A novel fragile watermarking scheme for image tamper detection and recovery

Shaomin Zhu (朱少敏)[1*] and Jianming Liu (刘建明)[2]

[1]*China Electric Power Research Institute, Beijing 100192, China*
[2]*State Grid Information & Telecommunication Co., Ltd., Beijing 100761, China*
[*]*E-mail: zhushaomin8888@sina.com.cn*

Received November 13, 2009

We propose a fragile watermarking scheme capable of image tamper detection and recovery with a blockwise dependency mechanism. Initially, the image is divided into blocks with size of 2×2 in order to improve image tamper localization precision. By combining image local properties with human visual system, authentication data are acquired. By computing the class membership degree of each image block property, data are generated by applying k-mean clustering technique to cluster all image blocks. The recovery data are composed of average intensity obtained by truncating the two least significant bits (LSBs) of each pixel within each block. Finally, the logistic chaotic encrypted feature watermark consisting of 2-bit authentication data and 6-bit recovery data of image block is embedded into the two LSBs of each pixel within its corresponding mapping block. Experimental results show that the proposed algorithm does not only achieve superior tamper detection and locate tiny tampered positions in images accurately, it also recovers tampered regions effectively.

OCIS codes: 100.2000, 100.2960, 100.5010.
doi: 10.3788/COL20100807.0661.

With the rapid development in multimedia and network technologies, digital multimedia, such as image, audio, video, and three-dimensional (3D) models, are easily acquired, transmitted, stored, and exchanged. Various powerful multimedia processing tools are capable of modifying and forging multimedia data with much ease and with effects that can hardly be noticed by human eyes. Fragile watermarking has emerged as a leading technique in solving problems related to the verification of the integrity and authenticity of digital multimedia[1].

Fragile watermarking schemes for image tamper detection and recovery have earned great attention because these algorithms can locate and detect image regions that have been tampered with, as well as recover the tampered properties of the image. For example, Zheng *et al.* proposed an auto-restorable fragile watermarking algorithm[2]. It used parities of cross-interleaved Reed-Solomon (RS) block-based codes as watermarks, embedding them into the lowest bits of the image. These cross-interleaved-RS codes were used to detect watermark, identify modified regions, and restore altered data. However, the tamper localization precision was low. Lin *et al.* proposed a hierarchical watermarking method for image content authentication, tamper localization, and recovery[3]. Authentication watermark was made up of parity check code, while recovery watermark was made up of average intensity information of Tour automorphism mapping block. In order to resist the collage and vector quantization (VQ) attacks, the method by Lin used the four-level hierarchical inspections, but the computational complexity was greater and the performance was inefficient. Chang *et al.* pointed out that Lin's method was vulnerable to four-scanning attack and had security problems[4]. Wang *et al.* considered all authentication data embedded in an image and utilized majority-voting technique to determine the legitimacy of image blocks[5]. However, the capacity of watermark embedding was reduced, and the accuracy of tamper localization was decreased. Zhang *et al.* proposed a fragile watermarking scheme capable of recovering the image from its tampered version[6]. A tailor-made watermark consisting of reference-bits and check-bits was embedded into the host image using the lossless data hiding method. However, the quality of watermarked image was lower than 30 dB; hence, the scheme did not achieve satisfactory imperceptibility. He *et al.* proposed a self-recovery fragile watermarking that incorporates block-neighborhood characterization for tamper detection and recovery[7]. However, the accuracy of image tamper localization required further improvements.

Hence, we propose an effective fragile watermarking scheme with image content restoration capability. The host image is segmented into small blocks as carrier of the feature watermark. The authentication data of each block is generated by exploiting the k-mean clustering and the local visual features of the image block. Similarly, recovery data are generated in order to recover the tampered region. The logistic chaotic encrypted feature watermark consisting of authentication data and recovery data of an image block is embedded into its mapping block. The proposed algorithm consists of three stages: feature watermark embedding, block-clustering tamper detection, and block-mapping tamper recovery. Experimental results show that the proposed scheme does not only have good visual quality, regardless of the nature of selected images, it also has superior capability in terms of locating tampered image regions, as well as in the detection and recovery of altered data placed under various malicious attacks.

Data clustering analysis is one of the most fundamental activities for grouping objects based on the measure of similarity, such that the similarities between objects in the same group are high, while similarities between objects in different groups are low. The k-mean clustering technique, one of the simplest unsupervised clustering algorithms, is frequently used in a number of applications, such as pattern recognition[8], data discovery and mining[9], and image hiding[10].

In this letter, we use $X = \{x_1, x_2, \cdots, x_N\}$ as a set of N-dimensional given points clustered into a set of cluster centers, $C = \{c_1, c_2, \cdots, c_K\}$. The basic idea of k-mean clustering is to partition N data points into K disjoint clusters c_j $(1 \leq j \leq K)$ until a stopping criterion is satisfied. The sum of the squared Euclidean distances between data points and the cluster center of the subset is widely used as clustering criterion. The fitness function is defined as[10]

$$F = \sum_{j=1}^{K} \sum_{i=1}^{N} \left\| x_i^{(j)} - c_j \right\|^2, \tag{1}$$

where $x_i^{(j)}$ represents data points that have been classified into the jth group, $\|\cdot\|$ is the Euclidean norm, and $\|x_i^{(j)} - c_j\|^2$ is calculated for measuring the distance between the ith data point of data samples $x_i^{(j)}$ and the jth cluster center c_j.

We combined the high sensitivity and low cross-correlation characteristics of the logistic chaotic map and the two-dimensional (2D) Tour automorphism in order to further extend and improve the security and effectiveness of the system against series of malicious attacks, such as the VQ attack, collage attack, and four-scanning attack. The embedding procedure was conducted as follows.

Step 1: We used I as a grayscale host image with a size of $M \times N$. The given image was divided into non-overlapping blocks I_i with a size of 2×2 pixels expressed as

$$I_i = \begin{bmatrix} i_{i0} & i_{i1} \\ i_{i2} & i_{i3} \end{bmatrix} (i = 1, \cdots, D) \left(D = \frac{M \times N}{4} \right). \tag{2}$$

Step 2: In order to implement recovery mechanism, a block mapping sequence, $A \rightarrow B \rightarrow C \rightarrow D \cdots \rightarrow A$, was needed. Each symbol denotes a non-overlapping block. The intensity feature of block I_A was embedded into block I_B, and the intensity feature of block I_B was embedded into block I_C, etc. Lin *et al.* used one-dimensional (1D) Tour automorphism to obtain a one-to-one mapping sequence[3]. The one-to-one block-mapping sequence for $(I_i, I_{i''})$ $(I_i = I_{XY}, I_{i'} = I_{X'Y'}\ i, i' = [1, D])$ was generated by the row secret key, $k1$, and column secret key, $k2$, to enlarge the 1D Tour automorphism key space:

$$X' = \left[f(X) = (k1 \times X) \mathrm{mod} \frac{M}{2} \right] + 1, \tag{3}$$

$$Y' = \left[f(Y) = (k2 \times Y) \mathrm{mod} \frac{N}{2} \right] + 1, \tag{4}$$

where X and X' are row block numbers, Y and Y' are column block numbers.

Step 3: The 8-bit feature watermark, denoted as F_i, consisted of 6-bit recovery data, $rd_i = \{f_7, f_6, f_5, f_4, f_3, f_2\}$, and 2-bit authentication data, $ad_i = \{f_1, f_0\}$. Recovery data were used to restore tampered blocks, while authentication data were used to identify any modifications made on the authentic image.

Step 3.1: To generate and embed the feature watermark, the two LSBs of all pixels within each block I_i were first set to zero.

Step 3.2: Human visual system (HVS) is one of the most complex biological systems; it includes the three stages of encoding, representation, and interpretation. Many factors cause human vision to have limited sensitivity. For example, the optical lens has chromatic aberration effects, and the mosaic of photoreceptors only employs the spatial sampling process[11]. In a natural image, HVS is more sensitive to noise in smooth area than in textured area, and is more sensitive to low luminance than to high luminance. Human vision perceptive redundancy allows us to choose proper k-mean clustering model for generating image authentication data. Given the image block I_i, its visual features were calculated by the following characteristics[12]:

$$\text{luminance sensitivity: } BS_i = \frac{1}{4} \sum_{v=0}^{3} I_{iv}, \tag{5}$$

$$\text{texture sensitivity: } TS_i = \sum_{v=0}^{3} |I_{iv} - BS_i|, \tag{6}$$

$$\text{contrast sensitivity: } CS_i = \max(I_i) - \min(I_i), \tag{7}$$

$$\text{entropy sensitivity: } ES_i = -\sum_{v=0}^{3} PI_{iv} \times \log(PI_{iv}),$$

$$\text{where } PI_{iv} = I_{iv} \Big/ \sum_{v=0}^{3} I_{iv}. \tag{8}$$

Chen *et al.* proposed that the relationship between image blocks can be created by applying the fuzzy c-mean clustering technique[13]. We generated the image authentication data by combining the membership matrix with a secret-key-generated random sequence. Since k-means is one of the simplest unsupervised learning algorithms, and since clustering speed is very fast, we used the k-mean clustering technique to obtain image authentication data. The feature of each block I_i can be regarded as a four-dimensional input vector, (BS_i, TS_i, CS_i, ES_i). The image matrix with a size of $M \times N$ was changed into a $4 \times D$ matrix, where the dimension of the clustering space is 4 and the number of the image blocks is D. The k-mean clustering was then applied to classify all blocks into C clusters. After performing k-mean clustering, 1D vector B containing the class membership of each block I_i is acquired. For each block, its corresponding 2-bit authentication datum ad_i is obtained by

$$ad_i = B(1, i). \tag{9}$$

Step 3.3: For each block I_i, its corresponding 6-bit recovery data consisted of the six most significant bits (MSBs) of average intensity avgI_i within each block I_i:

$$\mathrm{avg}I_i = \left\lfloor \sum_{v=0}^{3} I_{iv} \Big/ 4 \right\rfloor, \tag{10}$$

where $\lfloor \cdot \rfloor$ represents floor function.

Step 4: Chang *et al.* pointed out that the four-scanning attacker, which obtains the block-mapping sequence beforehand, can purposely tamper or modify easily the watermarked image[4]. In order to improve system security, this step was performed to encrypt the feature watermark against four-scanning attack by using logistic chaotic binary sequence z_i with exclusive-or (XOR) operation[14]:

$$FE_i = F_i \oplus z_i. \qquad (11)$$

The logistic chaotic private key of $k3$ was used to resist the four-scanning attack and enlarge key space.

Step 5: Image block I_i and its corresponding mapping block $I_{i'}$ were generated according to the block mapping algorithm described in Step 2. The watermarked image block was obtained by embedding the 8-bit encrypted feature watermark of FE_i into the two LSBs of each pixel in the corresponding block. The watermarked image I_{w} was obtained after all generated feature watermarks were embedded.

The proposed algorithm cannot only detect if the watermarked image was altered, it can also locate accurately the tampered regions. When the test image I_{w} is a distorted watermarked or unaltered watermarked image, detection could be achieved with the following steps.

Step 1: The test image was first segmented into non-overlapping blocks $I_{\mathrm{w}i}$ with a size of 2 × 2 pixels, as in the watermark embedding process.

Step 2: Secret row key $k1$ and column key $k2$ were used for block mapping. The paired block of $(I_{\mathrm{w}i}, I_{\mathrm{w}i'})$ was obtained by following Step 2 of the embedding process.

Step 3: The embedded watermark FE_i of block $I_{\mathrm{w}i}$ was extracted from the two LSBs of all pixels within mapping block $I_{\mathrm{w}i'}$. Then, the 8-bit feature watermark F_i of each block $I_{\mathrm{w}i}$ was reconstructed by

$$F_i = FE_i \oplus z_i. \qquad (12)$$

Step 4: For each block $I_{\mathrm{w}i}$, we set the two LSBs of each pixel in $I_{\mathrm{w}i}$ to zero. According to Step 3 of the embedding process, the authentication data ad'_i, the recovery data rd'_i, and the feature watermark F'_icould be obtained.

Step 5: For each block $I_{\mathrm{w}i}$, we compared the corresponding F'_i with F_i. If $F'_i = F_i$, $I_{\mathrm{w}i}$ is a valid block; otherwise, it is regarded as a tampered block.

After the block tamper detection was completed, all blocks of the test image were marked either as valid or tampered. Only tampered blocks needed to be recovered based on recovery information. We performed the same image tamper recovery mechanism, as presented in Ref. [3], as follows.

Step1: Supposing the tampered block is I_{wt}, we can find its mapping block $I_{\mathrm{wt}'}$. If block $I_{\mathrm{wt}'}$ is also marked as tampered block, tampered block I_{wt} will not be restored.

Step 2: If block $I_{\mathrm{wt}'}$ is marked as valid block, the 6-bit recovery data of the tamped block I_{wt} are extracted as two LSBs from each pixel within block $I_{\mathrm{wt}'}$.

Step 3: The 6-bit recovery data of tamped block I_{wt} are padded with two 0s to the end. The intensity of each pixel with tamped block is replaced with this recovery intensity. The recovered watermarked image I_{wr} is obtained after all tampered blocks are recovered.

A large number of experiments were conducted to validate the performance of the proposed algorithm on image tamper location, detection, and recovery. Figure 1 shows some host images and watermarked images. With respect to subjective evaluation, it seems difficult to distinguish the difference between the host and the watermarked images by human eye. With respect to objective evaluation, the peak signal-to-noise ratio (PSNR) was used to measure efficiently the visual fidelity of the host image and the watermarked image. Among the watermarked images, the image qualities measured by PSNR were greater than 40 dB. A comparison of the results in image quality is summarized in Table 1.

The probability of false acceptance (PFA) and the probability of false rejection (PFR) were adopted for measuring the accuracy of image tamper localization and detection[7]. The smaller the PFA and PFR, the better the performance of image tamper localization and

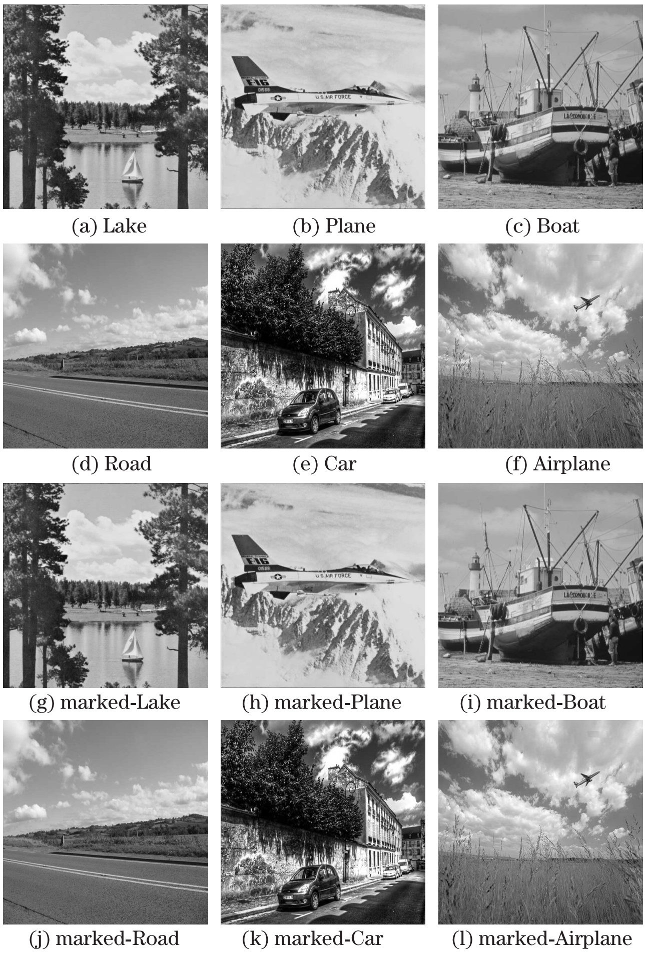

(a) Lake (b) Plane (c) Boat (d) Road (e) Car (f) Airplane (g) marked-Lake (h) marked-Plane (i) marked-Boat (j) marked-Road (k) marked-Car (l) marked-Airplane

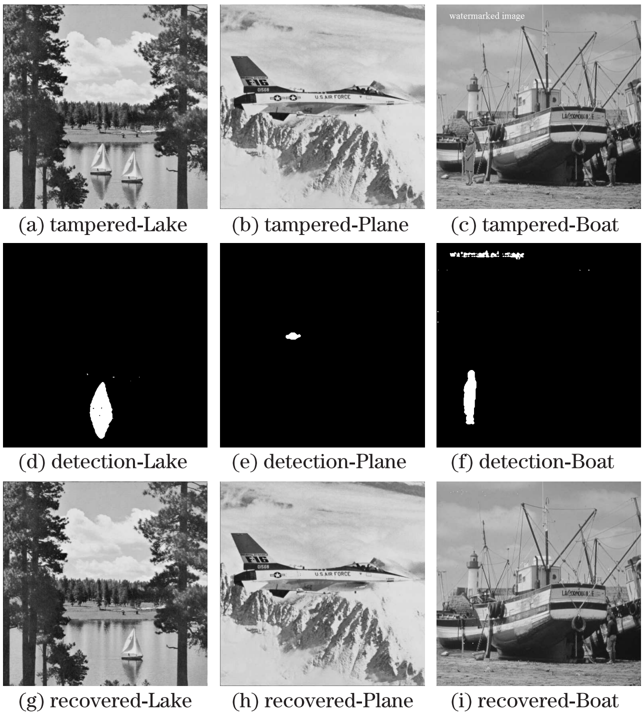
Fig. 1. (a)–(f) Host images and (g)–(l) watermarked images.

Table 1. Comparison Results of Image Quality

Image	Proposed		Ref. [13]	Ref. [7]
	PSNR (dB)	Average PSNR (dB)	Average PSNR (dB)	Average PSNR (dB)
Lake	44.22	44.50	44.22	44.15
Plane	44.35			
Boat	44.81			
Road	44.86			
Car	44.17			
Airplane	44.59			

Table 2. Experimental Results after Cut-and-Paste Attack

Image	Proposed		Proposed		Ref. [13]		Ref. [7]	
	PFA (%)	PFR (%)	Average PFA (%)	Average PFR (%)	PFA (%)	PFR (%)	PFA (%)	PFR (%)
Lake	0.23	0.21	0.26	0.24	0.32	0.4	1.13	0.05
Plane	0.21	0.17						
Boat	0.35	0.34						



(a) tampered-Lake (b) tampered-Plane (c) tampered-Boat
(d) detection-Lake (e) detection-Plane (f) detection-Boat
(g) recovered-Lake (h) recovered-Plane (i) recovered-Boat

Fig. 2. Cut-and-paste attack simulation.

detection techniques. PSNR highly depends on the size of tampered regions and the accuracy of tampered blocks identification. The greater the PSNR, the better the performance of image recovery techniques.

Below is the performance of proposed method on cut-and-paste attack. The watermarked Lake image was modified by inserting one sailboat into the image. The tampered image is shown in Fig. 2(a). The tamper detection result is shown in Fig. 2(d), and the recovered image is shown in Fig. 2(g). The "Star Logo" of the watermarked Plane image was copied and inserted into the image, as shown in Fig. 2(b). The tamper detection result and the recovered image are illustrated in Figs. 2(e) and (h). The modification made on the watermarked Boat image, as illustrated in Fig. 2(c), was done by adding some materials onto the watermarked Boat image. Figures 2(f) and (i) show the tamper detection result and the recovered image. The corresponding results of PFA, PFR, and PSNR are shown in Tables 2 and 3.

Table 3. PSNR Values after Cut-and-Paste Attack (dB)

Image	Proposed Scheme	Ref. [7]
Lake	42.66	41.85
Plane	43.13	
Boat	41.87	

Collage attack is a kind of Holliman-Memon counterfeiting attack. A counterfeiting image is constructed by combing portions of multiple watermarked images while preserving their relative spatial location within the target image[13]. The three images of Road, Car, and Airplane with a size of 512 × 512 pixels were used for simulating the collage attack. These images are shown in Figs. 1(d)–(f), and the corresponding watermarked images are illustrated in Figs. 1(j)–(l). The counterfeit image is shown in Fig. 3(a), and was constructed by copying the car from Fig. 1(k) and the plane from Fig. 1(l), and pasting them onto Fig. 1(j). The tamper detection result is shown in Fig. 3(b), and the recovered result is shown in Fig. 3(c). The corresponding results of PFA, PFR, and PSNR are shown in Tables 4 and 5.

Finally, we demonstrate the ability of the proposed method against vector quantization attack. Our method is not vulnerable to VQ attack since the block-wise dependency fashion, which is key to prevent VQ attack, is established through the block-clustering mechanism. We utilized and chose the iris images of ubiris3 database[15] for simulating VQ attack. Three sample iris images of

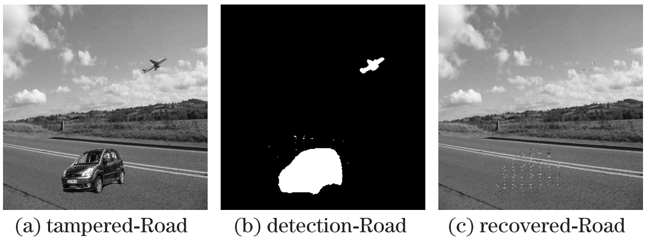

(a) tampered-Road (b) detection-Road (c) recovered-Road

Fig. 3. Collage attack simulation.

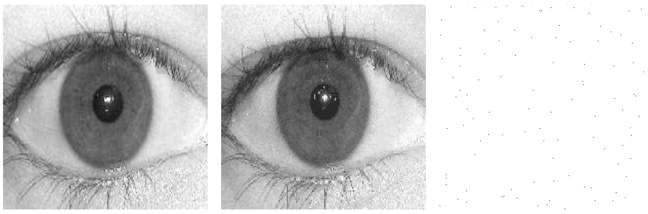
Fig. 4. Sample images of the iris database.

Table 4. Experimental Results after Collage Attack

Proposed		Ref. [13]		Ref. [7]	
PFA (%)	PFR (%)	PFA (%)	PFR (%)	PFA (%)	PFR (%)
0.43	0.66	0.7	0.81	3.82	0.01

Table 5. PSNR Values after Collage Attack (dB)

Proposed	Ref. [7]
33.68	31.94



(a) host-image (b) VQ-image (c) tampered-detection

Fig. 5. VQ attack simulation.

this database are shown in Fig. 4. Figures 5(a)–(c) represent the host image, VQ counterfeit image, and tamper-detection image, respectively. Since all blocks were disturbed by VQ attack, the PFR cannot be evaluated; the PFA is 0.52%, which is smaller than that of Chen's scheme (0.89%)[13].

In conclusion, we have presented a novel fragile watermarking scheme with superior image tamper localization, detection, and recovery. We have employed the k-mean clustering technique and image local visual features in order to establish new block-wise dependency mechanism in accordance with HVS. Multiple encryption techniques are used for improving system security. Experimental results show that the proposed scheme can preserve high perceptual quality and identify effectively tampered blocks in terms of lower PFA and PFR. The capability of tamper recovery is improved. In future work, we aim to focus on decreasing PFR while sustaining the superior location accuracy, and conduct research on recoverable semi-fragile watermarking scheme.

References

1. S. Zhu and J. Liu, Chin. Opt. Lett. **7,** 580 (2009).
2. J. Zheng, D. Feng, Y. Zhang, and R. Zhao, Chin. J. Computers (in Chinese) **27,** 371 (2004).
3. P. L. Lin, C.-K. Hsieh, and P.-W. Huang, Pattern Recognition **38,** 2519 (2005).
4. C.-C. Chang, Y.-H. Fan, and W.-L. Tai, Pattern Recognition **41,** 654 (2008).
5. M.-S. Wang and W.-C. Chen, Computer Standards & Interfaces **29,** 561 (2007).
6. X. Zhang and S. Wang, IEEE Trans. Multimedia **10,** 1490 (2008).
7. H.-J. He, J.-S. Zhang, and H.-M. Tai, Lecture Notes in Computer Science **5806,** 132 (2009).
8. M. Li, J. Tian, and F. Chen, Pattern Recognition Lett. **29,** 392 (2008).
9. Z. Jiang and Y.-X. Huang, Information Sci. **179,** 2002 (2009).
10. R.-Z. Wang and Y.-D. Tsai, Pattern Recognition **40,** 398 (2007).
11. J. F. Delaigle, C. Devleeschouwer, B. Macq, and I. Langendijk, in *Proceedings of 2002 IEEE International Conference on Multimedia & Expo* 489 (2002).
12. J. Wu, J. Xie, and Y. Yang, J. Harbin Institute Technol. **14,** 281 (2007).
13. W.-C. Chen and M.-S. Wang, Expert Systems with Applications **36,** 1300 (2009).
14. A. Kanso and N. Smaoui, Chaos Solitons and Fractals **40,** 2557 (2009).
15. H. Proença and L. A. Alexandre, Lecture Notes in Computer Science **3617,** 970 (2005).