Watermarking Shape Datasets with Utility and Distance Preservation

Anshika .V. Gupta
PG Student M B E Society's
College of Engineering,
Ambajogai Maharashtra, India

B. M. Patil
Professor M B E Society's
College of Engineering,
Ambajogai Maharashtra, India

V. M. Chandode
Associate Professor M B E
Society's College of
Engineering, Ambajogai
Maharashtra, India

ABSTRACT

Due to promulgation of data over internet significance of protection of one's intellectual property is the important topic with technological and legal aspects. Watermarking scheme is used for establishing the ownership of dataset containing multiple objects. As watermarking scheme distorts distance relationship graph, methodology preserves utility of dataset by preserving important distance properties such as nearest neighbor (NN) and minimum spanning tree (MST) of the original data set. We use fast algorithms for NN and MST which gives improved security without any sacrifice in distance relationships then NN and MST algorithms used earlier.

Keywords

Algorithm, fast nearest neighbor algorithm, minimum spanning tree algorithm, fast minimum spanning tree algorithm

1. INTRODUCTION

Due to wide and illegal spread of data over internet right protection of dataset has become inherent part of various companies for business and research practices to exchange data. Most popularly used tool for right protection is watermarking. As watermarking necessarily adds noise to the dataset, it distorts distance relationship graph. Hence goal is not only to provide right protection of data but also utility preservation of datasets containing multiple objects. Embed ownership via watermarking should satisfy properties of imperceptibility (no apparent visual distortion), delectability, preservation of distance relationship graph and robustness to malicious attacks.

Most of the mining algorithms and datasets depend upon the distance like nearest neighbor (NN) search[1][2], nearest neighbor classification, data clustering algorithms[4] (uses MST), visualization method and embedding techniques[5] (uses NN or MST), outlier detection (uses NN), dimensionality reduction (e.g. ISOMAP), phylogeny construction.

This paper presents the comparison done with fast algorithms for NN and MST based on analysis of [3][6].

2. LITRETURE SURVEY

Loads of watermarking research on multimedia datasets [7] has been done like on images, vector graphics audio and video [13]. A key drawback of multimedia watermarking is that it does watermarking on single digital vector object rather than multiple objects hence it lacks a distance relationship between the objects but in this work we considered overall topology of multiple objects after watermarking. Some privacy preservation techniques can be achieved through (a) protection via perturbation like adding noise [12] (b) condensation [15] (c) rotational perturbation [11] these perturbation techniques. These techniques did not work on actual altered data but try to reconstruct original data from distributed noise vector that has been added on the dataset [10] [14] privacy can also be achieved by partitioning of data horizontally and vertically over sites [16] this approach is different from previous approaches as one consider entire distributed dataset not by dividing datasets in sections.

They have seen watermark techniques for watermarking relational databases [10][14] but this technique is not useful for right protection of fast stream data [11] since watermarking relational data depend upon availability of entire dataset during watermarking process but stream data is available as soon as it is generated and also the lack of primary key so not applicable to [11] and hence numeric sequence is not applicable for this work as later work is on watermarking single numerical sequence not entire sequence to maintain pair wise distance relationships and also though it is resilient to attacks like sampling, summarizations, random alterations but not resilient to geometric transformations.

This work is different from just traditional watermarking which focused on privacy preservation, as scheme[3][6] not only works on perturbed data but also distance preservation by using NN and MST so that it can be relevantly used where distance relationship preservation is important property like in various data mining and machine learning algorithms.

3. RIGHT PROTECTION BY WATERMARKING

This section demonstrates how to embedded watermark key (secret information) in cover image. Image is extracted in the 2d sequence [6].The spread spectrum approach [7] is used to properly tune the power of watermark across multiple frequencies and over multiple objects in dataset for making difficult to remove watermark on attacks. Watermarked image with seal should satisfy following properties.

1. Imperceptibility
2. Detect ability
3. Preservation of minimum spanning tree and nearest neighbor
4. Robustness

3.1 Risk model

Risk model which is considered is: an adversary can make modification by removing the watermark in the watermark data and attacker can only make changes in the data without hindering its utility. Adversary is assumed to a) Make disturbances using noise addition, geometric transformations like scaling rotation, translation etc. b) May not have an idea of the secret key but can have knowledge of algorithm.

3.2 Embedding the seal

Consider each object represented as a vector of complex numbers,$x = \{x_1, \dots\dots, x_n\}$, where $x_k = a_k + b_k i$ (i is the imaginary unit, $i^2 = -1$), and where$a_k$and$b_k$are real and imaginary part respectively, describe the coordinates of the k-th point of object x in an imaginary plain. Such a model can capture coordinates of shape parameter in 2d plane. Team used a spread-spectrum approach which embeds the secret information (watermark) across multiple frequencies of each object and through multiple objects of the dataset. This makes removal of the watermark very difficult without destroying its utility.

3.3 Overview of watermarking technique

For protecting the data from malicious attack team embed the watermark in frequency domain rather than space domain. Mapping of an object x is into the frequency domain using its complex Fourier descriptors$X = \{X_1, \dots\dots, X_n\}$, is done by the regularized discrete Fourier transform, DFT(x), after embedding watermark key which is given by vector $W \in \{-1, 0, +1\}^n$ that is taking 3 distinct values. Then again watermark shape is gained by mapping from frequency domain to space domain by using its inverse IDFT(X).This technique is called multiplicative watermarking technique .Every coefficient $X_j$can be viewed in terms of its magnitude $\text{ɱ}_j$and phase$\emptyset_j$, as$X_j = \text{ɱ}_j e^{\emptyset j^l}$.

Defination1: (Multiplicative Watermark Embedding (W, p)). Let assume a sequence $x \in C^n$ with a set of Fourier descriptors X, a watermark $W \in R^n$ and power $p \in [0, 1]$ by which intensity of the watermark is specified. Watermarked sequence$\hat{x}$ is generated by multiplicative watermark embedding (W, p) by replacing the magnitudes of each Fourier descriptor of x with the watermarked magnitude$\hat{\text{ɱ}}$while not altering the phases, specifically:

$$\hat{\text{ɱ}} = \text{ɱ}_j * (1 + \rho W_j), \text{ and } \widehat{\phi_j} = \phi_j \quad (1)$$

Using original phases and modified magnitude backward mapping can be done from frequency domain to space domain by using IDFT.

While embedding the watermark the first Fourier descriptor has to be excluded as it is a DC component having center of mass of object x which is more susceptible for translation attack. A simple translation attack can shift the DC component and hence can erase the watermark at that part of watermarked image. So, we can say vector $W \in \{-1, 0, +1\}^n$ where,

$W_j = 0$ if (j=1 (DC component) or j=0)

$W_j = \{-1, 1\}$ if there is non-zero element.

And also $\sum W_j = 0$, so we can say only those l elements of $W_j \neq 0$ have secret information irrespective of sequence length n.

3.4 Efficiency of watermark under attacks

Some attacks are considered which adversary can perform to destroy watermark.

3.5 Geometric transmutations

Attacker may perform these types of geometric attacks [8][9] to remove embedded seal:

Rotation: Rotational transmutation can be an effective attack as it can possibly destroy secret watermark.

Translation: As watermark is computed from Fourier transformation and as during process team don't embed in the DC component, translation attack does not have any effect on watermarked image

Scaling: Adversary may try to scale the objects by some aspect which leads to scaling the magnitude of Fourier descriptors which can be normalized in detection process.

3.6 Noise addition

Adversary may add large amount of noise or Gaussian noise in frequency domain where watermark seal is embedded. But it has to add large amount of noise which will lead to destruction of usability of data. So this type of attack will not be of any use to attacker.

3.7 Lower and upper bounds on distance distortion

As watermark distorts distance relationship graph team proposed tight lower and upper bounds on contraction and expansion because of watermarking. Because of constraints provided by the lower and upper bound the two objects that are at some specific distance with each other cannot get far apart or too close after watermark embedding [3]. This is done by establishing restricted isometric property which makes watermark embedding power to lie in the interval $[\rho_{\min}, \rho_{\max}] \subseteq [0,1]$

Definition 2: (Tight lower and upper bound on the distance of watermarked objects). Given two objects $x, y \in D$ for any compatible watermark $W \in W(D)$ and any embedding power $[\rho_{\min}, \rho_{\max}] \subseteq [0,1]$ we have restricted isometric property

$$(1 - \rho_{\max})D(x, y) \leq \widehat{D_p}(x, y) \leq (1 + \rho_{\max})D(x, y) \qquad (2)$$

These bounds are tight and they help us to make foundations for fast algorithms for NN and MST preservation.

4. NN PRESERVATION

Given an object x and dataset D such that $x \in D$, x preserves its nearest neighbor after watermark w is embedded with power p and $NN(x) \neq x$ if

$$\widehat{D_p}\big(x, NN(x)\big) \leq \widehat{D_p}(x, y), \forall y \in D, y \neq x (3)$$

If it applies to all $x \in D$ then we can say that NN is preserved.

4.1 NN –P watermarking problem

Given a dataset D find a maximum feasible power when $\rho_{\min}$ and ,$\rho_{\max}$ are given and$\rho_{\min} \leq \rho \leq \rho_{\max}$ after watermark embedding such that at mostτ among the objects in the dataset D preserve their NN with watermark where $W \in W(D)$. The NN preservation algorithm used below is taken from [3][6].

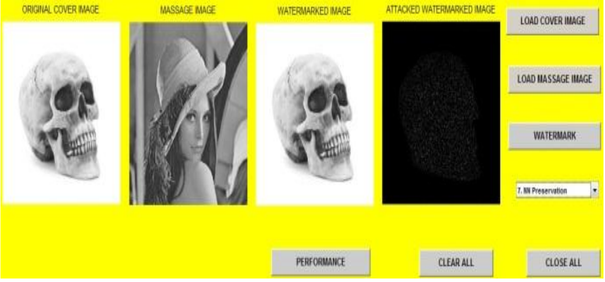
4.2 NN preservation algorithm

```
The inputs variables are: D, W,$\rho_{min}$, $\rho_{max}$, $\tau$
Get the output: $\rho^*$
NN(D)=find 1- Nearest Neighbors of D
for all x$\epsilon D$ do
    feasible_powers(x)=[$\rho_{min}$, $\rho_{max}$]
    for all y$\epsilon D$, $y \neq x$, y≠NN(x) do
        feasible_powers(x) =
            Solve($\widehat{D_p^2}(x, NN(x)) \leq \widehat{D_p^2}(x, y)$|
            feasible_powers(x))
```

```
8:  end for
9:   end for
10: $\rho^*$= max{p: |{x : p∉ feasible_powers(x)}| ≤ $\tau$.|D|}
```

This Algorithm retains the nearest neighbor of each object by calculating maximum feasible power for embedding within the range [$\rho_{\min}$, $\rho_{\max}$].

The results after applying NN algorithm is given in figure.



7. NN Preservation

Fig 3: Result of NN algorithm

5. MST PRESERVATION

5.1 MST-P watermarking problem

Given a dataset D find a maximum feasible power when $\rho_{\min}$ and ,$\rho_{\max}$ are given and

$\rho_{\min} \leq \rho \leq \rho_{\max}$ after watermark embedding such that atmost τ among the edges of MST in the distance graph D preserve MST with watermark W given, where W∈W(D).

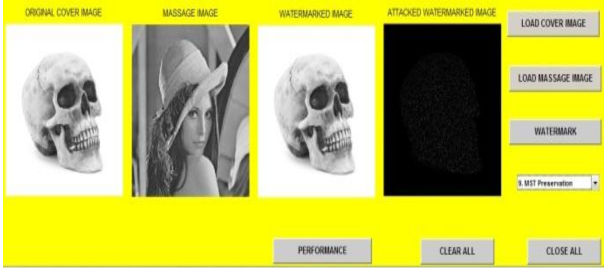
5.2 MST preservation algorithm

MST preservation algorithm used here is taken from [3][6]

```
1: The inputs variables are: D, W, $\rho_{min}$ , $\rho_{max}$ , $\tau$
2:  Get the output: $\rho^*$
3: T (D, E) = find MST of D (using kruskal's algorithm)
4: for all e$\epsilon E$ do
5:    feasible_powers (e) = [$\rho_{min}$ , $\rho_{max}$ ]
6:       for all u ∈ $U_e do$
7:      for all v ∈ $Vedo$
8:   feasible_powers (e) =
     (($\widehat{D_p^2}(e) \leq (\widehat{D_p^2}\ (u,v)$| feasible_powers (e))
9:        end for
10              end for
11: end for
12: $\rho^*$= max {p: | {e : p∉ feasible_powers(e)}| ≤ $\tau$.|D|-1)}
```

This algorithm removes those weak powers which violate MST properties.

The result after applying MST preservation algorithm is shown in figure2.



9. MST Preservation

Fig 4: Result of MST preservation algorithm

6. FAST ALGORITHM

The previous algorithms were not so resilient to attack. As compared to previous algorithms fast algorithms uses restricted isometric property to make watermarked image more secure that is better detect ability of watermark.

6.1 Fast NN preservation

As restricted isometric property states that if ratio of distance between object x and y D(x,y) to the distance between object x and its neighbor D(x, NN(x)) is greater than or equal to $\rho_{\max}$ then object y does not exploit the nearest neighbour of x NN(x) after watermark embedding.

Sufficient condition for NN preservation is

For x, y ∈ D, y≠x, NN(x), if

$$\frac{\mathrm{D(x,y)}}{\mathrm{D(x,NN\ (x)}} \geq \frac{1+\rho_{\max}}{1-\rho_{\max}} \qquad (4)$$

Then y does not exploit the nearest neighbor of x NN(x) after the watermark embedding, for all watermarks W ∈W (D) and embedding powers ρ ∈ [$\rho_{\min}$, $\rho_{\max}$]

6.1.1Algorithm

This algorithm is directly taken from [3].

```
1.   The inputs variables are: D,W∈W(D), $\rho_{min}$ , $\rho_{max}$ , $\tau$
2.   Get the output: $\rho^*$
3.   NN(D)=find 1- nearest neighbours of D
4.   for all x∈D do
5.   feasible_powers(x)= [$\rho_{min}$ , $\rho_{max}$ ]
6.   for all y$\epsilon D,\ y \neq x$, y≠NN(x) do
7.   If$\frac{D(x,y)}{D(x,NN(x)}\geq\frac{1+\rho_{max}}{1-\rho_{max}}$ then
8.   feasible_powers(x)=
     ($\widehat{D_p^2}(x,NN(x)) \leq \widehat{D_p^2}\ (x,y)$| feasible_powers(x))
9.       End if
10.    End for
11.  End for
12. $\rho^*$= max {p: |{x: p∉ feasible_powers(x)}| ≤ $\tau$.|D|}
```

The results after applying fast NN algorithm are in figure.

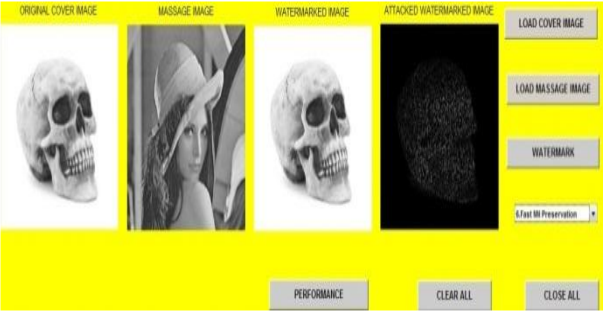


Fig 5: Result of fast NN preservation algorithm

6.2 Fast MST preservation

MST is preserved by this condition

For an edge e in an MST of graph D and object u∈ U_e, v∈ V_e, if

$$\frac{D(u,v)}{D(e)} \geq \frac{1+\rho_{\max}}{1-\rho_{\max}} \qquad (5)$$

Then edge (u,v) does not breach the MST at edge after the watermark embedding, for all watermarks W ∈ W(D) and embedding powers p ∈ [$\rho_{\min}$, $\rho_{\max}$]

6.2.1 Algorithm

Fast MST preservation algorithm is directly taken from [3].

```
The inputs variables are: D, W, $\rho_{min}$, $\rho_{max}$, $\tau$
Get the output: $\rho^*$
T (D, E) = find MST of D
for all e$\epsilon E$ do
   feasible_powers(e)=[$\rho_{min}$, $\rho_{max}$]
      for all u ∈ $U_e do$
      for all v ∈ $V_e do$
 If$\frac{D(u,v)}{D(e)} \geq \frac{1+\rho_{max}}{1-\rho_{max}}$ then
feasible_powers(x)=
(($\widehat{D_p^2}(e) \leq (\widehat{D_p^2}(u,v)$| feasible_powers (x))
          end if
end for
end for
$\rho^*$= max {p: | {e : p∉ feasible_powers(e)}| ≤ $\tau$.|D|-1)}
```

The result after applying fast MST preservation algorithm is shown in figure.

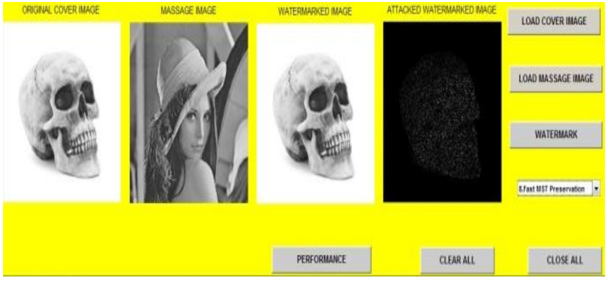


Fig 6: Result of fast MST preservation algorithm

7. RESULT AND ANALYSIS

Resilience of the watermarking scheme is analyzed with reference to the risk model presented in section 3. Robustness of the watermarking technique can be tested by scrutinizing detect ability of the watermark sealed in the digital data, attacked by an adversary. The attack can be geometric transformation, noise addition or removal of objects from the original image. One of the powerful tools to analyze the watermark detection is ROC curve. In a ROC curve, True Positive Rate (TPR) vs. False Positive Rate (FPR) is plotted to compare detection techniques. The curve more inclined towards top left corner of the graph implicitly proves the robustness and accuracy of detection algorithm. Also, ROC curve of the random binary detection scheme should be as far as possible from the curve of watermark detection scheme to provide better resilience against security attacks.

Graph 1 and Graph 3 shows the ROC curve of watermarking detection technique when attacked by noise addition and geometric transformation. The curve is far from the base (starting from bottom left corner) which proves the accuracy of the detection scheme. Other than this ROC curve also proves that a Random binary detection scheme implemented by the attacker cannot detect the watermark embedded in the image. Graph 2 and Graph 4 illustrates capability of the watermarking detection scheme when objects are removed by the attacker, Even with the 10% of the objects being removed, performance of the detection scheme is better than that of the random binary detection scheme, which further consolidates the resilience of the watermarking scheme.

Table 4: parameters for increasing percentage of removal of objects (5%, 10%) for "fish" database

Pa rameters	5% removal of object			10% removal of object		
True positive	0.6	0.7	0.9	0.6	0.7	0.9
False positive	0.6	0.65	0.8	0.6	0.65	0.8

Table 1: Parameters for watermark detection on "skull "database

parameters	Geometric transformation			Noise attack		
True positive	0.1	0.3	1	0.1	0.2	1
False positive	0.01	0.07	0.31	0.07	0.12	0.32

Table 2: Parameters for watermark detection on "fish" database

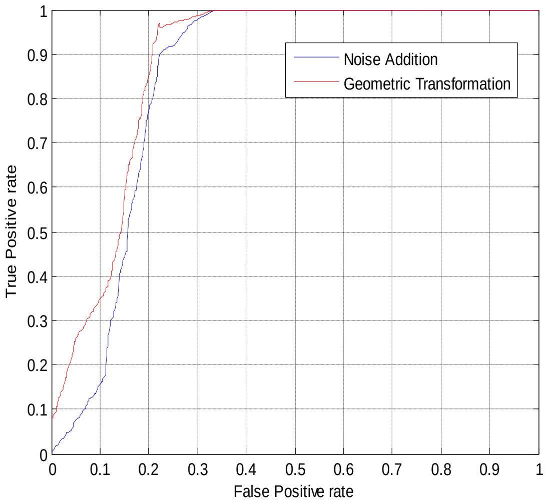
parameters	Geometric transformation			Noise attack		
True positive	0.4	0.5	1	0.4	0.5	1
False positive	0.32	0.14	0.31	0.14	0.15	0.32

Table 3: parameters for increasing percentage of removal of objects (5%, 10%) for "skull" database

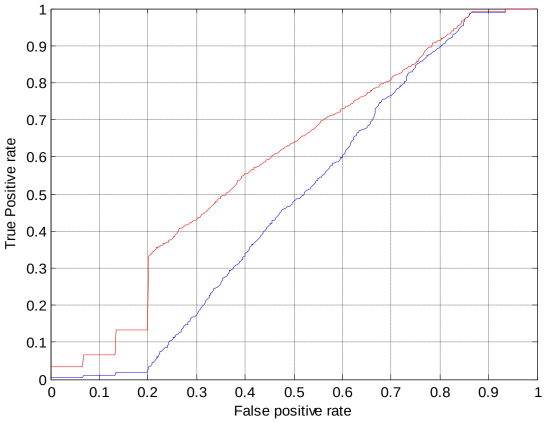
parameters	5% removal of object			10% removal of object		
True positive	0.1	0.3	1	0.1	0.2	1
False positive	0.25	0.35	0.9	0.02	0.2	10.9

8. CONCLUSION AND FUTURE WORK

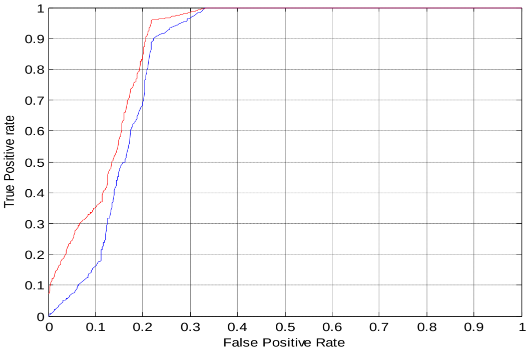
In this paper by graph analysis it is seen that fast algorithms are more resilient to attack then NN and MST algorithms. By using fast algorithms one can make the embedding more strong and robust to attacks. In future by increasing capacity ratio, embedding could be made stronger.



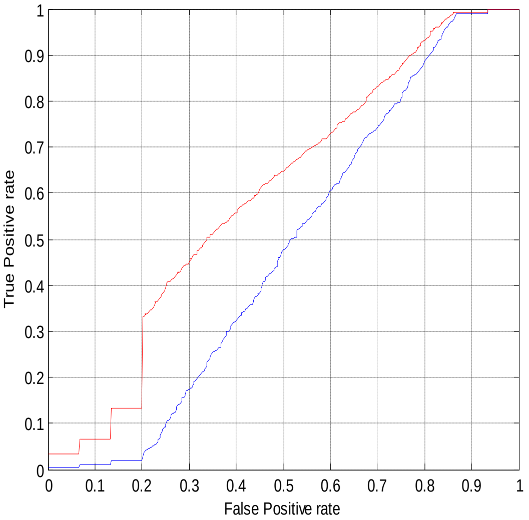
Graph1: ROC curve for watermarking detection on "skull" database



Graph2:ROC curve with increasing percentage of removal of objects (5%,10%) for skull database



Graph3: ROC curve for watermarking detection on"fish" database



Graph 4: ROC curve with increasing percentage of removal of objects (5%, 10%) for "fish" database

9. REFERENCES

[1] D.W. Aha, D. Kibler, and M.K. Albert, "Instance based learningalgorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, 1991.

[2] C. G. Atkeson, A. W. Moore, and S. Schaal, "Locally weightedlearning," *Artif. Intell. Rev.*, vol. 11, pp. 11–73, Feb. 1997.

[3] S.I.Zoumpoulis, N.M. Vlachos and M.Freris,"Right-protected data publishing with provable distance-based mining" IEEE Transactions on knowledge and data engineering, vol. 26, no.8, aug.2014.

[4] Y. Xu, V. Olman, and D. Xu, "Minimum spanning treesfor gene expression data clustering," *Genome Inform.*, vol. 12,pp. 24–33, 2001.

[5] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometricframework for nonlinear dimensionality reduction," *Sci.*,vol. 290, no. 5500, pp. 2319–2323, 2000.

[6] M. Vlachos, C. Lucchese, D. Rajan, and P. S. Yu, "Ownership protectionof shape datasets with geodesic distance preservation," in*Proc. 11th Int. Conf. EDBT*, Nantes, France, 2008, pp. 276–286.

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

[8] F. Hartung, J.K. Su, and B. Girod., "Spread spectrum watermarking:Malicious attacks and counterattacks," in *Proc. SPIE SecurityWatermarking Multimedia Contents,* vol. 3657, San Jose, CA, USA,1999.

[9] V. Solachidis and I. Pitas, "Watermarking polygonal lines usingFourier descriptors," *IEEE Comput. Graph. Appl.*, vol. 24, no. 3,pp. 44–51, May/Jun. 2004.

[10] R.S. Agrawal and J.A. Kiernan, "Watermarking relational databases,"in *Proc. 28th Int. Conf. VLDB*, Hong Kong, China, 2002,pp. 155–166.

[11] R. Sion, M. J. Atallah, and S. Prabhakar, "Rights protection fordiscrete numeric streams," *IEEE Trans. Knowl. Data Eng.*, vol. 18,no. 5, pp. 699–714, May 2006.

[12] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On theprivacy preserving properties of random data perturbation techniques,"in *Proc. 3rd IEEE ICDM*, Washington, DC, USA, 2003,pp. 99–106.

[13] K. Chen and L. Liu, "Privacy preserving data classification withrotation perturbation," in *Proc. 5th ICDM*, Washington, DC, USA,2005, pp. 589–592.

[14] R. Sion, M. Atallah, and S. Prabhakar, "Rights protectionfor relational data," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12,pp. 1509–1525, Dec. 2004.

[15] C. C. Aggarwal and P. S. Yu, "A condensation approach to privacypreserving data mining," in *Proc. Int. Conf. EDBT*, Crete, Greece,2004, pp. 183–199.

[16] H. Yu, J. Vaidya, and X. Jiang. Privacy-PreservingSVM Classification on Vertically Partitioned Data. InPAKDD, pages 647–656, 2006.