation. Significant prenegotiated distortions, including histogram equalization, blurring using a 7 × 7 averaging filter and JPEG compression with quality factor of 20, are sequentially employed during message embedding based on the prescribed informed-embedding spread-spectrum scheme.

Table I lists the PSNR values between marked images and corresponding cover images. All marked images show good objective and subjective fidelity performance. Since similar extraction results of all test images are obtained, only extraction results for Lena and Fruits are depicted in Figs. 6 and 7 due to the paper length limit. In Figs. 6 and 7, the horizontal axis represents different quality settings of JPEG compressions being applied to the stego works before extraction; the vertical axis is the similarity measure based on normalized correlation values between the extracted message sequence and the original message sequence. Figs. 6(a) and 7(a) show the extraction results that messages are extracted from stego works after performing different degree of JPEG compression only. Figs. 6(b) and 7(b) are results for the cases where histogram equalization is omitted before performing extraction. Figs. 6(c) and 7(c) are extraction results without performing the blurring operation. Figs. 6(d) and 7(d) are cases where blurring and histogram equalization are performed in reverse order; Figs. 6(e) and 7(e) are cases where JPEG compressions and blurring are performed in reverse order. Figs. 6(f) and 7(f) are extraction results when all the prenegotiated manipulations are performed in correct order before watermark extraction.

According to the experimental results, it is obvious that only when prenegotiated manipulations considered during message embedding are exactly reproduced before extraction will the message sequence be 100% correctly extracted. Note that the manipulations which cause relatively insignificant distortions, like blurring operations in this experiment, are consequently less relevant to whether the extraction results can be correctly extracted, as suggested by Figs. 6(c) and 7(c). Furthermore, the order of prenegotiated manipulations also contributes to the security of the hidden message. For example, as shown in Figs. 6(d) and 7(d), outputs of the histogram equalization operation may be quite different due to slight changes (caused by additional blurring) being performed on the input image.

F. Limitations of SS-DII

Though messages can be fully extracted when prenegotiated manipulations are exactly reproduced in the extraction side, the similarity values between the genuine secret message and the message sequences extracted using other configurations are still high. This is a natural limitation imposed by inherent characteristics of watermark vectors used in spread-spectrum schemes. As we have mentioned, the positive-contribution condition is only partially satisfied due to the pseudo-randomness of the watermark vector. Therefore, the ratio of valid messages bits (i.e., the message bits that can potentially demonstrate the invisible-ink behavior) is reduced to about 50%, and the positions of valid message bits are unpredictable. Moreover, the near-orthogonal characteristic between the watermark vectors and the noise vectors further reduces the probability that an arbitrarily chosen prenegotiated manipulation satisfies the significant-contribution requirement. Therefore, the SS-DII scheme is only suitable for protecting messages that even a slight change will lead to significantly different meanings. For example, nearby indices in a hash table may be mapped to very different meanings and thus satisfy such a constraint. The main advantage of SS-DII scheme over existing security measures (e.g., key-based or encryption/decryption modules) lies in that, except the message extractor, no explicit security modules like password input prompts or deciphering programs are required. Consequently, it is more difficult for the supervisor to reveal the correct hidden



(a) (b) (c)

(d) (e) (f)

Fig. 6. Different extraction results for the Lena image embedded with the proposed SS-DII data-hiding system.

message since the protection is "unsealed" by invoking media processing tools widely available in common computer platforms.

III. Digital Invisible Ink Based on Quantization Watermarking (Q-DII)

Unlike the SS-DII scheme that subjects to inherent limitations, the DII schemes based on quantization watermarking in specific embedding domains can bring the powers of digital invisible ink in fully play. Theoretical models, simulations and experimental results are illustrated to exploit the characteristics, applications and limitations of Q-DII schemes.

A. Model of Q-DII

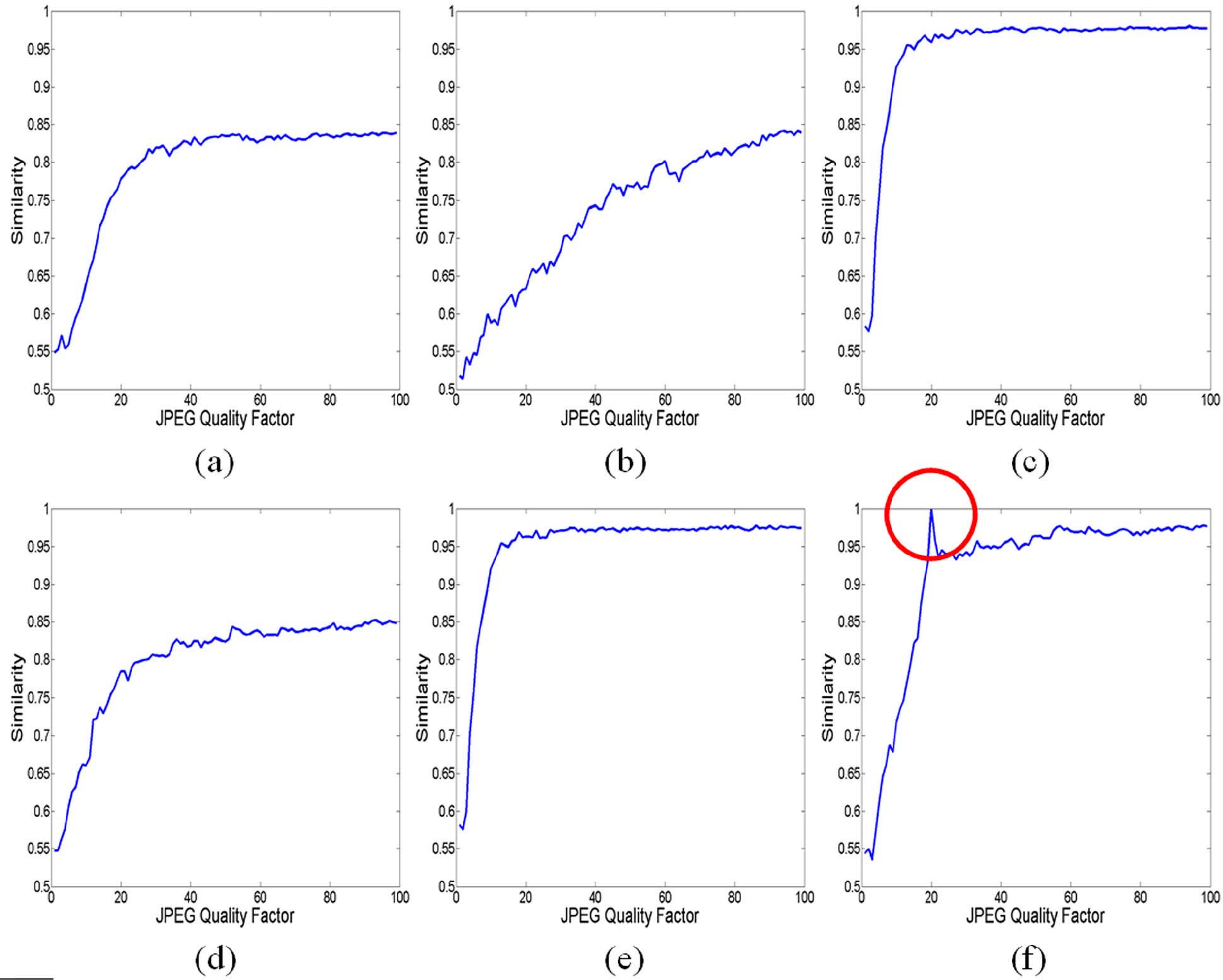
Quantization watermarking [19], [20] is another important class of blind-detection data-hiding schemes. In quantization watermarking methods, payload bits are embedded by quantizing components of the cover work according to some predefined quantizer. As shown in Fig. 8, a chosen component of the cover work, denoted as c, will be quantized to a reconstruction point larger or smaller than the predefined decision threshold T, depending on whether the watermark bit is positive $(b = 1)$ or negative $(b = -1)$.Therefore, the embedding process can be represented as:

$$c' = \begin{cases} T + D_1, & b = 1 \\ T - D_2, & b = -1 \end{cases} \quad (4)$$

where D_1 and D_2 are often determined according to perceptual models or heuristic rules. Note that in many schemes the decision threshold T closely depends on c, e.g., T may be a quantized version of c. During payload extraction, whether an embedded message bit is 1 or -1 can be easily read out by comparing the corresponding component in the marked work with T.

Note that, here, we discuss only the simplest case based on binary quantization for the ease of illustration. However, without loss of generality, the proposed DII principles can be applied to more quantization-based data hiding schemes.

Fig. 9 illustrates the model of DII data hiding schemes based on quantization watermarking. The original watermarking procedures are modified to satisfy the essential principle of DII data hiding—manipulating (distorting) the marked work is indispensable for the successful detection of payload bits. Note that, to prevent redundancies, only the case of $b = -1$ is considered in following discussions. Since the extractor has to output



(a) (b) (c)

(d) (e) (f)

Fig. 7. Different extraction results for the Fruits image embedded with the proposed SS-DII data-hiding system.

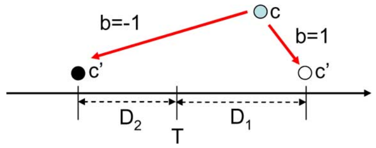


Fig. 8. Quantization watermarking using a simple binary quantizer.

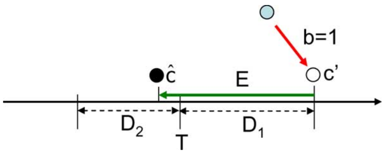


Fig. 9. Illustration of the Q-DII scheme.

an incorrect extraction result (as if $b = 1$) when the prenegotiated manipulation is not performed, c must be deliberately quantized to the wrong reconstruction point c' first. Then, the required manipulation must distort the marked work and direct it along the direction from the wrong reconstruction point to the correct one. The two procedures are helpful to satisfy the positive-contribution requirement of Q-DII scheme. Furthermore, since the manipulated content $\hat{c}$ should indicate the intended extraction result (i.e., $b = -1$), the magnitude of distortion caused by the manipulation, denoted as E in Fig. 9, must be larger than D_1 in order to satisfy the significant contribution condition of Q-DII data hiding.

B. The Q-DII System

To satisfy the positive-contribution condition of Q-DII, two sequential steps must be performed: quantizing the original work to the wrong reconstruction point and then distorting the marked work approaching to the correct direction. The first step is naturally implemented in a DII embedding module. However, the second step implies that the direction caused by the prenegotiated manipulation in the embedding domain must be controllable, which is unattainable in the prescribed SS-DII scheme. Fortunately, for Q-DII, it can be easily achieved by adequately selecting the embedding domain and the type of prenegotiated manipulations. As a trivial case, if message bits are embedded by altering the intensity value of image pixels and extracted by comparing with a decision threshold, adjusting the brightness would be an adequate manipulation for revealing the correct message bit. In another case, experimental results in [21] suggest that common image processing techniques can be classified into two classes according to whether an operation increases or decreases the magnitude of most DCT coefficients. To be more specific, operations like JPEG compression or

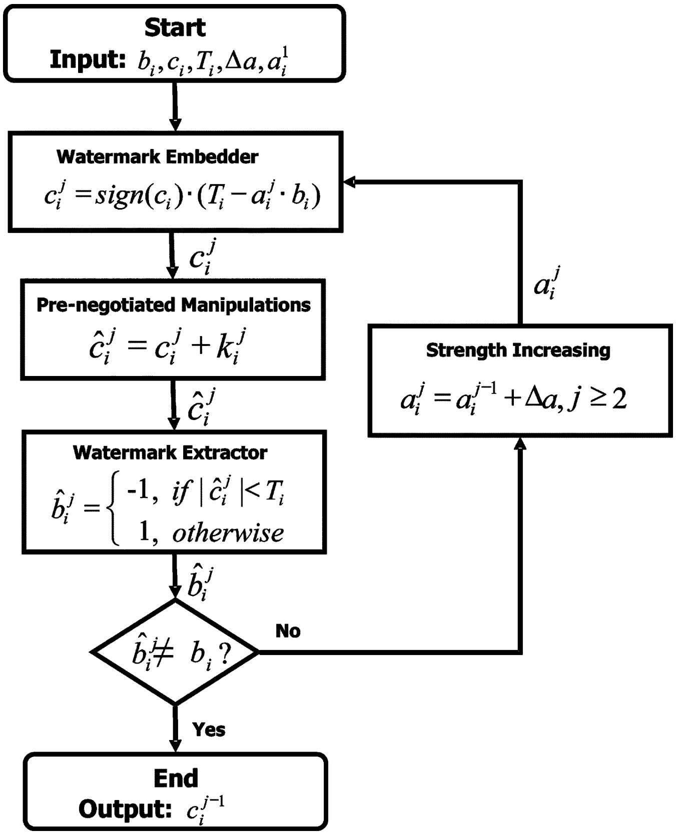

Start
Input: $b_i, c_i, T_i, \Delta a, a_i^1$
Watermark Embedder
$c_i^j = sign(c_i) \cdot (T_i - a_i^j \cdot b_i)$
c_i^j
Pre-negotiated Manipulations
$\hat{c}_i^j = c_i^j + k_i^j$
$\hat{c}_i^j$
Watermark Extractor
$\hat{b}_i^j = \begin{cases} -1, & if\ |\hat{c}_i^j| < T_i \\ 1, & otherwise \end{cases}$
$\hat{b}_i^j$
$\hat{b}_i^j \neq b_i$?
No
Yes
End
Output: c_i^{j-1}
a_i^j
Strength Increasing
$a_i^j = a_i^{j-1} + \Delta a, j \geq 2$

Fig. 10. Iterative-informed embedding process for Q-DII schemes.

blurring probably decrease the magnitude of DCT coefficients while adding noises or edge enhancement increases it. Therefore, the block-DCT coefficients, often adopted in existing watermarking schemes, can be readily utilized to implement Q-DII data hiding.

In subsequent discussions, we implement the Q-DII embedding module based on the frequency-domain watermarking schemes introduced in [22], [23]. Simply speaking, the magnitude of the scaled DC coefficient in every DCT block is chosen to be the decision threshold T, and the magnitude of each chosen DCT coefficients is set as values larger or smaller than T according to individual binary watermark bits. Moreover, JPEG compression is the required prenegotiated manipulation.

Finally, the significant-contribution condition can be easily satisfied by weak embedding, i.e., setting D_1 in Fig. 8 to small values. Since both conditions of the DII model can be fully satisfied, all message bits in a Q-DII scheme will show the invisible-ink liked behavior. Therefore, Q-DII can achieve better performance for securing the hidden message than SS-DII.

Similar to the iterative embedding process in SS-DII scheme, we adopt an informed-embedding process for Q-DII scheme so that the effect of prenegotiated manipulations can be exactly grasped. Fig. 10 shows the flowchart of the iterative process. Note that the manipulated content $\hat{c}_i^j$ should be at the same side of the decision threshold as the reconstruction point representing b_i in the underlying watermarking scheme. During the jth iteration for embedding b_i into c_i, an iteratively increasing watermark energy a_i^j is used to determine the distance between the decision threshold value T_i and the embedded content component c_i^j. Note that although this process seems to be different from the one shown in the SS-DII case, effects of the two processes are essentially the same. Simply speaking, a piece of content marked with adequate strength (so that the hidden message bits can be extracted exactly after the prenegotiated manipulations are performed) will be generated. During the iterations, c_i is deliberately quantized to values far and far away from T_i along the opposite direction as mentioned in Section III-A. As long as the extracted message bit $\hat{b}_i^j$ becomes different from b_i, the iteration process terminates immediately and then the marked work produced using the most adequate watermark strength a_i^{j-1} is obtained.

C. Covering Secret Messages With Insensitive Watermarks

Instead of implementing the same application of SS-DII, a plausibly deniable steganography scheme where secret messages are covered by insensitive watermarks is illustrated, as shown in Fig. 11. A short introduction to the plausibly deniability of steganography system has been presented in Section I-C. Note that the transmission of digital contents is monitored by a supervisor who has the right to cease any doubtful transmission. Moreover, using common watermarking schemes to facilitate insensitive applications, such as delivering copyright information or persistent metadata association, is allowed in this scenario. The supervisor can access the announced watermark extractor and may adopt additional statistic steganalysis tools to determine whether a piece of transmitted data shall be intercepted. In cases where the supervisor does not employ any steganalysis modules, the camouflage provided by the cover watermark can completely shadow the delivery of true secrets. As for the case where statistical steganalysis tools are adopted, we will show that DII-based scheme provides better secrecy against statistical steganalysis (for competing with systems hiding multiple watermarks) and eliminates the need to introduce additional extraction module (when compared with systems using more-than-one watermarking schemes).

To implement the steganography system shown in Fig. 11, binary watermark images are used to visualize the effect of Q-DII data hiding. Let both the cover message and the secret message be represented as black pixels in the binary image. Assume we denote all the pixels in the fully extracted watermark image as U, the set of black pixels comprising the cover message as B_C, and the set consisting black pixels of the secret message as B_S. The remaining white pixels is therefore represented as $U - B_C - B_S$. In our plausibly deniable steganographic system, B_S and B_C are designed to be non-overlapped. And the overall watermarking capacity is shared by the cover message and the secret message. Since semantic meaningful cover messages always leave blank areas, secret messages can always find their places. Later, we will demonstrate that, in some interesting message designs, B_C is even helpful for interpreting the semantics of B_S.

Since the secret message appears only when prenegotiated manipulations are performed before watermark extraction, the DII embedding methodology is employed only for watermark bits belonging to B_S. Other watermark bits are embedded using the underlying quantization watermarking scheme. Fig. 12

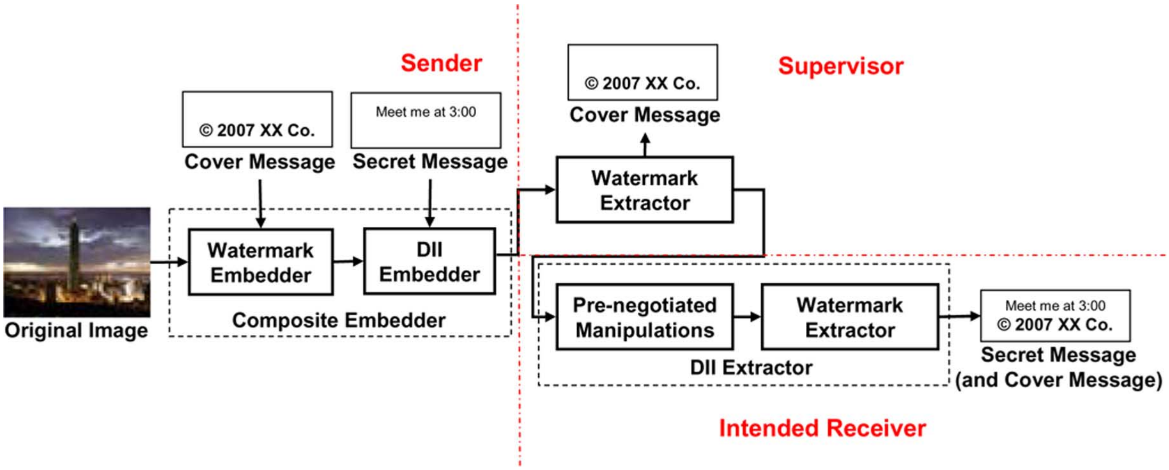


Fig. 11. DII-based plausibly deniable steganography system.

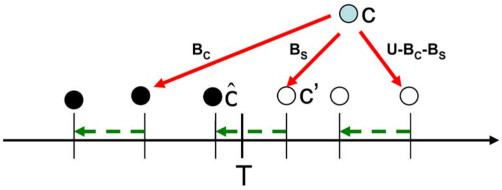


Fig. 12. Composite embedder of the plausibly deniable steganography system.

shows the model of the composite embedder. It is worth noting that the reconstructing points for watermark bits belonging to $U - B_C - B_S$ might be alternatively set as $c' + \Delta$, where c' for these watermark bits are also estimated using the procedure shown in Fig. 10 (as if the watermark bits are within B_S), and is a small value (empirically set as 4 in our experiments) rather than a significant amount in underlying robust watermarking schemes. The purpose of this extension is to eliminate the possibility that a non-intended user casually performs a severe manipulation (resulting in more degradations than the prenegotiated one) and accidentally discovers the hidden secret message. Since the watermarking strengths for the white pixels is only slightly larger than those for secret message bits, the extracted watermark image will comprise black pixels only when a user applies a severe attack stronger than the prenegotiated one.

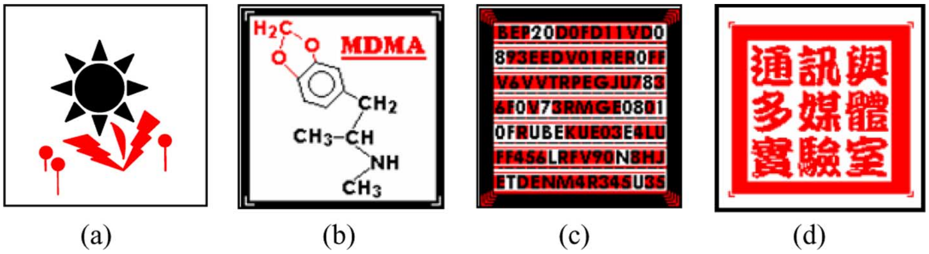
D. *Relationships Between the Cover Message and the Secret Message*

Since both the cover message and the secret message are represented by pixels within the same watermark image, various feasible watermark designs can be devised. Fig. 13 shows some sample watermark images. To clearly visualize the watermark designs, cover messages are represented as black pixels in each figure and the genuine one are shown in red (but both messages are viewed as black patterns in implementations). In Fig. 13(a), the semantic meaning of the watermark image (originally a sun pattern) changes (now a flower) after the genuine message is revealed. Fig. 13(b) illustrates the danger that an illegal message (the formula of MDMA, an addictive drug) may be disguised as an insensitive message (the formula of Methamphetamine, a valueless chemical compound). Fig. 13(c) demonstrates that the extraction of genuine message may eliminate unnecessary information so that a meaningful message (indicating a time/location pair in this case) can be revealed out of seemingly random patterns. To the extreme, the cover watermark may be a null pattern and all information capacity is reserved for the genuine watermark, such as the one shown in Fig. 13(d).

E. *Experimental Results*

Figs. 14 and 15 show the extraction results for cases that the 128 × 128 binary watermark images, shown in Figs. 13(c) and 13(d), are respectively embedded into the 512 × 512 Lena image. The PSNR values for the two marked works are 36.00 and 35.26, respectively. Experimental results using other test images show similar results and are therefore omitted here. According to the dimension ratio between watermarks and cover works, four watermark bits will be embedded into each 8 × 8 DCT block by altering predefined AC coefficients. Without loss of generality, AC coefficients located at the (5, 2), (4, 3), (3, 4) and (2, 5) coordinates within each DCT block are selected as places to embed messages. Both visual and statistical extraction results obtained from different versions of the marked images created by applying different degrees of JPEG compression (quality factors ranging from 99 to 1) are illustrated. As expected, the messages indicating the time/location information and the seal pattern (denoting the Communications and Multimedia Laboratory in traditional Chinese) are not visible until the manipulations very close to the prenegotiated one are performed. For manipulations moderately different from the prenegotiated one, either the cover message or a fully black pattern is extracted.

Note that though the prenegotiated attack is assumed to be the JPEG compression with quality setting of 40, the best extraction result occurs when JPEG compression with quality factor 35 is applied to the marked work. This is in fact the consequence caused by the undesired dependencies among chosen DCT coefficients within each DCT block. Though, theoretically, coefficients within each DCT block shall be independent to each other, altering a DCT coefficient might still change other coefficients. Fig. 16 shows the statistical extraction result using a



(a) (b) (c) (d)

Fig. 13. Watermark patterns consisting of both cover messages and genuine messages.

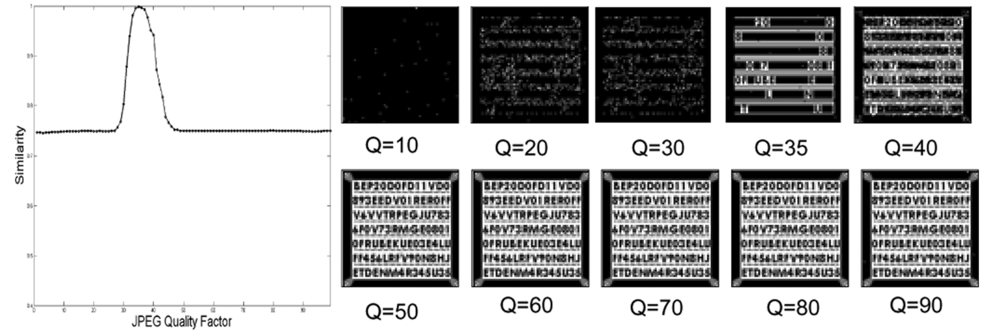


Fig. 14. Messages extracted from the Lena image marked with the watermark image shown in Fig. 13(c).

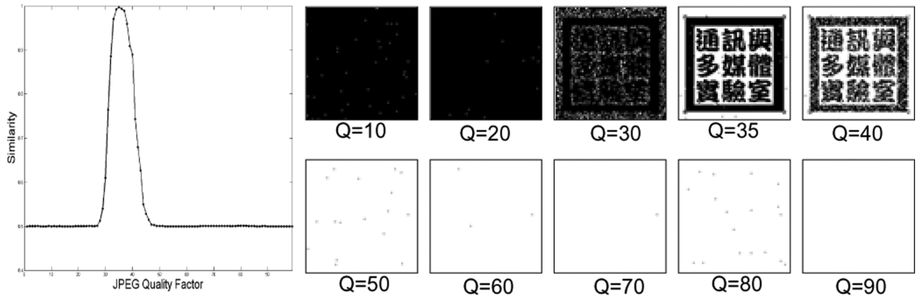


Fig. 15. Messages extracted from the Lena image marked with the watermark image shown in Fig. 13(d).

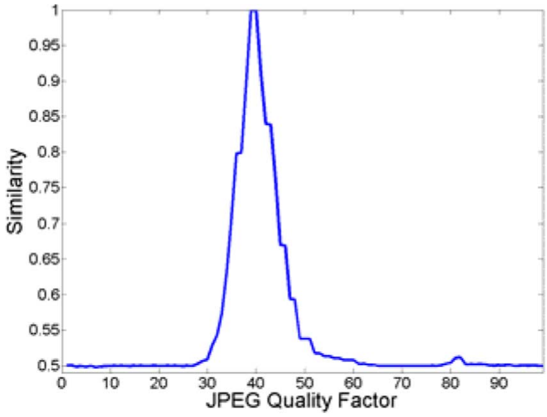


Fig. 16. Statistical extraction result when embedded components are completely independent.

64×64 watermark image where half of the consisting pixels are set as secret messages. Other experimental configurations are left unchanged. Now, in each DCT block of the original image, only single message bit is embedded. Obviously, the secret message can be perfectly visualized only when the prenegotiated manipulation (JPEG compressing with quality factor 40) is performed before extraction.

Moreover, though the secret messages are hidden in a weak fashion, it is capable of resisting limited noises introduced by an active warden who try to prevent secret communication using naïve steganography skills. Note that it is impractical for the warden to apply severe distortions to all incoming contents since performance degradation of communication channels may lead to serious inconvenience of innocent channel users. Fig. 17 shows more extraction results of the proposed Q-DII scheme. In these experiments, before performing the prenegotiated attack, several typical attacks are respectively issued by the warden, including adding of Gaussian noises $(\mathrm{SNR} = 25)$, moderate JPEG compression $(\mathrm{Quality} = 90)$, as well as Gaussian blurring (using a 3×3 filter). According to the results, the proposed scheme can survive the first two attacks imposed by the supervisor. However, when incorporating additional blurring operation, the extraction result is apparently worse. This phenomenon can be justified as follows. The proposed Q-DII scheme is a variant of the frequency-domain quantization watermarking scheme. In addition to the desired invisible-ink like behavior, its robustness is also significantly affected by

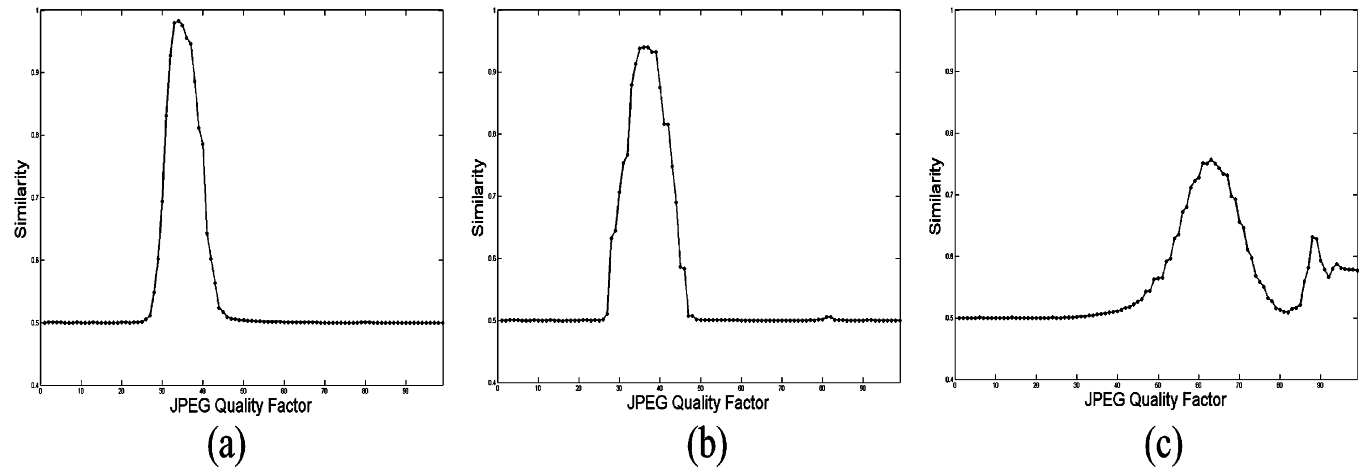


Fig. 17. Extraction results with additional warden-introduced attacks, including (a) Gaussian noises, (b) high-quality JPEG compression and (c) Gaussian blurring.

the inherent characteristics of the underlying watermarking approach. Since the underlying scheme is much more vulnerable to filtering operations like blurring, the message extracted from the slightly blurred image is inherently worse than those from images undergoing other mild operations.

From the viewpoints of channel supervisor in the active-warden scenario, this limitation gives useful clues for determining adequate modifications that can cease illegal communications. From the viewpoints of communication participants, this robustness deficiency can be alleviated by applying DII principles to other watermarking schemes with different robustness characteristics. Fortunately, for common multimedia communication channels like online-albums, introducing quality-degradation operations like blurring to all incoming contents solely for the purpose of eliminating illegal communications is infeasible and intolerable. Therefore, the feasibility of the proposed DII-methodology is not significantly lowered.

F. Limitations of Q-DII

The Q-DII scheme surpasses the SS-DII scheme in that both the positive-contribution and the significant-contribution conditions can be fully satisfied. Therefore, it can be readily applied to far more applications than SS-DII can. However, one of the major limitations of Q-DII scheme is its heavy dependency on the adequately chosen embedding domain, e.g., watermarking by altering and comparing DCT coefficients or pixel values. For embedding domains where controllable prenegotiated manipulations cannot be easily found, such as the pseudo-randomly projected domain used in [15], the feasibility of Q-DII schemes is reduced. In such schemes, the Q-DII approach suffers from similar performance penalties to that of the SS-DII scheme.

The iterative-informed embedding procedure precisely grasps the effect of prenegotiated manipulations and leads to better secrecy against a brute-force steganalysis. However, it also leads to higher computational complexity of the embedding module. Fortunately, the extractor of the proposed scheme is simply a normal extractor of the underlying watermarking scheme. Therefore, for most watermarking applications, the widespread watermark extractors will never suffer from a long-time extraction or have additional deployment costs.

IV. Discussions

A. Comparisons With Existing Plausibly Deniable Steganography Approaches

Various implementations of steganography systems can be used to enable plausible deniability. Roughly speaking, in addition to the proposed DII schemes, plausibly deniable schemes can be classified into three types: (A) hiding multiple watermarks with single watermarking scheme in which each watermark is associated with a different embedding key, (B) hiding more-than-one watermarks and each has its own watermarking scheme, and (C) hiding an additional watermark in allowed watermarks. Each class has its pros and cons. For (B), additional watermark extractors other than the allowed/announced one need to be distributed and deployed in the receiving end, leading to higher risk of being discovered by the channel supervisor. For (C), since most existing watermarks are of simple encoding formats (such as binary images) and limited capacity, hiding additional watermarks that carry useful information in the original watermarks and preserving the fidelity of "watermarked watermarks" simultaneously are quite difficult. In the proposed DII scheme, no additional watermarking-related module, except the announced watermark extractor, is required. Furthermore, the indispensable prenegotiated manipulations can be provided by common content processing tools with an innocuous looking. In addition, the fidelity of the original watermark extracted under normal configurations is not affected at all. Therefore, the DII scheme apparently surpasses plausibly deniable schemes of type (B) or (C).

When comparing the DII schemes with approaches classified as class (A)—multiple watermarks are embedded using a single watermarking scheme with different embedding keys, we illustrate the superiority of DII schemes by testing the secrecy against some existing statistical steganalysis tools. Since Q-DII scheme embeds watermarks in the block-DCT domain, the steganalysis scheme based on JPEG compatibility introduced

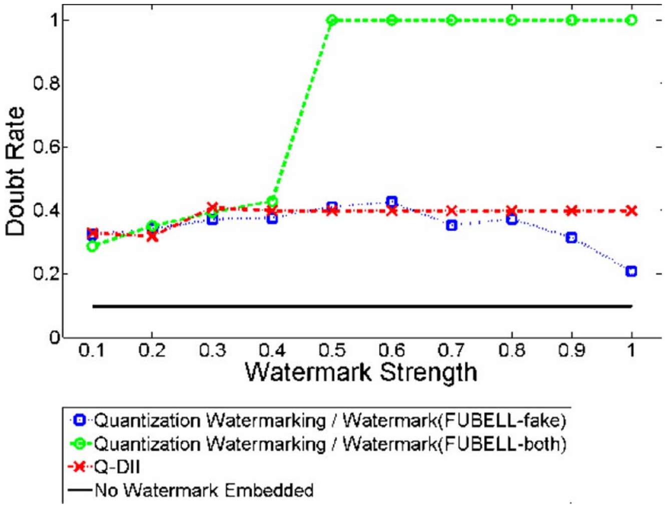


Fig. 18. Statistical steganalysis results illustrating the superiority of the DII scheme over the approach that simply hides more messages with one scheme (denoted as FUBELL-both).

by [24] is adopted. All the secret messages are embedded into JPEG-compressed images and then the marked images are saved in loseless PNG format in order to meet the requirement of the adopted steganalysis system. Steganalysis results are obtained by testing the images marked using the underlying quantization watermarking scheme with a watermark image consisting of the cover message only, the image marked using the proposed DII scheme with the watermark image containing both the cover message and the secret message, as well as the image marked using the underlying quantization watermarking scheme with two watermark images containing the cover message and the secret message, respectively. Fig. 18 lists the experimental results using the watermark image shown in Fig. 13(c). Clearly, the DII-based scheme provides plausible detection results as if only the cover message is embedded, while hiding multiple watermarks leads to more uncertain detection values when feasible watermark strength is employed. This experimental result is not surprising, since it is just a nature outcome of a rule-of-thumb in steganalysis: the higher the degree of alternation is, the more statistical traces it will leave in the marked work. Therefore, the DII-based scheme can provide better plausible deniability than existing deniable steganographic approaches.

B. Potential Security Holes

The fact that current statistical steganalysis schemes cannot differentiate the DII scheme from the underlying watermarking scheme hiding cover message only implies a potential danger caused by insensitive watermarking applications. Well-designed plausibly deniable steganography schemes can take cover of legal watermarking applications and cheat steganalysis schemes as if only insensitive watermark messages are hidden. For example, danger messages related to illegal activity or homeland-security threats may be delivered via legal data-hiding applications like copyright indication or metadata association. Therefore, steganalysis specifically targeting on plausibly deniable schemes shall be developed. Otherwise, general watermarking applications cannot be widely accepted without questioning about security.

C. Security Concerns With Known Algorithm Details

Experimental results in Section IV-A demonstrate the security of DII schemes against statistical steganalysis without any additional information. However, a marked work generated using the DII algorithm will never be exactly the same as the one generated using the underlying watermarking scheme with the cover watermark solely. Therefore, there might be doubts about the security of DII schemes when the algorithm details of underlying watermarking scheme are announced, i.e., when the chosen-stego attack is performed. In fact, whether the security of all watermarking schemes can be kept when algorithm details are revealed is always a controversial issue among watermarking researchers. According to discussions and conclusions obtained in [25], revealed algorithm details at least lead to improved efficiency of watermark removal. Therefore, it seems reasonable for users of practical watermarking applications kept some information as secret keys, e.g., the embedding positions or the applied watermark strength, so that the security of messages can be guaranteed. Though watermarking schemes illustrated in this paper are chosen to be simple for the ease of illustration, most existing watermarking schemes determine the embedding strength based on sophisticated perceptual models or predefined rules. Therefore, it would be more difficult to differentiate whether a work is marked by the DII scheme or by a common watermarking scheme even only a small amount of necessary information is kept.

V. Conclusions and Future Works

Models and implementations, as well as applications, of invisible ink-liked data hiding are illustrated in this paper. Secret messages can be embedded without introducing additional alternation to the original work and no additional sensitive module deployment in the receiving end is required. The proposed digital-invisible-ink scheme can provide better secrecy by enabling plausible deniability. Furthermore, the adopted statistical steganalysis scheme cannot differentiate works marked using the DII scheme from the one produced using common watermarking schemes with the cover message only. Potential dangers of delivering secrets in cover of legal watermarking applications are also pointed out. Currently, we are exploiting the capacity issues of digital-invisible-ink schemes and try to achieve better robustness against geometric and filtering attacks in the active-warden scenario. Moreover, we also try to devise schemes that can provide better secrecy.

Acknowledgment

The authors gratefully acknowledge the precious comments from the associate editor, Dr. M. Kankanhalli, and the anonymous reviewers. And the authors also thank for the revision suggestions provided by Dr. Y.-T. Lin of National Taiwan Normal University.

References

[1] F. L. Bauer, *Decrypted Secrets: Methods and Maxims of Cryptology*, 2nd ed. Berlin, Germany: Springer, 2000, ch. 1.

[2] D. Rigden*, SOE Syllabus: Lessons in Ungentlemanly Warfare, World War II*. East Sussex, U.K.: Gardners, 2004.

[3] S. Katzenbeisser and F..A. P. Petitcolas, Eds*., Information Hiding Techniques for Steganography and Digital Watermarking*. Norwell, MA: Artech House, 2000.

[4] I. J. Cox, M. L. Miller, and J. A. Bloom*, Digital Watermarking*. San Mateo, CA: Morgan Kaufmann, 2002.

[5] M. Wu and B. Liu*, Multimedia Data Hiding*. New York: Springer-Verlag, 2003.

[6] C. S. Lu*, Multimedia Security: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property*. Hershey, PA: Idea Group Publishing, 2004.

[7] B. Furht and D. Kirovski*, Multimedia Security Handbook, Part III and IV*. Boca Raton, FL: CRC, 2005.

[8] N. F. Johnson and S. Jajodia, "Steganalysis: The investigation of hidden information," in *IEEE Information Technology Conference*, New York, Sep. 1998.

[9] N. F. Johnson and S. Jajodia, "Steganalysis of images created using current steganography software," *Lecture Notes in Computer Science*, vol. 1525, 1998.

[10] E. T. Lin and E. J. Delp, "A review of data hiding in digital images," in *Proc. Image Processing, Image Quality, Image Capture Systems Conf., PICS '99*, Apr. 25–28, 1999.

[11] J. Fridrich and M. Goljan, "Practical steganalysis of digital images—State of the art," in *Security and Watermarking of Multimedia Contents*, 2002, vol. SPIE-4675.

[12] G. J. Simmons, "The prisoner's problem and the subliminal channel," in *Advances in Cryptography: Proceedings of CRYPTO '83.*.

[13] S. Craver, "On public-key steganography in the presence of an active warden," in *Proc. IH' 98*, Portland, OR, Apr. 1998.

[14] N. Provos, "Defending against statistical steganalysis," in *10th USENIX Security Symp.*, Washington, DC, USA, 2001.

[15] C. H. Huang, Y. F. Kuo, and J. L. Wu, "Digital invisible ink: Revealing true secrets via attacking," in *Fast Abstract Session, ACM Symp. Information, Computer and Communication Security (ASIACCS '06)*, Taipei, Taiwan, R.O.C., Mar. 2006.

[16] C. H. Huang, S. C. Chuang, and J. L. Wu, "Digital invisible ink and its applications in steganography," in *ACM Multimedia and Security Workshop 2006 (ACM MM&Sec '06)*, Geneva, Switzerland, Sep. 2006.

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

[18] H. S. Marvar and A. F. Florencio, "Improved spread spectrum: A new modulation technique for robust watermarking," *IEEE Trans. Signal Process.*, vol. 51, no. 4, pp. 898–905, Apr. 2003.

[19] M. D. Swanson, B. Zhu, and A. H. Tewfik, "Data hiding for video-in-video," in *Proc. Int. Conf. Image Processing*, 1997.

[20] B. Chen and F. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[21] C. S. Lu, S. K. Huang, C. J. Sze, and H. Y. Liao, "Cocktail watermarking for digital image protection," *IEEE Trans. Multimedia*, vol. 2, no. 4, pp. 209–224, Dec. 2000.

[22] 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.

[23] C. H. Huang and J. L. Wu, "A blind watermarking algorithm with semantic meaningful watermarks," in *ACSSC 2000*, Pacific Grove, CA, Oct. 2000.

[24] J. Fridrich, M. Goljan, and R. Du, "Steganalysis based on JPEG compatibility," in *SPIE MMSA*, 2001.

[25] "Is knowledge of the watermarking algorithm useful for watermark removal?," in *Challenge I in the 2nd Wavila Challenges (WaCha '06)*, Geneva, Switzerland, 2006.

Chun-Hsiang Huang received the B.S., M.S., and Ph.D. degrees in computer science and information engineering from National Taiwan University (NTU), Taipei, Taiwan, R.O.C., in 1997, 1999 and 2004, respectively.

He is now a post-doctoral research fellow in the Department of Computer Science and Information Engineering, NTU. His current research interests include information hiding, multimedia security and content analysis.

Shang-Chih Chuang received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taiprei, Taiwan, R.O.C., in 2005 and 2007, respectively. His current research interests include video compression, image processing, stereo imaging, multimedia application, and multimedia security.

Ja-Ling Wu (SM'98–F'08) received the Ph.D. degree in electrical engineering from Tatung Institute of Technology, Taipei, Taiwan, R.O.C., in 1986.

From 1986 to 1987, he was an Associate Professor of the Electrical Engineering Department, Tatung Institute of Technology. In 1987, he moved to the Department of Computer Science and Information Engineering (CSIE), National Taiwan University (NTU), Taipei, where he is presently a Professor. From 1996 to 1998, he was assigned to be the first Head of the CSIE Department, National Chi Nan University, Puli, Taiwan. During his sabbatical leave (from 1998 to 1999), he was invited to be the Chief Technology Officer of the Cyberlink Corp. In this one-year term, he involved with the developments of some well-known audio-video softwares, such as the PowerDVD. Since August 2004, he was appointed to head the Graduate Institute of Networking and Multimedia, NTU. He has published more than 200 technique and conference papers. His research interests include digital signal processing, image and video compression, digital content analysis, multimedia systems, digital watermarking, and digital right management systems.

Prof. Wu was the recipient of the Outstanding Young Medal of the Republic of China in 1987 and the Outstanding Research Award three times of the National Science Council, Republic of China, in 1998, 2000 and 2004, respectively. In 2001, his paper "Hidden Digital Watermark in Images" (co-authored with Prof. Chiou-Ting Hsu), published in IEEE TRANSACTIONS ON IMAGE PROCESSING, was selected to be one of the winners of the "Honoring Excellence in Taiwanese Research Award", offered by ISI Thomson Scientific. Moreover, his paper "Tiling Slideshow" (co-authored with his students) won the Best Full Technical Paper Award at ACM Multimedia 2006. he was selected to be one of the lifetime Distinguished Professors of NTU, November 2006. He was elected IEEE Fellow in January 2008 for his contributions to image and video analysis, coding, digital watermarking, and rights management.