Multimedia Security

Watermark extraction

Mauro Barni
University of Siena
Watermark extraction: summary

• Optimum decoding/detection
  – Additive SS watermarks
• Decoding/detection of QIM watermarks
• The dilemma of de-synchronization attacks
  – Geometric attacks: the most dangerous attacks
  – Possible remedies
Detection vs decoding

• Watermark detection
  – Classical hypothesis testing problem
  – Bayes optimum decision theory
  – Neyman-Pearson criterion

• Watermark decoding
  – Classical digital communication problem
  – Optimum decoding
  – Channel coding
Watermark detection
Statistical decision theory

- Observation variable: \( y = \{y_1, y_2 \ldots y_N\} \)
  - \( y = f \) or \( f_w \) (plus noise) according to whether \( f \) contains the watermark or not
- Parameter: \( w = \{w_1, w_2 \ldots w_N\} \)
- Two possible hypotheses
  - \( H_0 \): the system does not contain \( w \)
  - \( H_1 \): the system contains \( w \)
- Test of \( H_1 \) versus \( H_0 \) which is optimum with respect to a criterion
- Bayes Theory
Bayes decision theory

- Criterion: minimum Bayes risk
- Decision rule

\[ \Phi(y) = \begin{cases} 
1, & y \in S_1 \\
0, & y \in S_0 
\end{cases} \]

- \( S_1, S_0 \) = acceptance and rejection regions for \( H_1 \)

\[ S_1 = \{ y : \frac{f(y | H_1)}{f(y | H_0)} > \frac{p_0 L_{01}}{p_1 L_{10}} \} \]
Bayes decision theory

$$S_1 = \{ y : \frac{f(y | H_1)}{f(y | H_0)} > \frac{p_0 L_{01}}{p_1 L_{10}} \}$$

$$\ell(y) = \text{Likelihood ratio}$$

$$\ell(y) = \frac{f_Y(y | H_1)}{f_Y(y | H_0)} = \frac{f_Y(y | w^*)}{f_Y(y | w = 0)}$$
Bayes decision theory

$S_1 = \{y : f(y \mid H_1) / f(y \mid H_0) > \frac{p_0 L_{01}}{p_1 L_{10}}\}$

Decision threshold: $\lambda = \frac{p_0 L_{01}}{p_1 L_{10}}$

$P_0 = P(w = 0), P_1 = P(w = w^*)$

$L_{01} = L[H_0, \Phi(y) = 1], L_{10} = L[H_1, \Phi(y) = 0]$

Letting $\lambda = 1$ minimizes the overall $P_e$
Neyman-Pearson criterion

- Bayes decision theory requires that the a-priori probabilities of $H_0$ and $H_1$ are known
  - This is rarely the case in practical applications
- Setting the cost of the two kinds of errors may be very difficult
  - In most cases extremely asymmetric costs would be needed
- It is preferable to adopt a different optimization criterion
Neyman-Pearson criterion

• Neyman-Pearson criterion: minimize $P_{MD}$ subject to a fixed false alarm rate $P_{FA} \leq P_{FA}^*$

• The same test function is adopted (the likelihood ratio) but a different threshold is used

$$P_{FA} = P[\ell(y) > \lambda \mid H_0] = \int_{\lambda}^{+\infty} f(\ell \mid H_0) \, d\ell$$
The Add-SS, AWGN case

- In order to go on a particular embedding strategy has to be considered
- Additionally, a model to describe host features and attacks is needed
- We focus on the simplest case: the AWGN channel
  - Additive SS watermarking
  - Gaussian host features
  - Additive Gaussian noise as attack
The Add-SS, AWGN case

- It is easy to show that in the AWGN case optimum detection corresponds to correlation detection

\[ \rho = \frac{1}{n} \sum_{i=1}^{n} y_i w_i \]
\[
\begin{cases}
\rho > T_\rho & H_1 \\
\rho < T_\rho & H_2
\end{cases}
\]

- The false and missed detection probabilities can be computed, and the detection threshold fixed, if the variance and mean of the host features is known

\[ f_\rho(\rho \mid 0/1) = N(\mu_{\rho\mid0/1}, \sigma_{\rho\mid0/1}^2) \]
The case of antipodal Add-SS (AWGN)

\[
f_\rho(\rho) = N(\mu_\rho, \sigma_\rho^2)
\]

\[
\left\{
\begin{array}{l}
\mu_{\rho|0} = 0 \\
\sigma_{\rho|0}^2 = \frac{\sigma_x^2 + \sigma_n^2}{n}
\end{array}
\right.
\]

\[
\left\{
\begin{array}{l}
\mu_{\rho|1} = \gamma \\
\sigma_{\rho|1}^2 = \frac{\sigma_x^2 + \sigma_n^2}{n}
\end{array}
\right.
\]
ROC curve

\[
P_m(P_f) = \frac{1}{2} \left( \sqrt{\frac{n \cdot SNR}{2}} - erfc^{-1}(2P_f) \right)
\]

\[
SNR = \frac{\gamma^2}{\sigma_x^2 + \sigma_n^2}
\]
Practical considerations

• Practical implementation of the detector requires that the variance of non marked, possibly attacked, host features is known

• By assuming the watermark strength is very low ($\gamma << 1$), such variance can be estimated on the to-be-inspected image

• It can be shown that this is equivalent to using a detector based on the correlation coefficient, leading to triangular (hyper-conic) detection regions

\[ \rho_n = \frac{y \cdot w}{||y|| \cdot ||w||} \]
Watermark decoding
Add-SS, AWGN decoding

- In this, very simple, case we have

\[ z_i = y_i + n_i = x_i + \alpha b + n_i \]

- If the host features and attack noise are normally distributed, we have an AWGN channel and optimum decoding reduces to correlation decoding. The bit error rate is

\[
P_e = \frac{1}{2} \text{erfc}\left( \sqrt{r} \frac{\alpha^2}{2(\sigma_x^2 + \sigma_n^2)} \right)
\]
QIM watermarks

- QIM is used mainly for multibit watermarking
- We consider the simplest case, i.e. DM watermarking
- In the absence of attacks decoding is straightforward and $P_e = 0$
- When attacks are present, optimum decoding (m.a.p.) goes through the calculation of $f(H_0|y)$ and $f(H_1|y)$
- For equiprobable bits, optimum decoding reduces to maximum likelihood decoding, i.e. to verifying whether $f(y|H_0)$ is larger or smaller than $f(y|H_1)$
Decoding of DM watermarks

- In general codebook entries are not equiprobable leading to the situation depicted in the figures.
- We have assumed that only an AWGN attack is present.
- The upper plot represents $f(y|H_0)$, whereas the lower one corresponds to $f(y|H_1)$. 
Decoding of DM watermarks

• If the quantization step is sufficiently small we can assume that codebook entries are locally equiprobable, leading to the situation given in the figure (top).

• In this case, optimum decoding reduces to minimum distance decoding (bottom figure).
Decoding of DM watermarkss

- When bit repetition is used, \( r \)-dimensional decoding regions are obtained, making the computation of bit error rate more cumbersome

- Minimum distance decoding is often used for sake of simplicity
Detection of QIM watermarks

- No general theory is available
- The most common approach consists in
  - Use a multibit (known) watermark
  - Decode the hidden message
  - Compare it against the known message
  - Decide on the presence of the watermark according to the hamming distance
  - The false alarm and missed detection probabilities are ruled by a binomial pdf and hence can be easily calculated once the bit error rate is known
Watermark desynchronization
Desynchronization attacks

- Robustness against DAs is a big challenge
- The watermark is not destroyed but the decoder (detector) is de-synchronized
- In most cases DAs are introduced because of geometric distortions
  - geometric distortions do not introduce significant perceptual degradations (despite the PSNR)
  - in most cases geometric distortions are non-malevolent manipulations
- We focus on the case of still images
Geometric attacks

• The list of the most common global geometric attacks include:
  – Shifting
  – Scaling
  – Cropping
  – Rotation
  – General affine transformation

• Local deformations are more difficult to cope with
  – The most famous one is the RBA implemented in the Stirmark package
Geometric attacks

• Geometric manipulations are easily understood in the spatial domain
• What does happen in the frequency domain?
  – shifting: phase change
  – resizing: zero padding
  – cropping: re-sampling
  – rotation: inverse rotation
The random bending attack (RBA)

- A random displacement field is applied to the image resulting in local geometric distortion
- In order to limit the degradation a locally smooth field is applied
- Coupled with global warping the RBA is one of the most powerful attack against any watermarking scheme
RBA: an example

Original

Stirmark 1

Example 1

Stirmark 2

Example 2
Multiresolution RBA

DF subsampled by 2

DF subsampled by 8
Multiresolution RBA

DF subsampled by 32

DF subsampled by 64
Coping with (global) geometric attacks (1)

• Use of the original image in detection

• In this case robustness against geometric attacks is more easily achieved

• The original image can be used to register the attack image prior to watermark detection/decoding

• Random local distortions may still be difficult to cope with (efficient registration schemes are needed)
Coping with (global) geometric attacks (2)

• Exhaustive search

• A set of admissible geometric distortions is defined

• The watermark is looked for by considering all admissible geometric configurations

• Problems:
  – Computational complexity
  – $P_f$ increases
  – Not applicable to multibit watermarking
Coping with (global) geometric attacks (2)

- The false detection ($P_f$) and the missed detection ($P_m$) probabilities are (Add-SS)

$$P_m^{(ES)} = P_m^{(S)} (1 - P_f^{(S)})^{n-1} \approx P_m^{(S)}$$

$$P_f^{(ES)} = 1 - (1 - P_f^{(S)})^n \approx nP_f^{(S)}$$

$P_f$ is increased by a factor $n$

$$P_m^{(S)} = \frac{1}{2} \text{erfc} \left( \sqrt{\frac{n(\gamma_w - T_\rho)^2}{2\sigma_x^2}} \right)$$

Both terms go to zero exponentially fast with $n$

$$P_f^{(S)} = \frac{1}{2} \text{erfc} \left( \sqrt{\frac{n T_\rho^2}{2\sigma_x^2}} \right)$$
Coping with (global) geometric attacks (3)

- Use of synchronization patterns (most common)
- Two watermarks are embedded
- The second (auxiliary) watermark is used to re-synchronize the detector/decoder
- Then the first watermark is looked for (decoded)

Problems:
- Additional distortion
- Loss of security
- $P_m$ may increase significantly
Coping with (global) geometric attacks (3)

• The false detection \( (P_f) \) and missed detection \( (P_m) \) probabilities are given by \( (P_{se} = \text{probability of a synch. error}) \)

\[
P_m^{(TM)} = (1 - P_f^{(S)}) P_{se} + P_m^{(S)} (1 - P_{se}) \\
P_f^{(TM)} = P_f^{(S)}
\]

• For large values of \( n \) we have

\[
P_m^{(TM)} \approx P_{se} + P_m^{(S)}
\]

\[
P_{se} \approx (\text{union bound}) \frac{n - 1}{2} \text{erfc} \left( \frac{\sqrt{n \gamma_s^2}}{4\sigma_x^2} \right)
\]

• As \( n \) increases the term \( P_{se} \) becomes predominant
Coping with (global) geometric attacks (4)

- Self-synchronizing watermarks
- Geometric properties of the watermark are exploited to recover synchronism
- Most typical case: spatial, 2D-periodic watermarks
- It improves the performance of the template matching approach

- Problems
  - Loss of security
Coping with (global) geometric attacks (5)

• Exploitation of invariants

• The watermark is inserted in a domain that is invariant with respect to geometric manipulations
  – Magnitude of DFT spectrum (shift invariance)
  – Histogram (invariant to everything but cropping)
Example of geometric invariance

By embedding the watermark at a fixed position in the frequency spectrum, invariance to scaling is automatically achieved (M. Barni, F. Bartolini V. Cappellini and A. Piva, “A DCT-domain system for robust image watermarking”, *Signal Processing*).
Example of geometric invariance

By always extending the image to the same size through zero padding (and by exploiting the translation invariance of the magnitude of DFT) the watermark can be recovered even in presence of moderate cropping.

marked coefficients

Marked spectrum

marked coefficients

Spectrum of cropped signal

marked coefficients

Spectrum of cropped and zero-padded signal
Example of geometric invariance

- J. O Ruanaidh et al., “Rotation, scale and translation invariant digital image watermarking”, *ICIP’97*, S.Barbara, 1997;
- The watermark survives resizing only if aspect ratio is preserved
- As to cropping, spectrum resampling must be considered
Coping with (global) geometric attacks (6)

• Feature-based geometric normalization

• The watermark is always inserted in the same geometric configuration

• The reference geometric configuration is tied to the image content

• Problems
  – Robustness ultimately depends on feature robustness and stability
Feature-based normalization

Extract the edges from the image, then compute the central inertial axis and the central moments of edges
Feature-based normalization

Deform the image so that central inertial axis and central moments of edges assume a reference value, then insert the watermark and go back to original geometry.
Coping with (local) geometric attacks

• Why the other solutions do not work:
  – original image: actually it works
  – exhaustive search: no good model, exponential size of DA class
  – geometric invariance: does such a domain exist?
  – template matching: as for ES
  – self-synchronizing watermarks: as for ES
  – feature-based normalization: insertion and deletions

• Open problem: available solutions apply the previous techniques locally
  – example: movies camcorded in theaters
References

• I. J. Cox, M. Miller, J. Bloom, *Digital watermarking*, Morgan Kaufmann