Content Relevant Subspace Watermarking Methods

Wen-Liang Hwang
Institute of Information Science
Academia Sinica, Taiwan

Jengan Tzeng and I Liang Chern
Mathematics Department
National Taiwan University, Taiwan

Introduction

- **Embed digital signatures, called "Watermarks" in contents are important for copyright protection, copyright control, and information hiding in multimedia applications.**
- **In the sense of copyright protection, watermarking is a detection problem :**
- **Given a test image T we are testing whether it comes from a random source :**

$$\begin{aligned} \mathbf{T} &\sim \mathbf{X} + \mathbf{M} + \textbf{distortion noise} \\ &= \mathbf{X}^{\mathbf{M}} + \textbf{distortion noise}, \end{aligned}$$

  where X is host image and $\mathbf{X}^{\mathbf{M}}$ is watermarked image, and M is watermark image.
- **I. Cox** *et. al* **: In watermark detection, X is** *not purely a noise*, **since the media content is known completely to the watermark embedder at embedding stage.**
- **One should embed watermark according to the information of the content. But, How ?**

Motivation

- **In a batter field, soldiers tend to hide in places that are least likely to be attacked.**
- **Where are the save places in X to hide watermark against attacks ?**
  If we can guess the attacks $\{A\}$ of a pirate on X, can we
  1. **Find the places in the image X that are least likely to be modified by the attacks ? and**
  2. **Hide watermark information in the places.**
- **We call the places the** *Watermark Space* **of the image with respect to the attacks $\{A\}$.**
- **The** *Wavelet Space* **is content-dependent.**

Problem Model

- **$\underline{\mathbf{X}}^{\mathrm{M}}$: A variation of the watermarked image X^{M}. May be a version after attacks.**

$$\begin{aligned}\underline{\mathbf{X}}^{\mathrm{M}} &= \sum_{i,j} < \underline{\mathbf{X}}^{\mathrm{M}}, \Phi_{i,j} > \tilde{\Phi}_{i,j} \\ &= \sum_{i,j} < \mathbf{X}, \Phi_{i,j} > \tilde{\Phi}_{i,j} + \sum_{i,j} < \mathbf{X}^{\mathrm{M}} - \mathbf{X}, \Phi_{i,j} > \tilde{\Phi}_{i,j} + \sum_{i,j} < \underline{\mathbf{X}}^{\mathrm{M}} - \mathbf{X}^{\mathrm{M}}, \Phi_{i,j} > \tilde{\Phi}_{i,j}.\end{aligned}$$

- **Host feature vector:**

$$[< \mathbf{X}, \Phi_{i,j} >]$$

- **Watermark feature vector :**

$$\mathbf{m} = [< \mathbf{X}^{\mathrm{M}} - \mathbf{X}, \Phi_{i,j} >]$$

- **Variations from Watermark feature :**

$$\underline{\mathbf{e}}^{M} = [< \underline{\mathbf{X}}^{\mathrm{M}} - \mathbf{X}^{\mathrm{M}}, \Phi_{i,j} >]$$

Problem Model (Continue)

- Given $\underline{\mathbf{e}}^M$.
- Can we select
  - the watermark feature m of X and
  - a sub-feature space W

  such that for the feature t of a test image:
  - *High Detection Prob.* -
    If t is from our random source, then the correlation measurement $sim(\mathbf{m}, P_W(\mathbf{t}))$ will as large as possible, and
  - *Low False Alarm Prob.* -
    If t is not from our random source, then $sim(\mathbf{m}, P_W(\mathbf{t}))$ will as small as possible, where

  P_W is the projection to W, our *watermark space*.

Watermark Subspace Selection

- Suppose our feature space is R^N, and that our watermark feature is $\mathbf{m} \in W \subseteq R^N$.
- $\underline{\mathbf{e}}^M$ can be rewritten as

$$\underline{\mathbf{e}}^M = \underline{\alpha}\mathbf{m} + \underline{\mathbf{v}},$$

  where

  1. $\underline{\alpha}$ is a scalar random variable, obtained by projecting $\underline{\mathbf{e}}^M$ onto $\mathbf{m}$, and
  2. $\mathbf{m} \perp \underline{\mathbf{v}}$.

- If W is chosen such that most of the realizations of $\underline{\mathbf{e}}^M$:

$$\begin{cases} ||P_W(\underline{\mathbf{v}})|| << ||\mathbf{m}|| \\ |\underline{\alpha}| \text{ is close to } 0, \end{cases} \tag{1}$$

  then for most of $\underline{\mathbf{e}}^M$, we will have ***high*** detection probability and ***low*** false alarm probability.

- If W is perpendicular to most of the realizations of $\underline{\mathbf{e}}^M$, then the conditions in (1) will be satisfied.

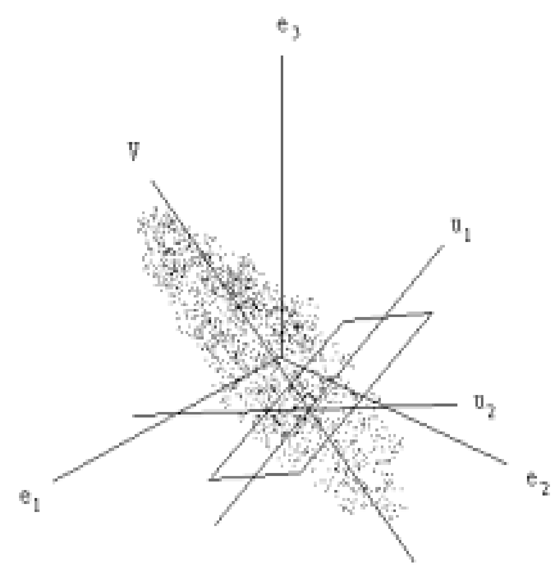
Watermark Space Selection (Continue)

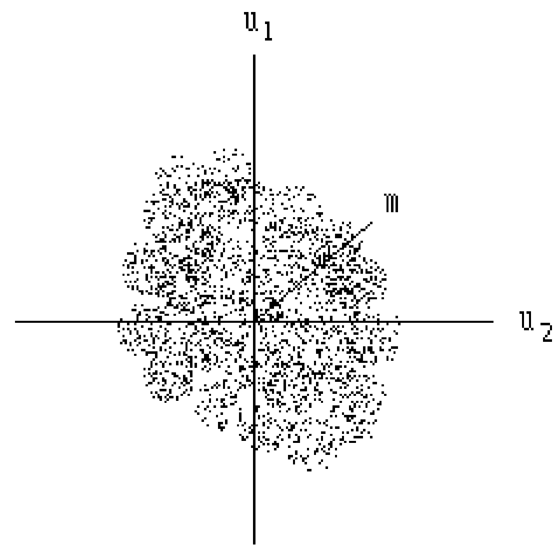
- **Detection Prob.:**

$$\begin{aligned} sim(\mathbf{m}, P_W(\mathbf{m} + \underline{\mathbf{e}}^M)) &= sim(\mathbf{m}, P_W(\mathbf{m}) + P_W(\underline{\alpha}\mathbf{m} + \underline{\mathbf{v}})) \\ &= sim(\mathbf{m}, (1 + \underline{\alpha})\mathbf{m} + P_W(\underline{\mathbf{v}})) \\ &\approx sim(\mathbf{m}, (1 + \underline{\alpha})\mathbf{m}) = 1. \end{aligned}$$

- **False Alarm Prob.:**

$$\begin{aligned} sim(\mathbf{m}, P_W(\underline{\mathbf{t}})) &\approx sim(\mathbf{m}, P_W(\underline{\mathbf{e}}^M)) \\ &= \underline{\alpha} + P_W(\underline{\mathbf{v}}) \\ &= \underline{\alpha} \approx 0. \end{aligned}$$





Our Watermarking Strategy

Selection by means of Second Order Statistics

- **We can find W such that the inner product of any vector $\mathbf{m} \in W$ to $\underline{\mathbf{e}}^M$ is small by means of statistics**

$$min_{\mathbf{m}\in\mathbf{W}}\ E\{(\mathbf{m}'\underline{\mathbf{e}}^M)(\mathbf{m}'\underline{\mathbf{e}}^M)'\},$$

where $\mathbf{m}'$ is the transpose of $\mathbf{m}$.

- $$\begin{aligned} E\{(\mathbf{m}'\underline{\mathbf{e}}^M)(\mathbf{m}'\underline{\mathbf{e}}^M)'\} &= \mathbf{m}'E\{(\underline{\mathbf{e}}^M)(\underline{\mathbf{e}}^M)'\}\mathbf{m} \\ &= \mathbf{m}'\mathbf{U}\Sigma\mathbf{U}'\mathbf{m} \\ &= \sum_{i=1}^{N} \sigma_i^2(\mathbf{m}'\mathbf{u}_i)^2, \end{aligned}$$

where

1. $\Sigma = diag(\sigma_1^2, \sigma_2^2\ \cdots, \sigma_N^2)$**.**

2. $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_N]$ **is the matrix of eigenvectors.**

- **Optimal solution is assigning our watermark feature m to the subspace spanned by eigenvectors whose corresponding eigenvalues are zeros.**

Fixed-Dimension Watermark Subspace

- **In practice, it is convenient to fix the dimension of W, say D, and to choose W such that it is spanned by the eigenvectors corresponding to the D smallest eigenvalues.**
- **This corresponds to finding a linear transformation of $\underline{\mathbf{e_c}}^M$ with a matrix A as**

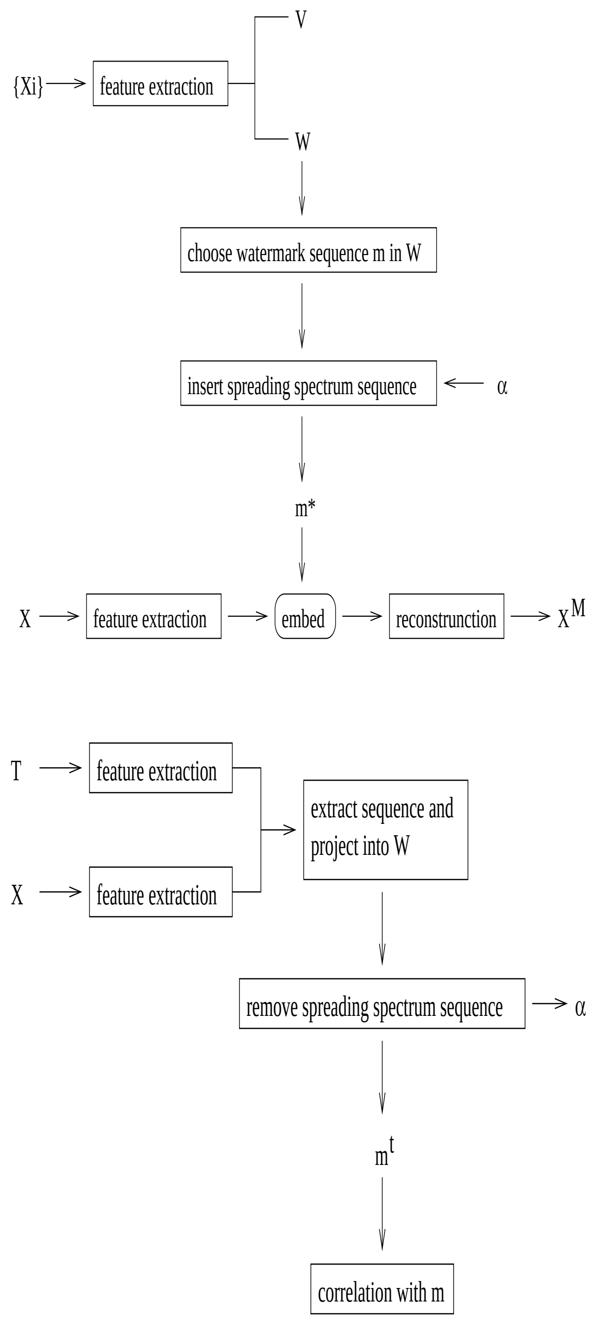
$$\mathbf{A}'\underline{\mathbf{e_c}}^M,$$

where A is an N by D matrix, whose rank is D with $D \leq N$ and where each column of A has only one non-zero element with a value of 1

- **such that the following objective function is minimized:**

$$\min_{\mathbf{A}} trace(\mathbf{A}'\mathbf{U}\Sigma\mathbf{U}'\mathbf{A}),$$

where $trace$ is the trace operation on a matrix.



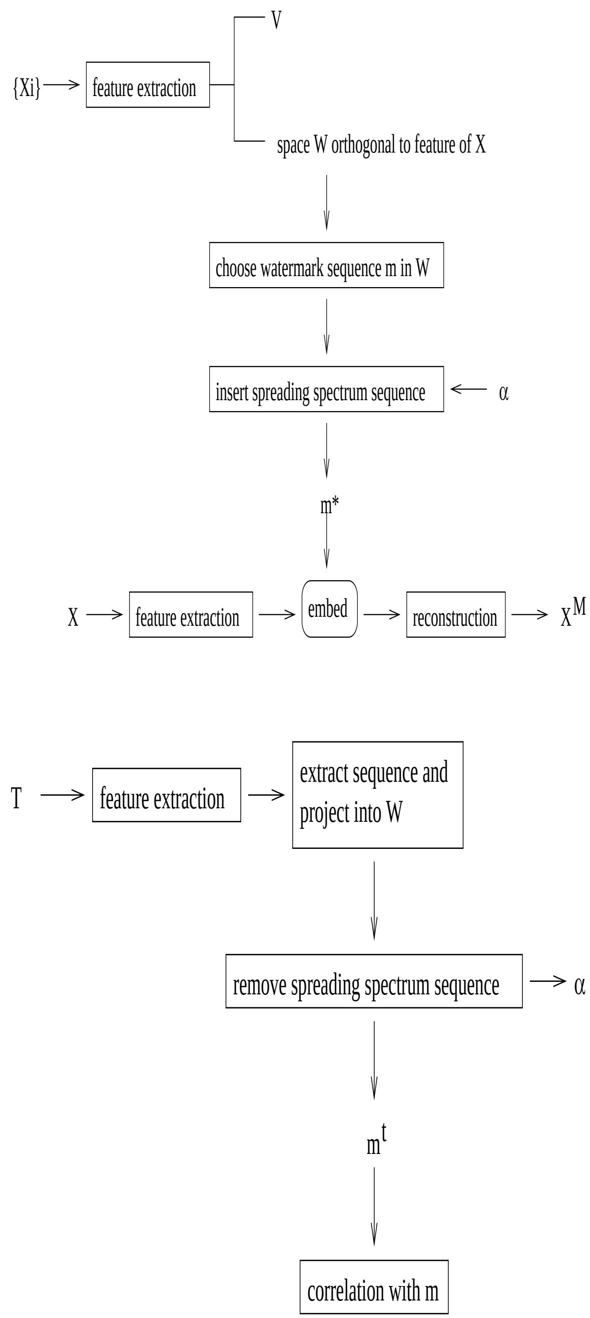
Watermark Encoding/Decoding : W is released

Blind Watermark Subspace

- **We can embed our watermark feature such that the extraction of the feature uses no host image.**
  - **Let W' be our watermark subspace.**
  - **We find a subspace of W' such that the subspace W is perpendicular to the feature vector of the host image (Gram-Schmidt):**

$$P_W([< \mathbf{X}, \Phi_{i,j} >]) = 0.$$

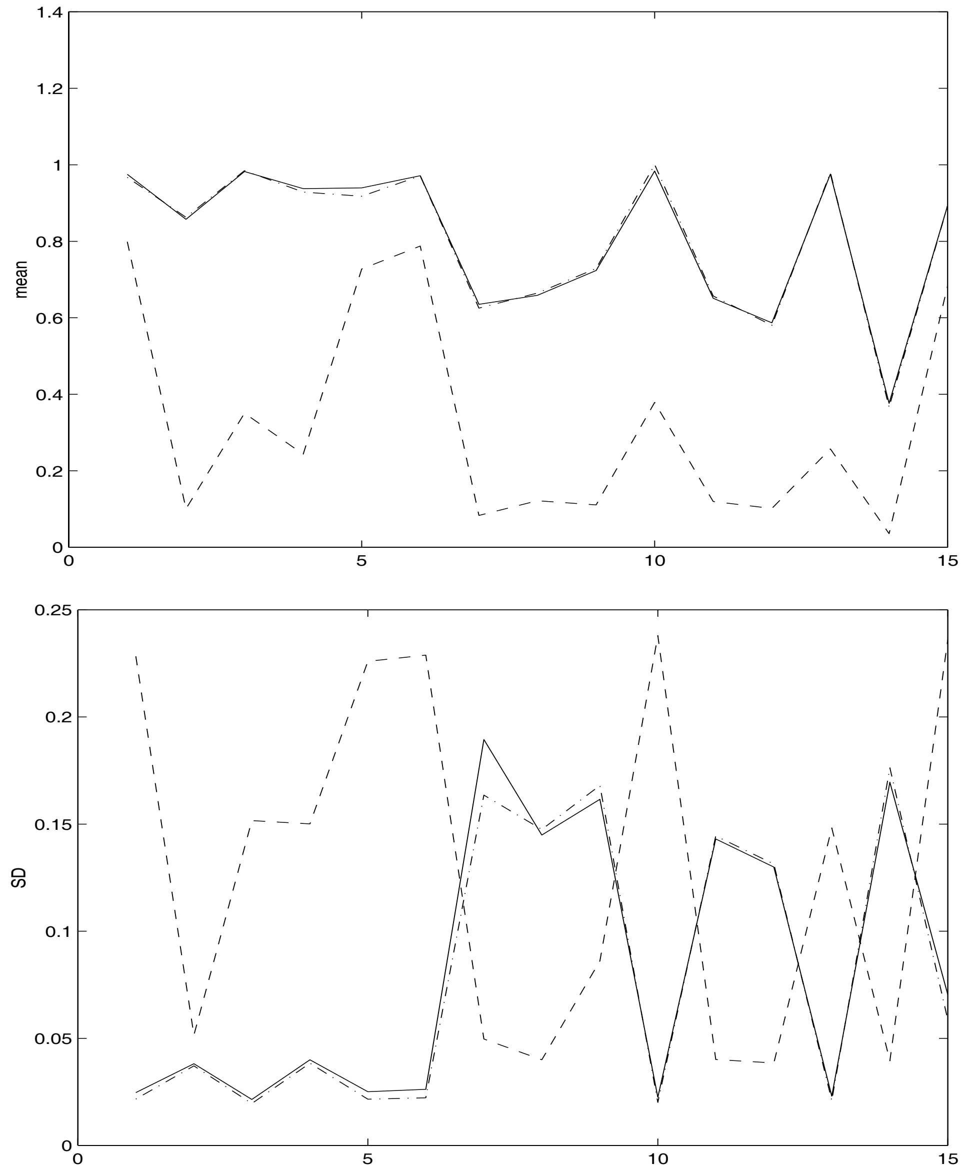
- **Blind Watermark Subspace is the subspace of W' perpendicular to $[< \mathbf{X}, \Phi_{i,j} >]$.**



Blind Watermark Encoding/Decoding : W is released

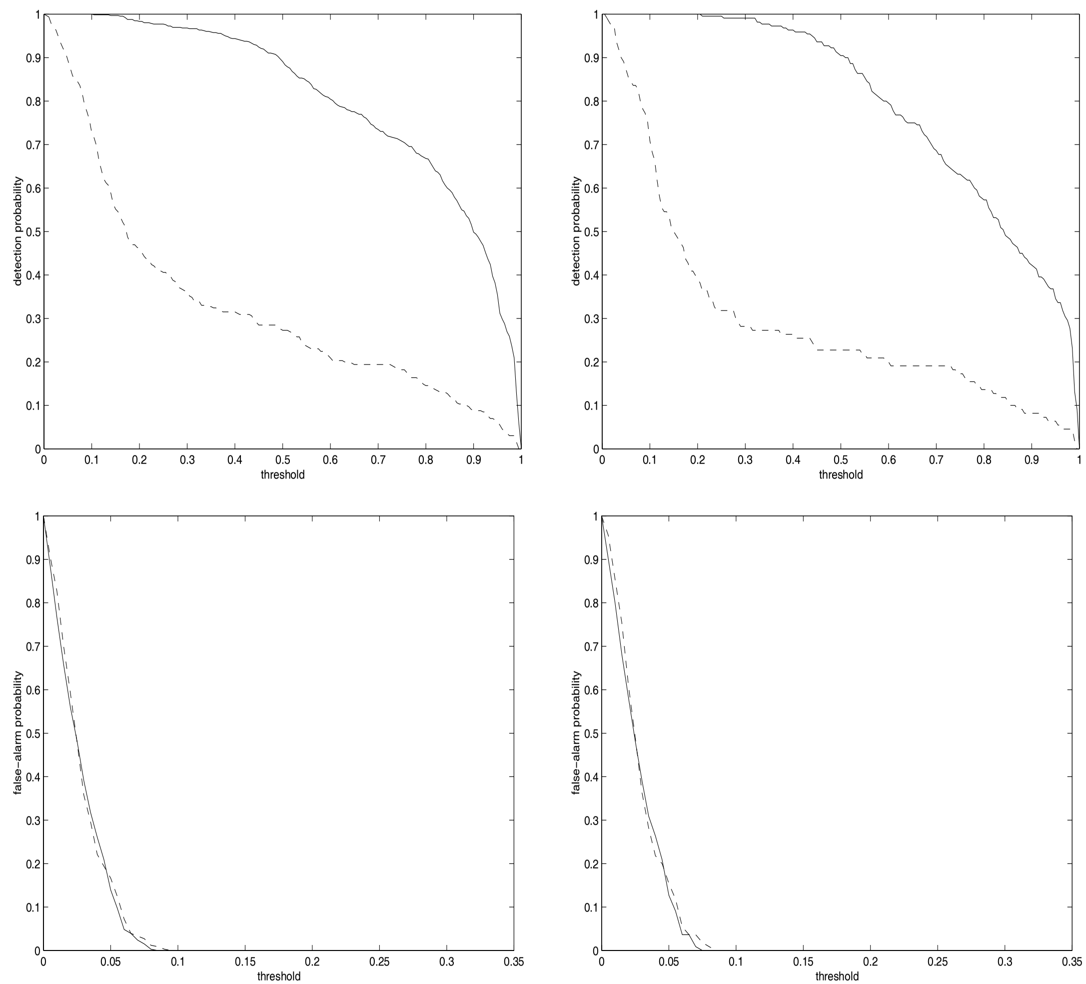
Experimental Results

- **Apply full frame DCT to a set of 22 images.**
- **Feature: Select the combinations of 32 horizontal low frequency bands and 32 vertical low frequency bands.**
- **Operations on each image :**
  - **blurring,**
  - **compression with JPEG,**
  - **small rotations (by $\pm 0.1^{\circ}$,$\pm 0.2^{\circ}$),**
  - **small translations (by shifting 1 pixel either up, down, left or right),**
  - **geometrical deformation,**
  - **adding random noise,**
  - **other image operations in Matlab and Microsoft Photo Editor.**
- **In total, we obtained 183 forged images for each image.**
- **The dimension of watermark space for each image is 900 (released to attackers).**
- **Each image has a watermark space and a blind watermark space (released to attackers).**

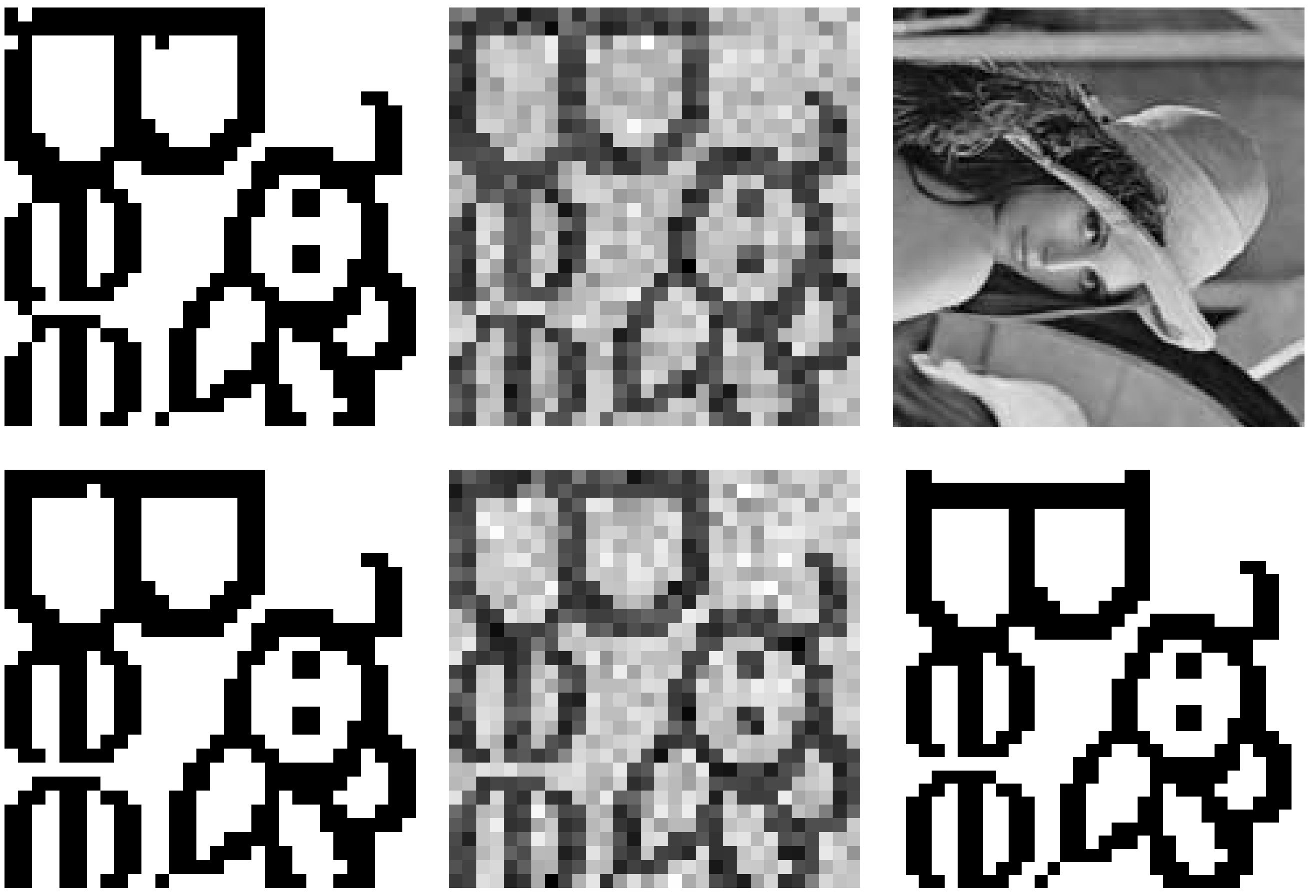

mean
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
5
10
15
SD
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15

(Continued)

- **Comparisons of the mean and standard derivation of various attacks on our methods (solid line with reference, dash-dot lines without reference) and on Cox's method (dash lines) with 22 test images.**
- **Each image was subjected to 15 attacks.**
- **The first 5 were operations that were intended to obtain our watermark space W, while the middle 5 were not, and the last 5 were combinations of attacks with one of them from 1 to 5 except for Attack 13.**
- **Attacks 1 to 5 were respectively: 1. Jpeg(60%): Jpeg Compression with a quality setting of 60%, 2. Stirmark(with small values for its parameters); 3. Small rotation 0.02^o; 4. Small translation (1 pixel in either direction); 5. Small random noise. Attacks 6 to 10 were: 6. Jpeg(53%): Jpeg Compression with a quality setting of 53%, 7. Stirmark(with larger values than Attack 2); 8. Rotation 1^o; 9. Translation 2 pixels in either direction; 10. Blur(cubic): Smooth by cubic spline. The last 5 were, respectively: 11. Jpeg 60% + Rotation 1^o; 12. Translation 1 pixel + Blur(cubic); 13. Rotate 10^o and then rotate 10^o back + blur(quadratic); 14. Stirmark(with the same parameters used in Attack 2) + Translation (2 pixels); 15. Random noise (more noise than in Attack 5) + Jpeg53%.**



The mean detection probability (top) and the mean false alarm probability (bottom) of our method (solid lines) compared with those of Cox's method (dash lines). Left: Attacks 1 to 5 are included. Right: Attacks 1 to 5 are excluded. The horizontal axes of these figures are thresholds. The false alarm probabilities are approximately the same for both methods. Given a threshold, our method has a higher mean detection probability.



Other Attacks (Continued)

- **Blind Attack: attack image**
  - **5 subjectives, knowing image processing, attacked our watermarked Lena image.**
  - **They attacked hard but kept the attacked images visually acceptable.**
  - **Obtain a total of 120 images. Among them, 85% has $sim > 0.5$, and 80% has $sim > 0.7$.**
- **Malicious Attacks - attack watermark space**
  - **Jamming our watermark space by means of spreading random noise. As much as noise but still visually acceptable.**
  - **Copy Attack - assuming that the attackers know our watermarking modulation and demodulation processes and their parameters.**

Conclusion

- **In our approach of watermark detection, the media content is not viewed purely as noise.**
- **We derive from second order statistics the watermark space for an image. The watermark space is robust to attacks and any where in the space can hide our watermark feature.**
- **Our watermarking methods are applicable to watermark detection whether a reference image is given or not.**