EE 6886: Topics in Signal Processing
-- Multimedia Security System

Lecture 5: Security Attacks

Ching-Yung Lin
Dept. of Electrical Engineering
Columbia University, New York, NY 10027

Outline -- Introduction

- Multimedia Security :
  - Multimedia Standards – Ubiquitous MM
  - Encryption – Confidential MM
  - Watermarking – Uninfringible MM
  - Authentication – Trustworthy MM
- Security Applications of Multimedia:
  - Audio-Visual Person Identification – Access Control, Identifying Suspects
  - Surveillance Applications – Abnormality Detection
  - Media Sensor Networks – Event Understanding, Information Aggregation

Major Issues in Watermarking

- Invisibility:
  - Least-Significant Bits
    - Spatial Domain
    - Compression-Compliant Block-Frequency Domain
    - Global Frequency Domain
  - Human Perceptual Models
    - Domain-Specific Models
    - Generic Models
- Robustness:
  - Lossy Compression
  - Format Transformation
  - Scaling, Translation, Cropping
  - Rotation, Scan-and-Print
- Embedding Information Payload:
  - Information Theory
  - Writing on Dirty Paper
  - Zero-Error Embedding Capacity
- Security:
  - Attacks

Security Requirements

Security Requirements Depend on Applications – Example (1)

- Scenario 1: Alice is an advertiser who embeds a watermark in each of her radio commercials before distribute them to 600 radio stations.
  - Alice monitors radio station broadcasts with a watermarking detector.
  - She matches her logs with the 600 invoices.
  - [Attack]:
    - Bob secretly embed Alice's watermark into his own advertisement and airs it in place of Alice's commercial.

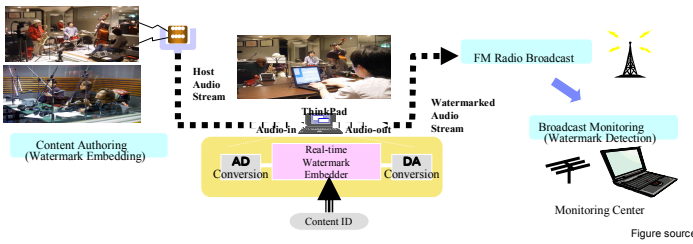


Figure source: Dr. Tachibana, IBM TRL

➔ Unauthorized Embedding / Forgery Attack

Security Requirements Depend on Applications – Example (2)

- Scenario 2: Alice owns a watermarking service that, for a nominal fee, adds an owner identification watermark to images that will be accessed through the Internet.
  - Alice provides an expensive reporting service to inform her customers of all instances of their watermarked images found on the Web.
  - [Attack]: Bob builds his own web crawler that detects watermarks embedded by Alice and offers a cheaper reporting service.

➔ Unauthorized Detection / Passive Attack

Security Requirements Depend on Applications – Example (3)

- Scenario 3: Alice owns a movie studio, and she embeds a copy-control watermark in her movies before they are distributed.
  - She trusts that ll digital recorders capable of copying these movies contain watermark detectors and will refuse to copy her movie.
  - [Attack] Bob is a video pirate who has a device designed to remove the copy protection watermark

➔ Unauthorized Removal

Operational Table of the Three Scenarios

	Embed	Detect	Remove
Broadcast Monitoring			
Advertiser	Y	Y	-
Broadcaster	N	N	-
Public	N	N	-
Web Reporting			
Marking Service	Y	Y	-
Reporting Service	-	Y	-
Public	N	N	N
Copy Control			
Content Provider	Y	Y	-
Public	-	Y	N

Y: must be allowed, N: must not be allowed, - : don't care

Assumption about the Adversary

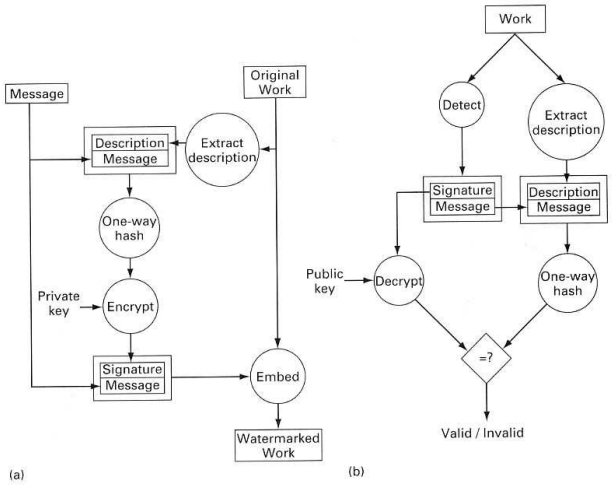
- If the attacker knows nothing:
- If the attacker has more than one watermarked work
- If the attacker knows the algorithm
- If the attacker has a detector

Categories of Attack (1)

- Unauthorized Embedding:
  - Being able to composing and embedding an original message..
    - Another example, in Scenario 2, Alice charges for embedding and gives away the monitoring tool..
    - Possible Solution: using standard cryptographic techniques.
  - Being able to obtain a pre-composed legitimate message and embeds this message in a Work.
    - E.g., in Scenario 1, Bob extract the reference pattern and then use it to his work – called *copy attack*.
    - Possible Solution: using content-related watermarks.

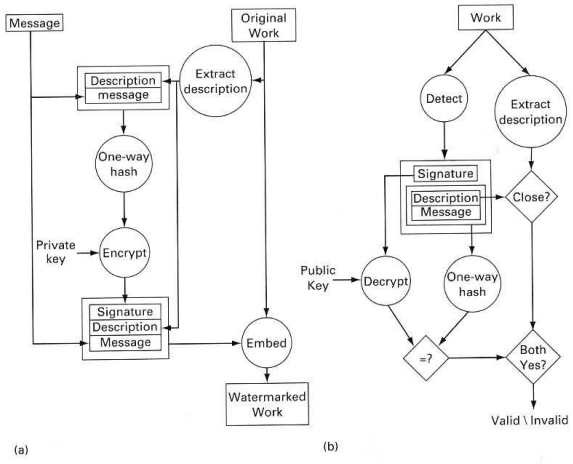
Methods to Prevent Unauthorized Embedding (1)

- Make the embedding codes:
  - Content dependent
  - Signer dependent



Methods to Prevent Unauthorized Embedding (2)

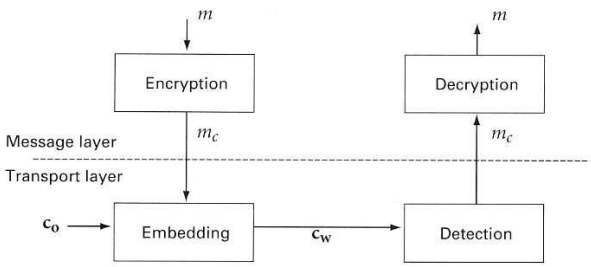
- Another method can be used in the detector.



Categories of Attack (2)

- Unauthorized Detection:
  - A hospital might embed the names of patients into their X-rays.
  - Knowing whether or not a watermark is present → Steganography.
  - Intervention on the transmission process.

Methods to Prevent Unauthorized Detection



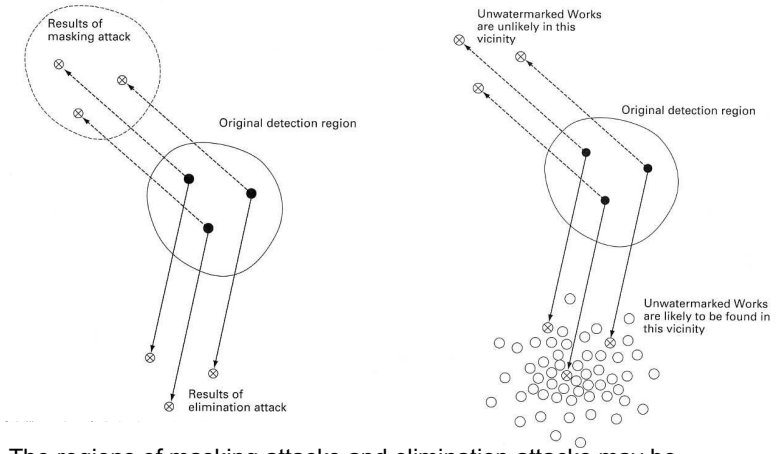
- Encryption and Decryption can be used.

Categories of Attack (3)

- Unauthorized Removal:
  - Attackers try to modify the watermarked Work such that it resembles the original and yet does not trigger the detector.
  - Two types of attacks:
    - Elimination attacks ➔ The watermark is truly gone.
    - Masking attacks ➔ The watermark is still present but is weakened.

Unauthorized Removal

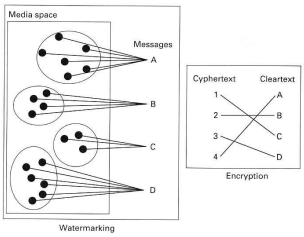
- Masking Attack and Elimination Attack



➔ The regions of masking attacks and elimination attacks may be predictable

Methods to Prevent Unauthorized Removal

- Spread Spectrum Techniques are suggested.
- One-line of researching is based on the belief that watermarking can be made secure by creating something analogous to asymmetric-key encryption. ➔ The detection key is not sufficient to remove a watermark. ➔ May not survive sensitivity analysis.
- There are some fundamental differences between watermarking and cryptography that make the standard asymmetric-key encryption systems unsuitable.
  - In watermarking, the mapping between Works and messages must be many-to-one, so that a given message may be embedded in any given Work.
  - In asymmetric-key cryptography, the mapping between cleartext and ciphertext is always one-to-one.
  - In watermarking, small changes in the Works should map to similar messages.
  - In assymmetric-key cryptography, a small change in cleartext results in large change in the ciphertext.



Categories of Attack (4)

- System-level Attacks:
  - Attackers exploit the weakness in how the watermarks are used.
  - For instance, in a copy-control application, an attaker might open the recorder and just remove the chip.
  - Forge identification.

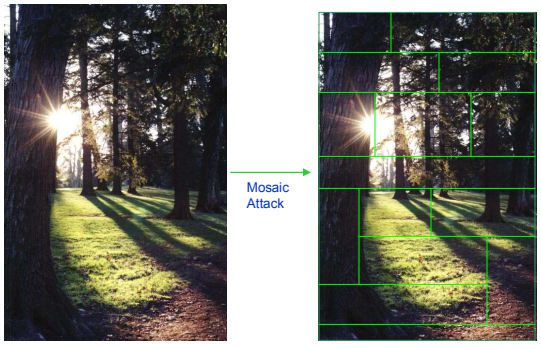
Some Significant Known Attacks

- Scrambling attacks
- Pathological distortions:
  - Synchronization attacks
  - Linear filtering and Noise Removal Attacks
- Copy attacks
- Ambiguity attacks
  - Ambiguity attacks with informed detection
  - Ambiguity attacks with blind detection
- Sensitivity analysis attacks
- Gradient descent attacks

Scrambling Attack

- System-level attack
  - An additional device is applied to scramble watermarked multimedia work to make the watermark undetectable by the detector.
  - Using a descramble device to invert the work.
- Example:
  - *Mosaic Attack:* partition the watermarked image into several individual smaller images that are organized with table when displayed.
- Effectiveness: avoid on-line image crawling

➔ Solutions?



Pathological Distortions (I)

- Synchronization Attacks:
  - Most watermarking techniques are sensitive to synchronization
  - Audio and Video: delay and time scaling
    - Pitch-preserving scaling
    - Sample removing
  - Image and Video: rotation, scaling and translation
    - Shearing
    - Horizontal reflection
    - Column or line removal
    - Nonlinear warping
  - Some of these attacks are applied by the StirMark – a watermark benchmarking system.

Pathological Distortion (II)

- Linear Filtering and Noise Removal Attacks:
  - May be effective while many watermarking system embed significant energy in the high frequencies.
  - Wiener filtering is an optimal linear-filtering/noise-removal attack. It is effective when:
    - The added pattern is independent of the work.
    - Both the work and the watermark are drawn from zero-mean Gaussian distribution.
    - Linear correlation is used as the detection statistic.
  - The security of a watermark against Wiener filtering can be maximized by selecting the power spectrum of the added pattern to be a scaled version of the power spectrum of the original work, as

$$|W_a|^2 = \frac{\sigma_{w_a}^2}{\sigma_{C_0}^2}|C_0|^2$$

Power spectrum of the watermark

Variations of the distribution of the watermark pattern and the work

Power spectrum of the work

Copy Attack

- An adversary copies a watermark from one work to another. It is a form of unauthorized embedding.
- Example: (Kutter et al., 2000) given a legitimately watermarked work, c_{1w}, and an unwatermarked target work, c_2, this method begins by
  - applying a watermark removal attack to c_{1w} to obtain an approximation of the original, c_1', by using a nonlinear noise-reduction filter.
  - Estimate the added watermark pattern by subtracting the estimated original from the watermarked work:

    $$w_a' = c_{1w} - c_1'$$
  - The estimated watermark pattern is added to the unwatermarked work:

    $$c_{2w} = c_2 + w_a'$$

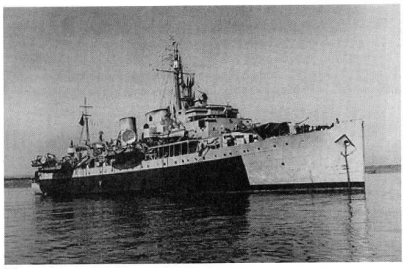
➔ Solutions?

Ambiguity Attacks

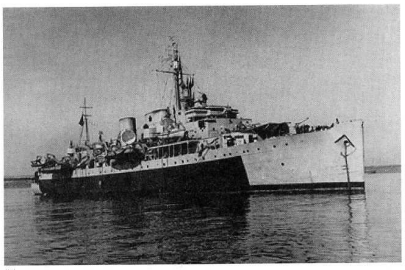
- Ambiguity attacks (or called the Cover attack, *Craver et al., 1998*) create the appearance that a watermark has been embedded in a work when in fact no such embedding has taken place.
- Objectives: claiming false ownership.
- Two situations:
  - ambiguity attacks with informed detection
  - ambiguity attacks with blind detection

Ambiguity Attacks with Blind Detection

- Examples of Ambiguity Attack: (a) True original Image, (b) Distributed Watermarked Image.

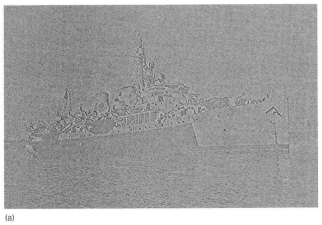


(a)

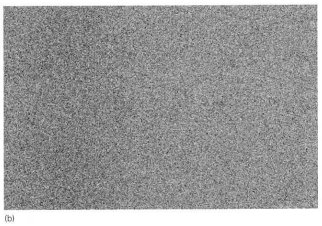


(b)

Ambiguity Attacks with Blind Detection

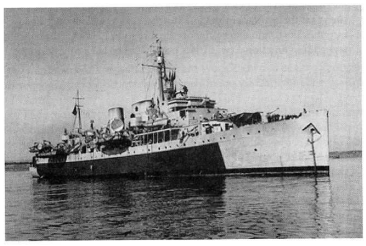


(a)



(b)

- Ambiguity Attack (a): Adding some random noise into the Fourier phase; (b) Add noise to the image and then scale Fourier coefficients with random magnitude changes



Faked original image constructed by subtracting 99.5% of the fake reference pattern

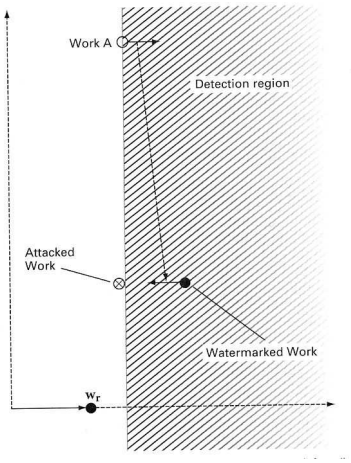
➔ How to defeat ambiguity attacks?

Defending Ambiguity Attacks

- The true owner of the Work uses a watermarking technique that can ensure that his original could not have been forged.
- Invertibility: a watermarking scheme is invertible if the inverse of the embedding is computationally feasible.
- Ambiguity attacks cannot be performed with non-invertible embedding techniques. For instance, the reference pattern should be dependent on the content of the original work.

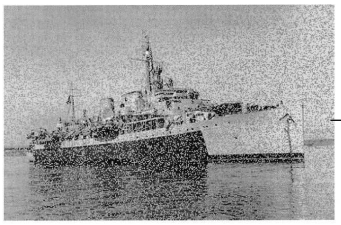
Sensitivity Attack (I)

- A technique used for unauthorized removal of a watermark when the adversary has a black-box detector.
- Three steps:
  - First, find a work that lies very close to the detection region boundary. It need not be perceptually similar to the watermarked work (e.g., distortions).
  - Second, approximation of the direction of the normal to the surface of the detection region at Work A.
  - Third, estimate the N-dimensional normal vector (Linnartz and van Dijk, 1998)



Sensitivity Attack (II)

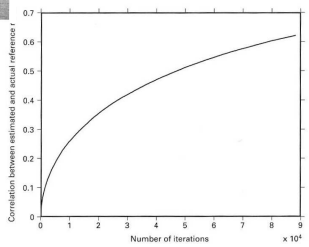
- Examples

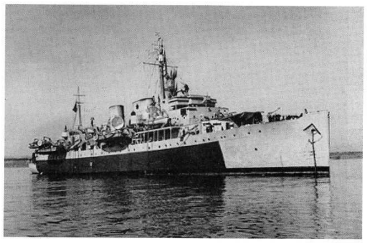


An artificial image A

Iterative removal

Watermarked work

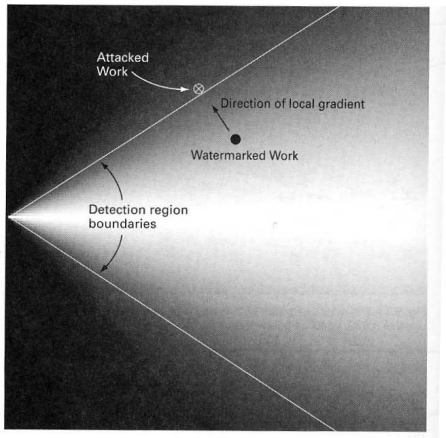




After watermark removal

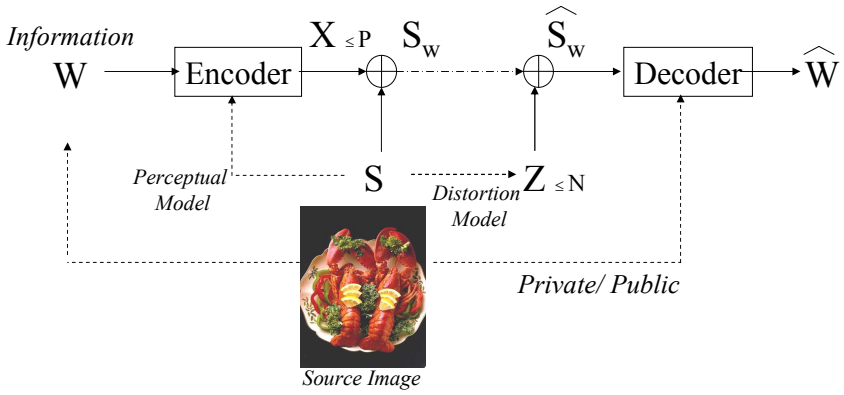
Gradient Descent Attack

- The attacker knows the actual detection values of detectors.
- Similar to the scenario of sensitivity attack.



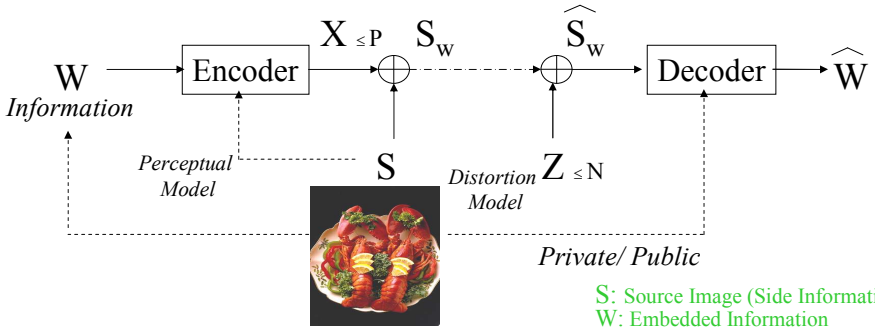
Information Hiding Capacity

Watermarking on Multimedia Content



S: Source Image (Side Information)
W: Embedded Information
X: Watermark (Power/Magnitude Constraint: P)
Z: Noise (Power/Magnitude Constraint: N)

Information Hiding Capacity



S: Source Image (Side Information)
W: Embedded Information
X: Watermark (Power Constraint: P)
Z: Noise (Power Constraint: N)

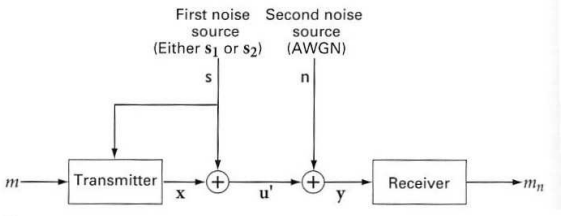
- Channel Capacity:
  Shannon (1948) (private watermarking)

$$C = \tfrac{1}{2} \log_2 (1 + P/N) \quad \text{bit/sample}$$

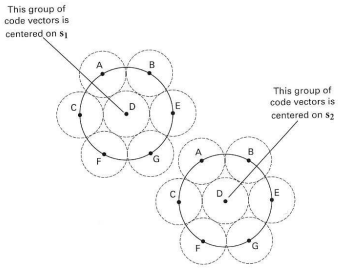
- Assumptions:
  - Uniform Power Constraints on Watermark and Noise
  - An image is a channel

-- Is this the final answer or just a beginning?

Writing on Dirty Paper



- Embedding watermarks on images is similar to writing codes on a dirty paper.
- Costa (1983): with or without the original data at the detector, the channel capacities are the same.



Watermarking Capacity based on Legend Works

- Assumptions:
  - Uniform Power Constraints on Watermark and Noise
  - An image is a channel
- Channel Capacity:

Shannon (1948) (for private watermarking) , Costa (1983) (for public watermarking)

$$C = \tfrac{1}{2} \log_2 (1+ P/N) \quad \text{bit/sample}$$

Why Costa got the same hiding capacity regardless of the existence of the source signal?

➔ Information is shifted from the modulation coefficients to the content-dependent modulation bases.

Watermarking Capacity based on Non-uniform Power Constraints

Previous Propositions:

- Image as parallel channels?
  - Parallel Gaussian Channels – Akansu (1999), Servetto (1998), Kundur
    -- Possible Drawback: A channel needs infinite codeword length !!

Image as one channel

- Arbitrary Varying Channel (AVC, *Csiszar 1989*) – Possible Drawback: arbitrary varying

We consider:

- Image as a variant-state channel
- Image coefficient values are discrete

If not power-constraint, but amplitude constraints on noises,

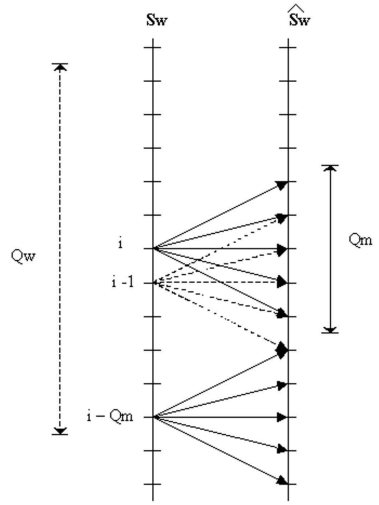
➔ Zero-Error Watermarking Capacity

Image as Communication Channel(s)

An M×N digital image can be considered as

- *Case 1*: a *variant-state discrete memoryless channel* (DMC). Transmission utilizes this channel for *M×N times*.
- *Case 2*: a mixture of Case 1 and 3. If an image is divided into B blocks with K coefficients in each block, then this image can be considered as *B parallel channels with K transmissions* in each channel.
- *Case 3*: a product of M×N *static-state* DMCs, in which each coefficient forms a DMC. Each channel can be at most transmitted *once*.

Adjacency-Reducing Mapping of Discrete Values given Bounded Noises



Adjacency-reducing mapping
two input nodes are adjacent if there is a common output node which can be caused by either of these two.
-- Shannon *The zero-error capacity of a noisy channel*, Trans. on IT, 1956)

$$C(Q_w, Q_m) = \log_2 (\lfloor Q_w / Q_m \rfloor + 1) \text{ bits}$$

➔ *A bound for private/public watermarking*

Zero-Error Capacity for Various Cases

A Mimimum bound for public watermarking

$$C'(Q_w, Q_m) = \log_2 (\lfloor \max(Q_w - Q_m , 0) / Q_m \rfloor + 1) \text{ bits}$$

Capacity bound for Case 2

$$C = \lfloor B \times \Sigma C'(Q_w, Q_m) \rfloor \text{ bits}$$

Capacity bound for Case 3

$$C = B \times \Sigma \lfloor C'(Q_w, Q_m) \rfloor \text{ bits}$$

Watermarking using Exactly Reconstructable DCT Coefficients

Fp(i,j) : The original DCT coefficient at the position (i,j) of block p

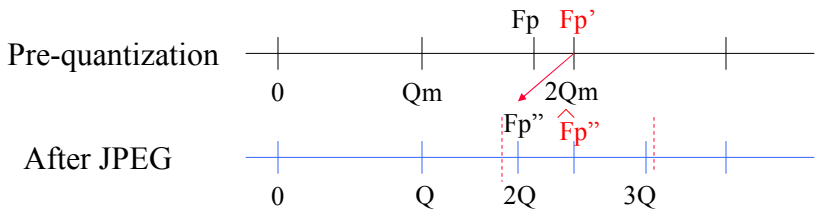
Fp'(i,j) : The pre-quantized DCT coefficient by Qm(i,j)

Fp"(i,j) : The quantized result of Fp'(i,j) after lossy compression using Q(i,j).

- Theorem 2: For all $Q(i,j) \leq Qm(i,j)$

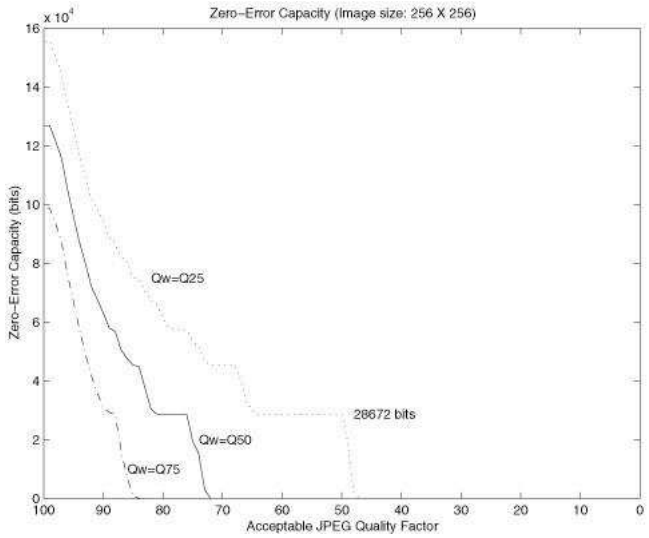
  Fp"(i,j) ≡ Integer Rounding[Fp"(i,j) / Qm(i,j)] Qm(i,j)

  = Fp'(i,j)

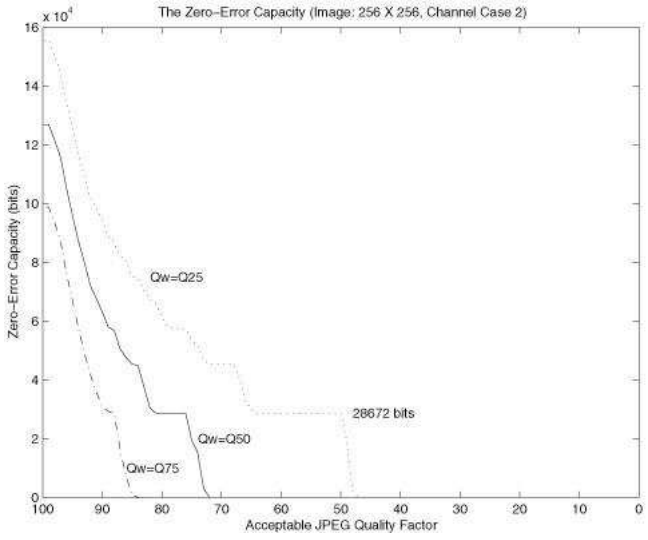


capacity for watermark embedding.

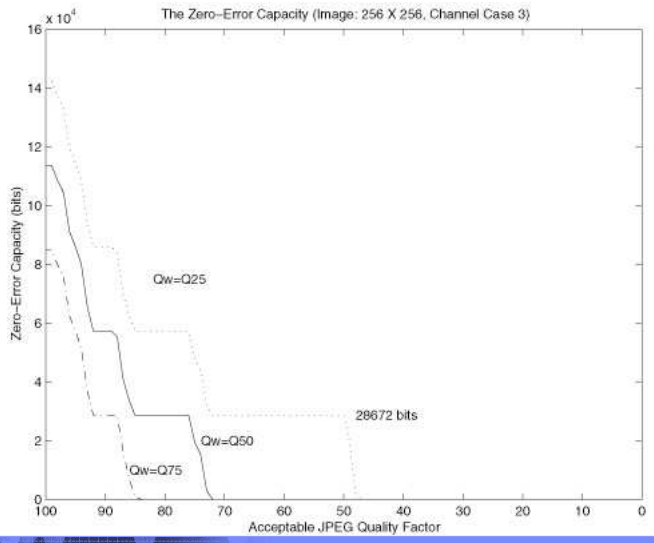
Zero-error capacity of amplitude-constrained noisy environments



Zero-error capacity of amplitude-constrained noisy environments



Zero-error capacity of amplitude-constrained noisy environments



Example of watermark embedding achieving zero-error capacity

(a) (b)

(a) Original image (256x256); (b) watermarked image with 28672 hidden bits surviving JPEG QF=75 compression.

Reference Papers

- C.-Y. Lin, "Digital Signature and Watermarking Techniques for Multimedia Authentication and Copyright Protection," *Columbia Ph.D. Thesis*, Chapter 4 and 5, Dec. 2000.
- I. J. Cox, M. L. Miller and J. A. Bloom,, "Digital Watermarking," *Morgan Kaufmann*, Chapters 3, 5, 9, 2001.