Principles of Digital Image and Video Watermarking

Ton Kalker

Philips Research Eindhoven

ton.kalker@ieee.org

adapted from ICIP-2000 tutorial
with contributions from Jonathan Su

Outline

- Introduction
- Spread-spectrum watermarking
- Attacks and robustness
- De- and re-synchronization
- JAWS & Millennium
- Millennium System Aspects

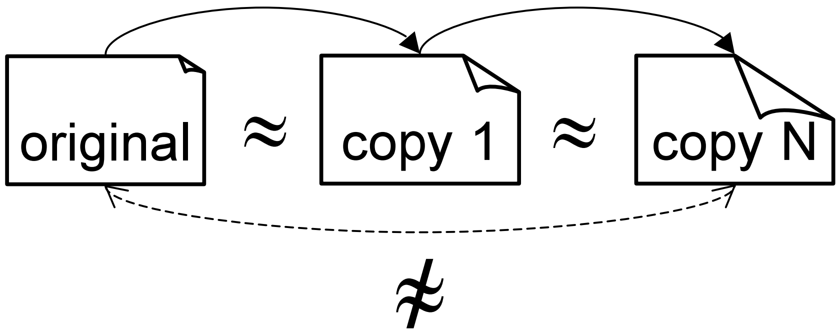
INTRODUCTION

- Motivation
- “How can information be hidden in digital data?”
- “What is the watermark?”
- Watermarking as communications
- Desired properties
- Limitations

Analog and Digital Multimedia

Analog Media

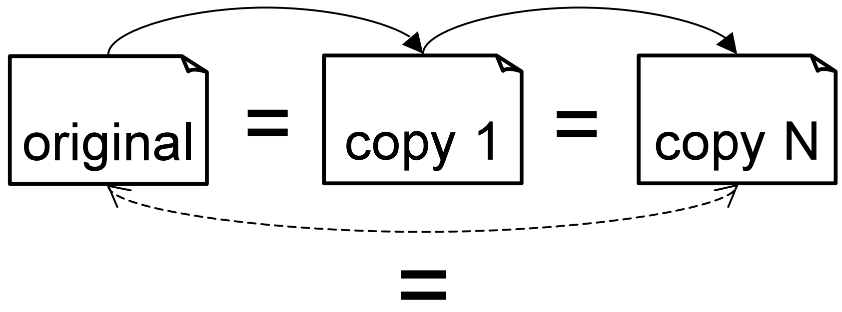
photocopies
audio cassettes
photographs
VHS videotapes



- “Built-in” protection against copying and redistribution
- Distribution net required

Digital Media

ASCII, PostScript, PDF
CDs, MP3 audio
JPEG images
DVDs, MPEG video



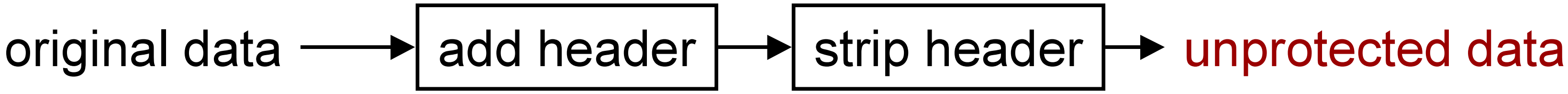
- **No inherent protection** against copying and redistribution
- **“Free” distribution net**: Internet

Unauthorized Use of Digital Data

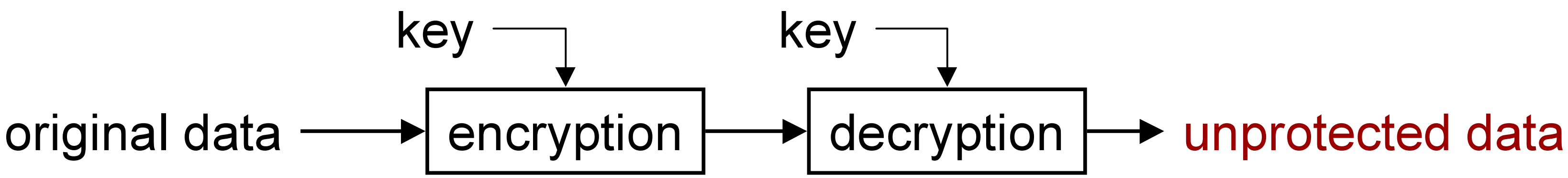
- Digital multimedia
  - can be stored, copied, and distributed easily, rapidly, and with no loss of fidelity
  - can be manipulated and edited easily and inexpensively
- Are these properties always advantageous?
  - Some Hollywood studios will not release DVDs unless copyright protection can be ensured
  - USA Today, Jan. 2000: Estimated lost revenue from digital audio piracy: US$8,500,000,000.00
  - Recent examples: MP3.com, Napster

Traditional Methods of Protecting Data

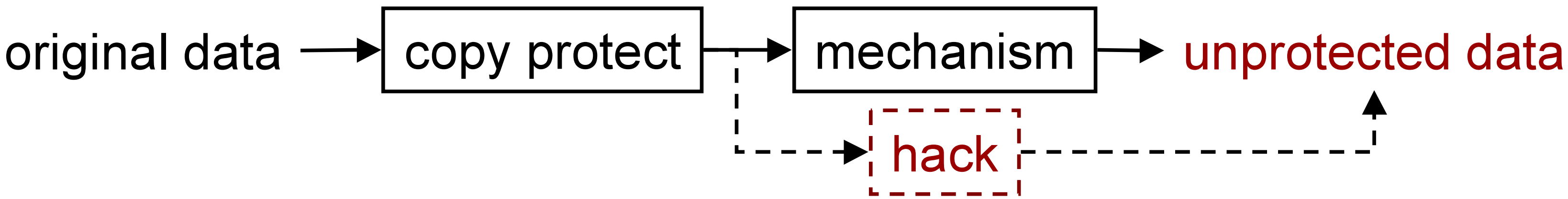
- Access-control headers: easily removed/altered



- Encryption: decrypted data unprotected



- Copy protection: susceptible to hacking

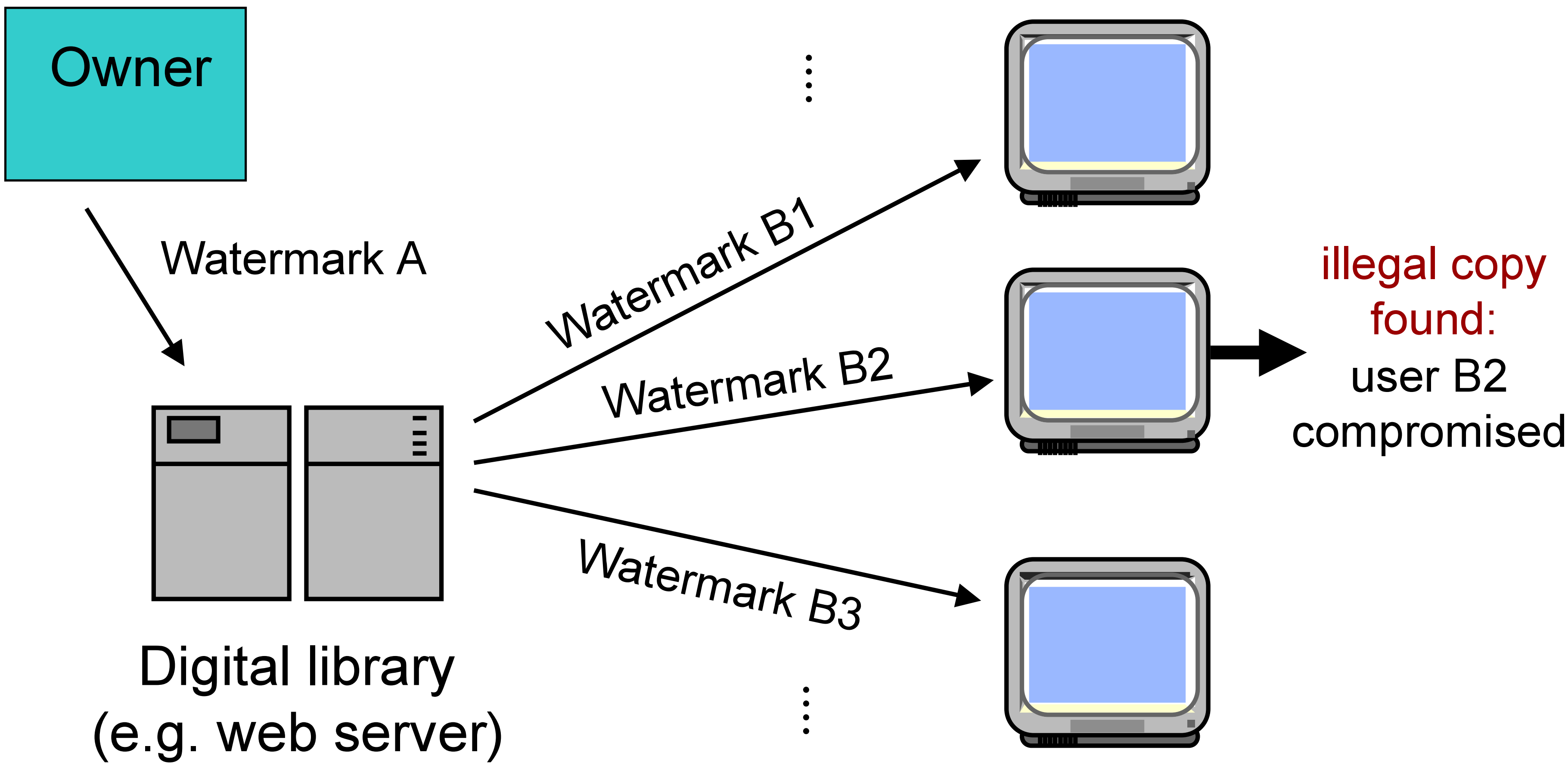


Motivation for Digital Watermarking

- Imperceptibly embed information directly into original data ("host data", "cover data") to produce "watermarked data"
- Principle: Embedded information travels with the watermarked data, even after copying and redistribution

  - "last line of defense"
  - loosely analogous to watermarks in paper

Example: Distribution from a Library

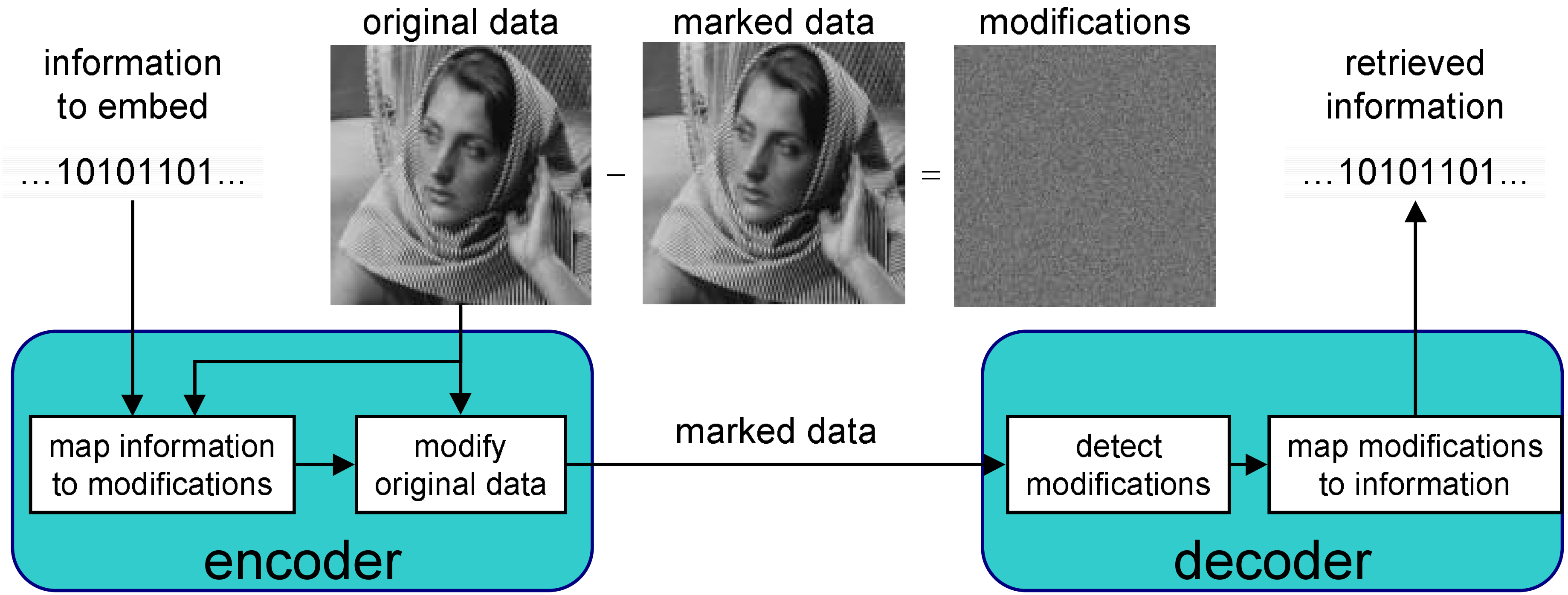


Watermarking Applications

- Access control
  - playback, copy-generation control (DVD)
  - copyright protection, proof of ownership (?)
- Distribution tracing
  - fingerprinting
  - identification of compromised parties
- Broadcast monitoring
- Media authentication (fragile watermarking)
- Covert communication (steganography)
- Added value via meta-information
  - e.g., SmartImages by Digimarc Corp. [Alattar 2000]

Two Basic Questions

- How can information be hidden in digital data?
- What is the watermark, actually?



"How can information be hidden in digital data?"

- By exploiting "perceptual headroom."
  - human perception is imperfect
  - make modifications to the original data without changing it perceptually
  - modifications can be detected via signal processing

"What is the watermark, actually?"

- No standard definition, two common viewpoints
  - "watermark-as-signal"
    - watermark = modifications to original convey information
    - applies regardless of implementation details (e.g., spatial- or frequency-domain methods)
  - "watermark-as-information"
    - watermark = information that is embedded and retrieved

watermark-as-signal

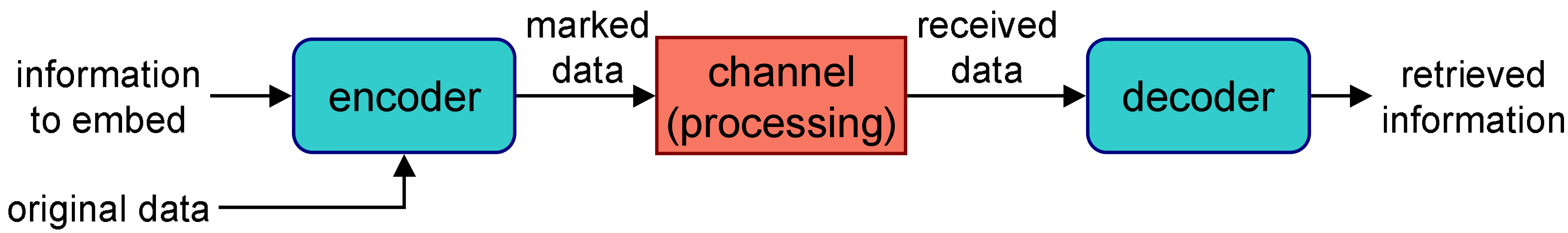
original data   marked data

watermark-as-information

information to embed		retrieved information
…10101101…	→	…10101101…

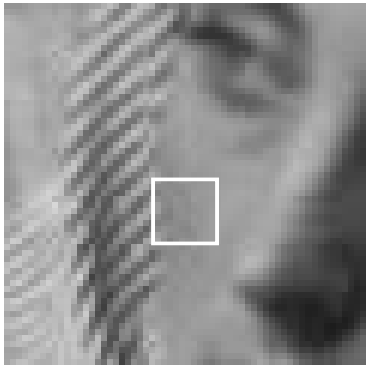
"What is digital watermarking?"

- Watermarked data is likely to be processed
  - view processing as a communications channel
- Digital Watermarking: The *imperceptible, robust, secure communication* of information by embedding it in and retrieving it from other digital data.



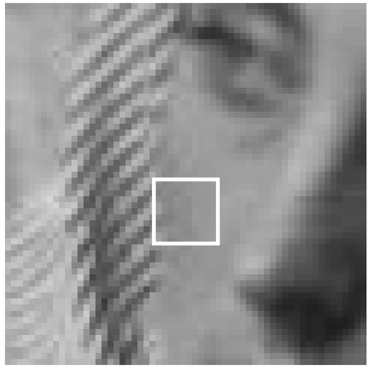
Desired Properties: Imperceptibility

- Watermarked data and original data should be perceptually indistinguishable
- Use low-amplitude modifications and/or perceptual modeling



```
115 154 180 …
158 183 174 …
177 168 144 …
```

Original image



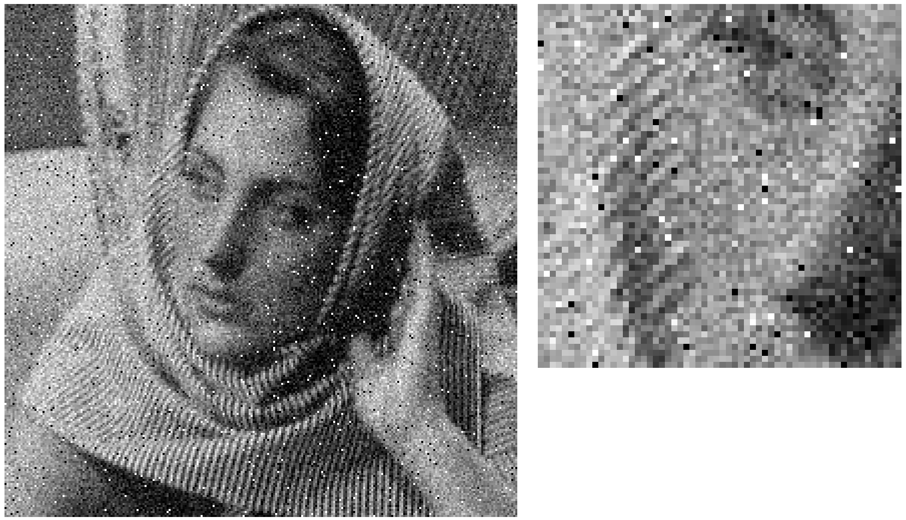
```
114 150 180 …
156 186 172 …
177 170 144 …
```

After embedding

Desired Properties: Robustness

- Processing of the watermarked data cannot damage or destroy the embedded information without rendering the processed data useless

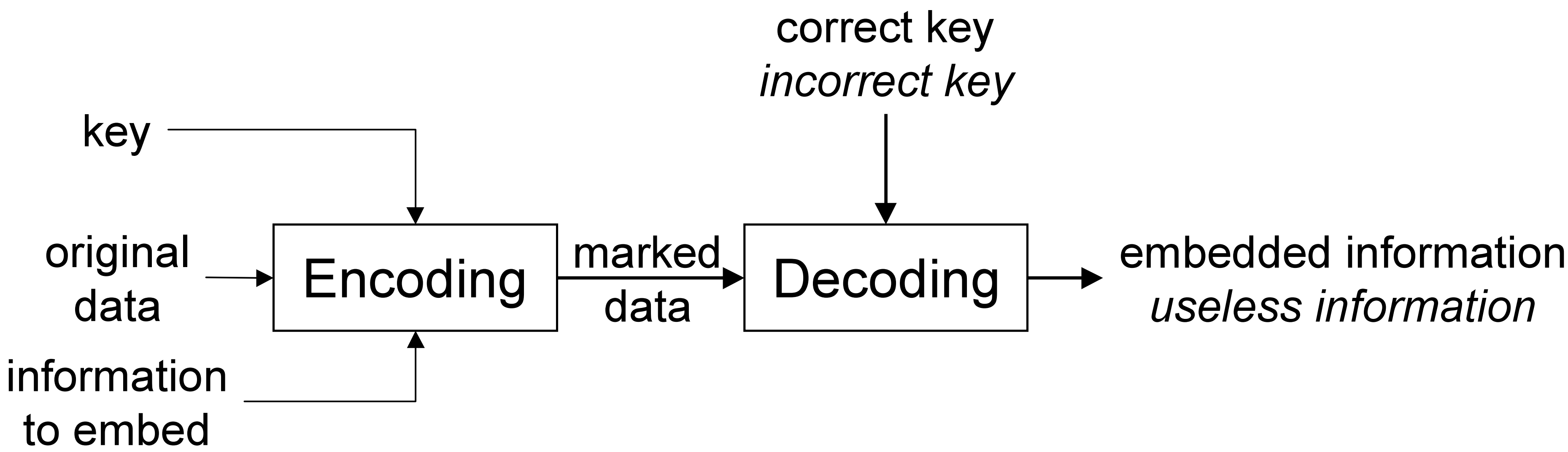
JPEG compression



Additive noise & clipping

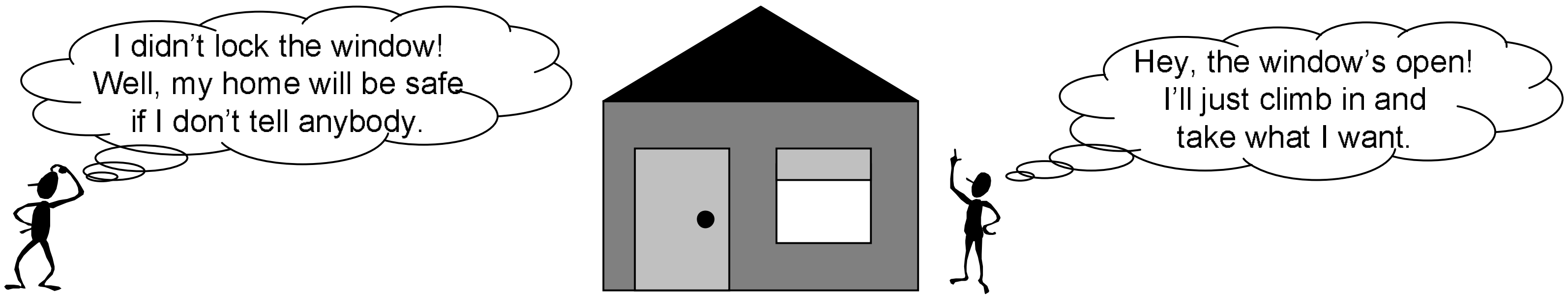
Desired Properties: Security

- Embedded information cannot be detected, read, and/or modified by unauthorized parties
- Kerckhoff's principle: Security resides in the secrecy of the key, not in the secrecy of the algorithm.



Kerckhoff's Principle

- Security resides in the secrecy of the key, <u>not</u> in the secrecy of the algorithm
- Assume your opponent has complete knowledge of your strategy but lacks a secret.
  - strategy = algorithm & implementation
  - secret = key
- Otherwise: <u>False sense of security</u>!



Additional Desired Properties

- "Blind" watermarking
  - no reference to original data during decoding
  - possible interference from original data
- Multiple watermarks
  - one copy with several information streams
  - different information in different copies
- Compressed-domain processing
  - combined watermarking and compression
  - bit-rate constraint
- Implementation concerns
  - speed, computational load, footprint, cost

Additional Desired Properties

- Low False Positive Rate
  - a positive detection on non-marked content
- Low Granularity
  - minimal spatio-temporal interval for reliable embedding and detection
- Large Capacity
  - related to payload
  - #bits / sample
- Layering & Remarking Capabilities
  - watermark modification

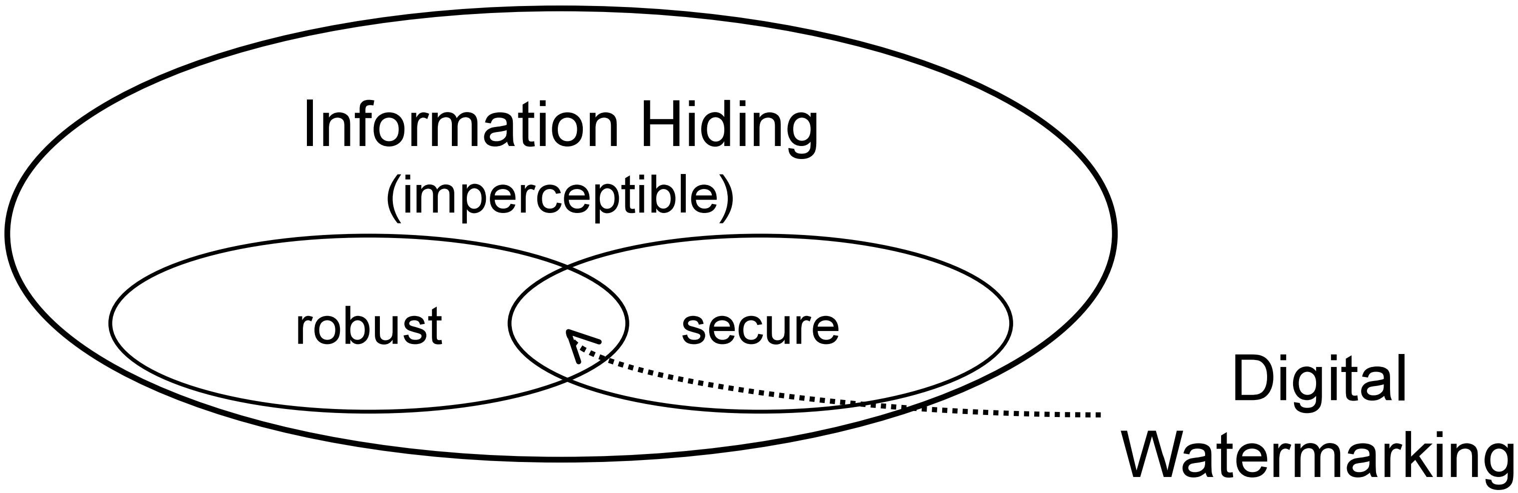
Relation to Information Hiding

- Information Hiding (steganography)

  The imperceptible communication of information by embedding it in and retrieving it from other digital data.

- Digital Watermarking

  The imperceptible, **robust**, **secure** communication of information by embedding it in and retrieving it from other digital data.



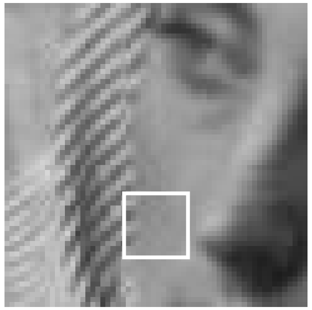
Other Types of Watermarks

- Imperceptible → Perceptible but unobtrusive
  - closer analogy with paper watermarks
  - less robust? since watermark is easily located
- Robust → Fragile
  - watermark should “fail” even after slight modifications to watermarked data
  - applications: media authentication, tamper detection

Low-bit Modulation: Not Watermarking

- Early scheme: alter LSB or low-order bits

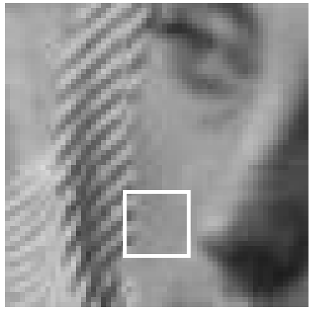
Original



```
130=1000 0010
123=0111 1011
117=0111 0101
  :     :     :
```

Embed
101…

After embedding



```
131=1000 0011
122=0111 1010
117=0111 0101
  :     :     :
```

✓ imperceptible (modify only LSBs)

✓ secure (encrypt embedded information)

not robust (e.g., randomly set LSBs to 0 or 1)

- More accurate: secure info-hiding method

Limitations

- Digital watermarking does <u>not</u> prevent copying or distribution
  - (but embedded information remains in copied data)
- Digital watermarking alone is <u>not</u> a complete solution for access/copy control or copyright protection!
- Digital watermarking is a <u>part</u> of a larger system for protecting digital data against unauthorized use

SPREAD-SPECTRUM WATERMARKING

- Principle
- Relation to watermarking
- Direct-sequence spread spectrum
- Possible drawbacks

Spread-Spectrum Principle

- Transmit information via pseudo-random modulation that uses a (much) larger bandwidth than the minimum necessary
- Common techniques
  - direct-sequence spread spectrum
    - multiply information bits directly by a “spreading sequence”
    - statistically, spreading sequence resembles white noise
  - frequency-hopping spread spectrum
    - rapidly change carrier frequency
    - carrier frequencies selected in pseudo-random order

Direct-Sequence Spread Spectrum I

- Repeat message bit $b \in \{-1,+1\}$ N times
  - "chip rate" = N
  - rectangular window $r[n]$

$$r[n] = \begin{cases} 1, & 0 \le n \le N-1; \\ 0, & \text{otherwise.} \end{cases}$$

- Spreading sequence $c[n] \in \{-1,+1\}$
  - noise-like statistical properties

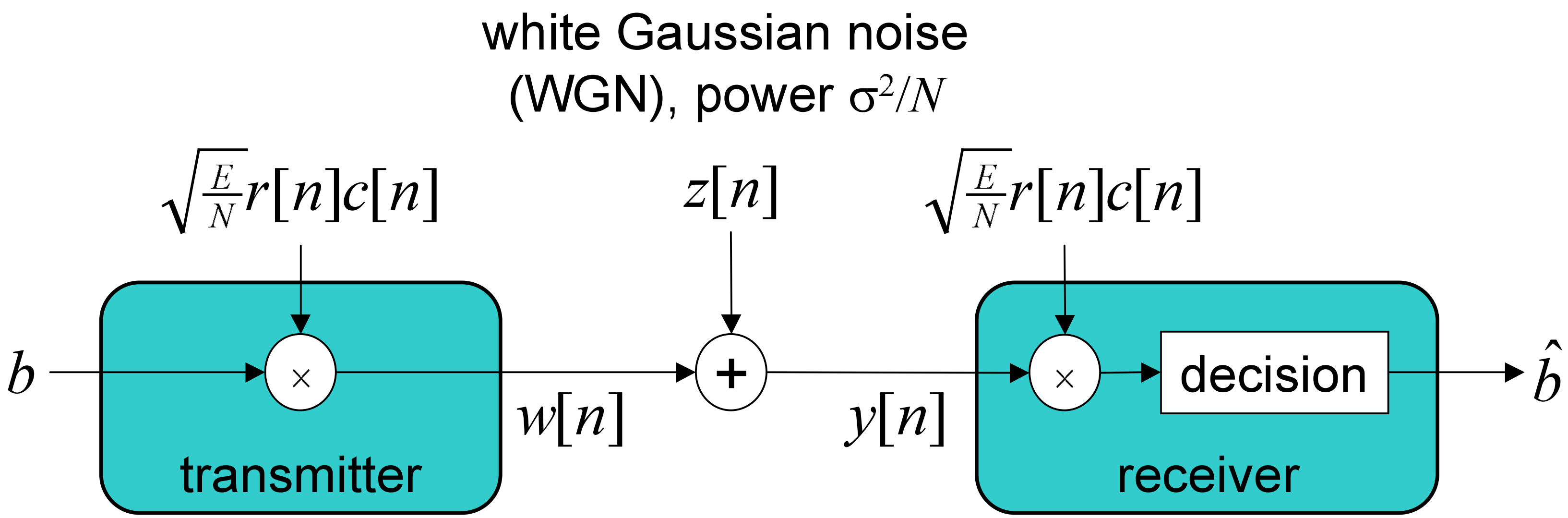
$$\frac{1}{N}\sum_{n=0}^{N-1} c[n] \approx 0$$ zero mean

$$\frac{1}{N}\sum_{n=0}^{N-1} c[n]c[n+k] \approx \delta[k]$$ autocorrelation

  - Gaussian, uniform, other sequences possible
  - generated by a secret key (seed) $\Rightarrow$ SECURITY

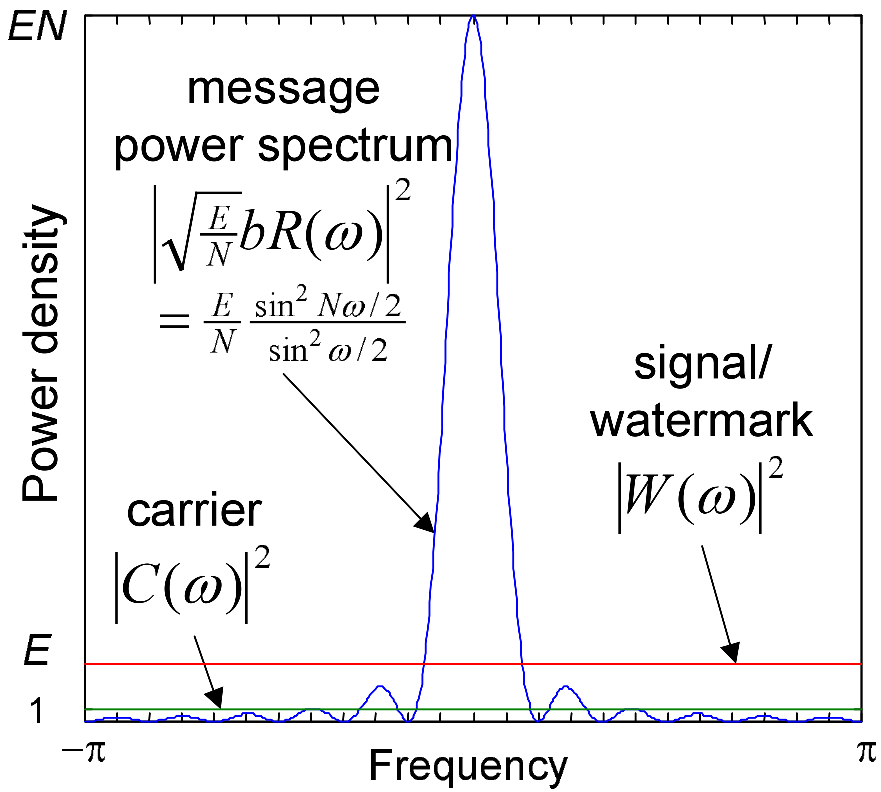
Direct-Sequence Spread Spectrum II

- Standard additive white Gaussian noise (AWGN) channel model



Spreading the Spectrum

- Modulate repeated message bit $br[n]$ with noise-like carrier $c[n]$
- Convolve their spectra
- Result: "spread" the message spectrum over (much) wider bandwidth
- Signal acts like noise and is conveyed via many small modifications $\Rightarrow$ IMPERCEPTIBILITY



Processing Gain

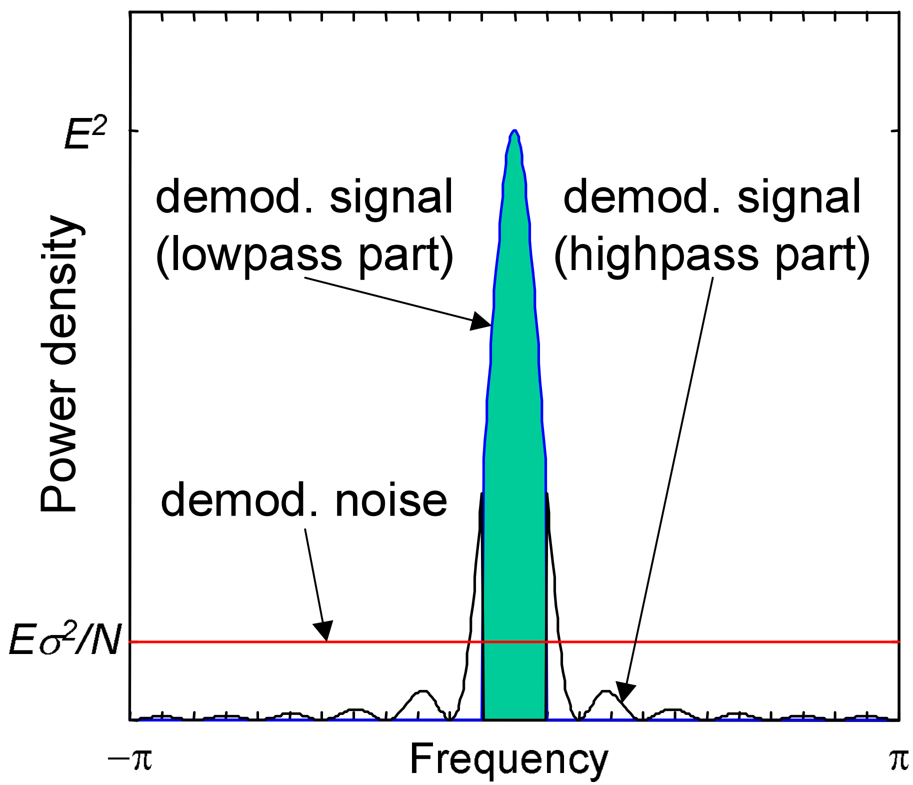
- After demodulation,

$$\text{SNR} = E / \sigma^2$$

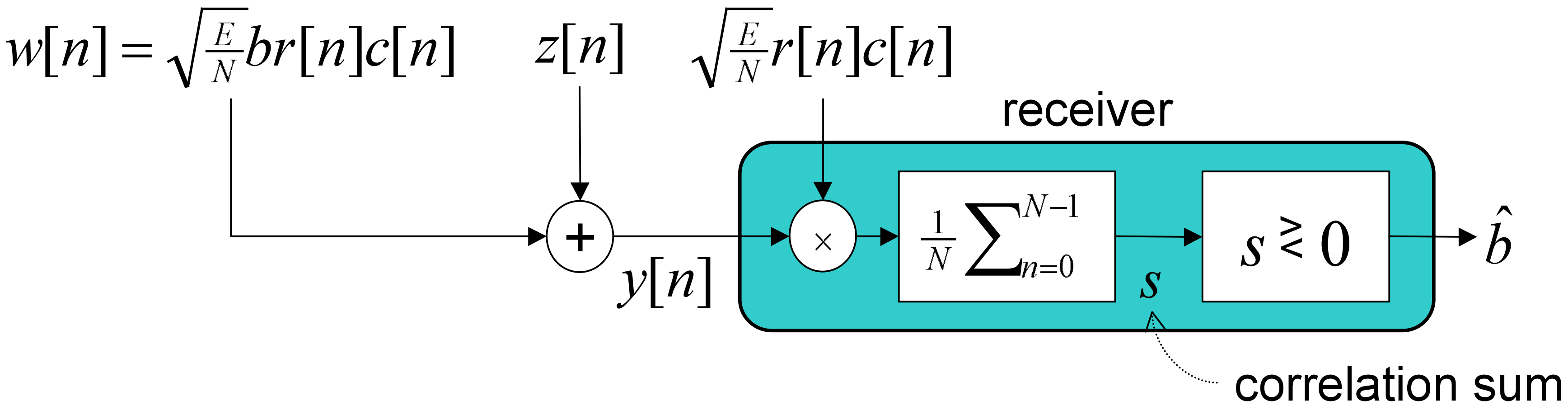
  - message signal is lowpass
  - noise remains white
- Ideal lowpass filtering
  - most of signal passes
  - $1/N$-th of noise passes

$$\text{SNR}_{\text{proc}} \approx N \times \text{SNR}$$

- "Processing gain" $= N$
  - increase SNR by factor of N
- Anti-jamming property $\Rightarrow$ ROBUSTNESS



Correlation Detection I



- Correlation sum s
- Sample correlation of $y[n]$ and $c[n]$

$$s = \frac{1}{N}\sum_{n=0}^{N-1}\left(\frac{E}{N} b r[n] c^2[n] + \sqrt{\frac{E}{N}} r[n] c[n] z[n]\right)$$

$$= \underbrace{\frac{E}{N} b}_{\text{signal}} + \underbrace{\sqrt{\frac{E}{N}} \cdot \frac{1}{N}\sum_{n=0}^{N-1} c[n] z[n]}_{\text{noise}}$$

Correlation Detection II

$$s = \underbrace{\tfrac{E}{N} b}_{\text{signal}} \quad + \quad \underbrace{\sqrt{\tfrac{E}{N}} \cdot \tfrac{1}{N} \sum_{n=0}^{N-1} c[n] z[n]}_{\text{noise}}$$

- AWGN or Central Limit Theorem: s is Gaussian
- Conditional mean and variance of s

$$\mathrm{E}\left[s \middle| b = b_0\right] = \frac{E}{N} b_0 \qquad \rightarrow \text{signal power} = \frac{E^2}{N^2}$$

$$\mathrm{var}\left[s \middle| b = b_0\right] = \frac{E\sigma^2}{N^3} \qquad \rightarrow \text{noise power} = \frac{E\sigma^2}{N^3}$$

  – result: processing gain N

$$\mathrm{SNR}_{\mathrm{proc}} = N \frac{E}{\sigma^2} = N \cdot \mathrm{SNR}$$

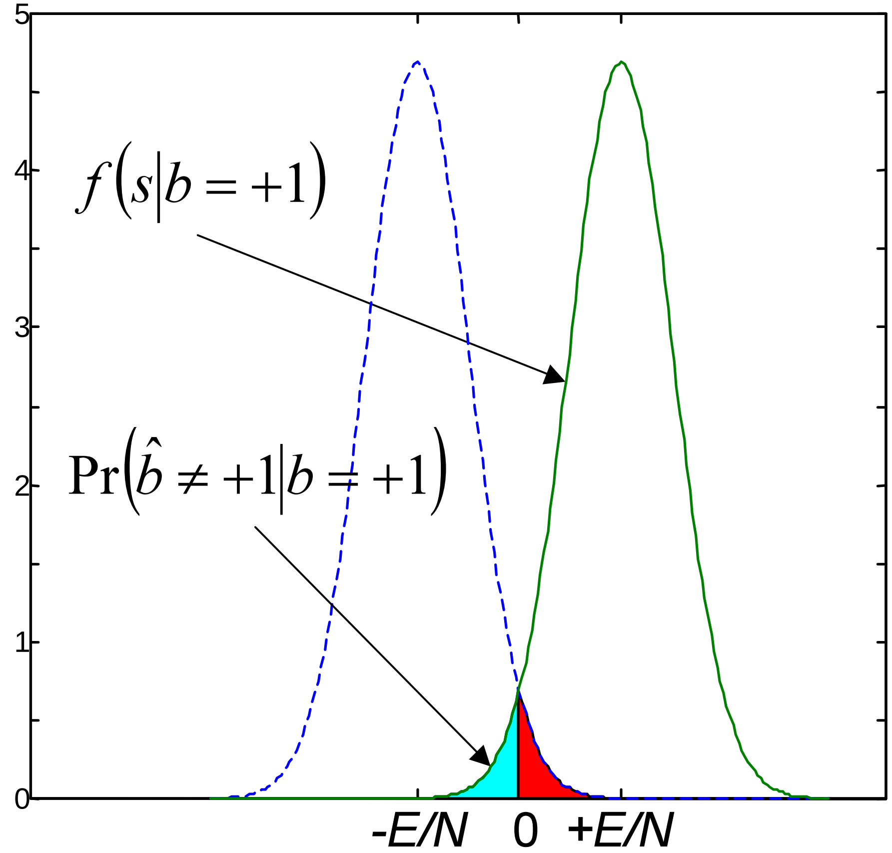
Correlation Detection III

- Correlation sum s
  - assumed Gaussian
  - mean Eb_0/N
  - variance $E\sigma^2/N^3$
- Decision rule becomes

$$\hat{b} = \begin{cases} +1, & \text{if } s > 0; \\ -1 & \text{if } s < 0. \end{cases}$$

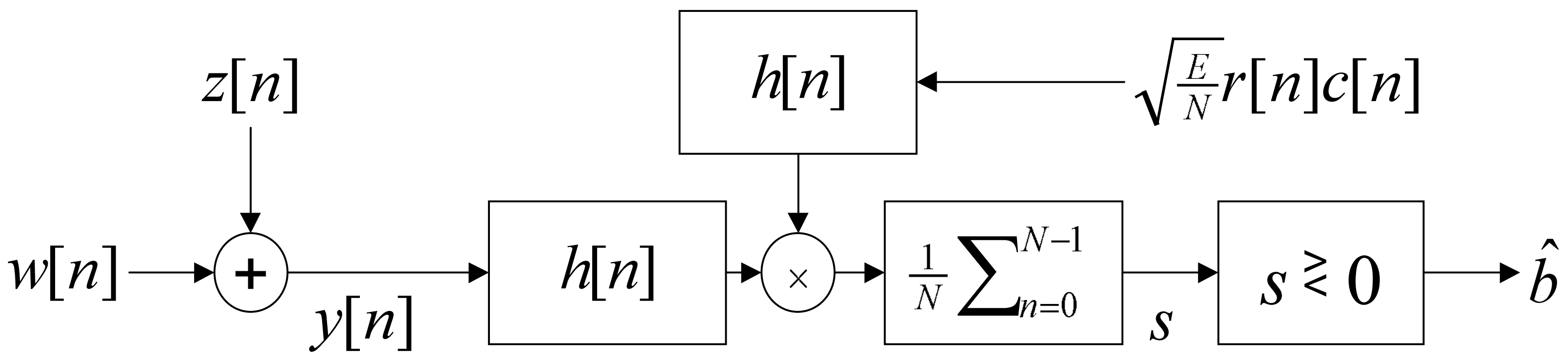
- Probability of error

$$P_E = \Pr\left(\hat{b} \neq b_0 \middle| b = b_0\right) \approx Q\left(N\sqrt{\frac{E}{\sigma^2}}\right)$$

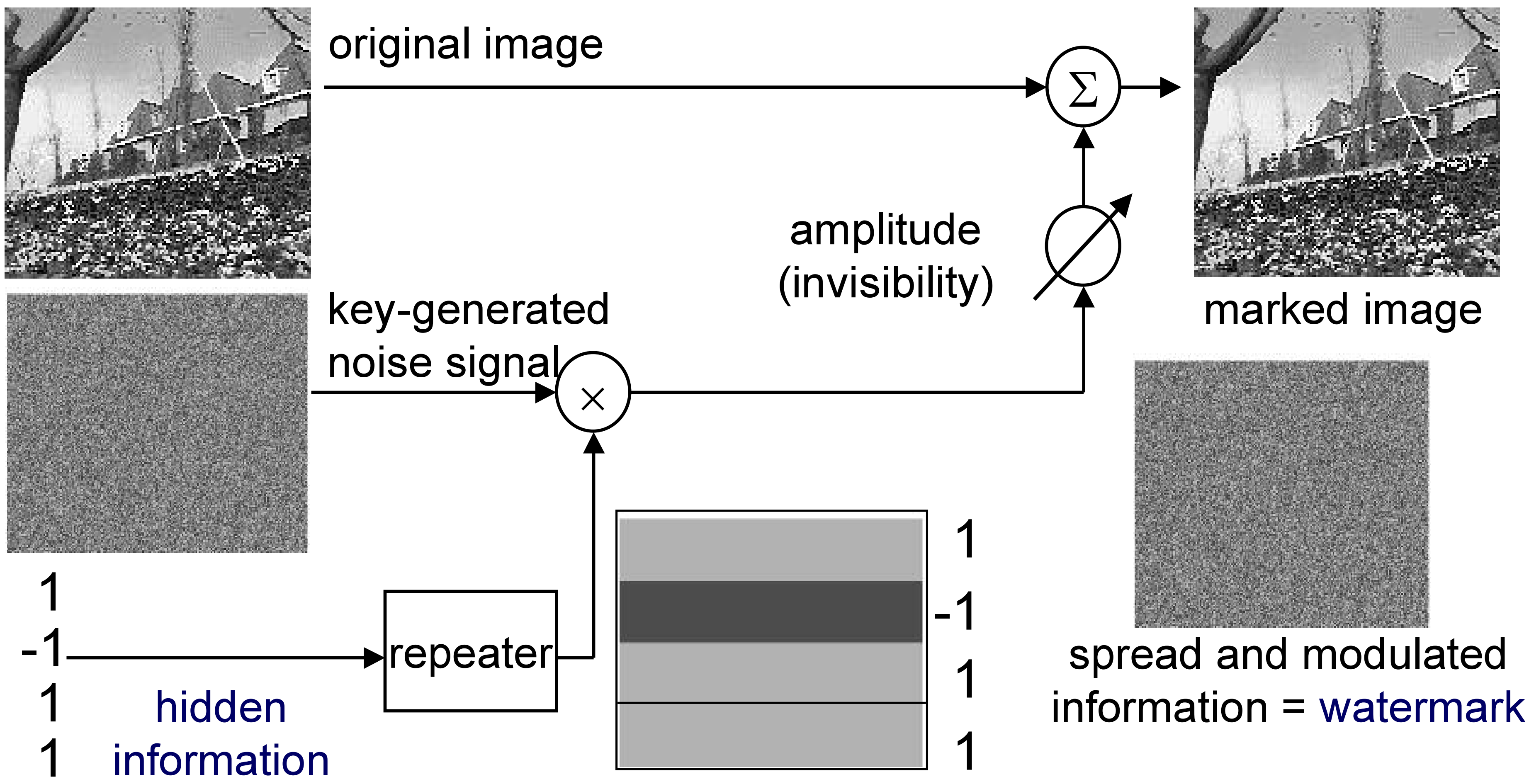


Colored Noise

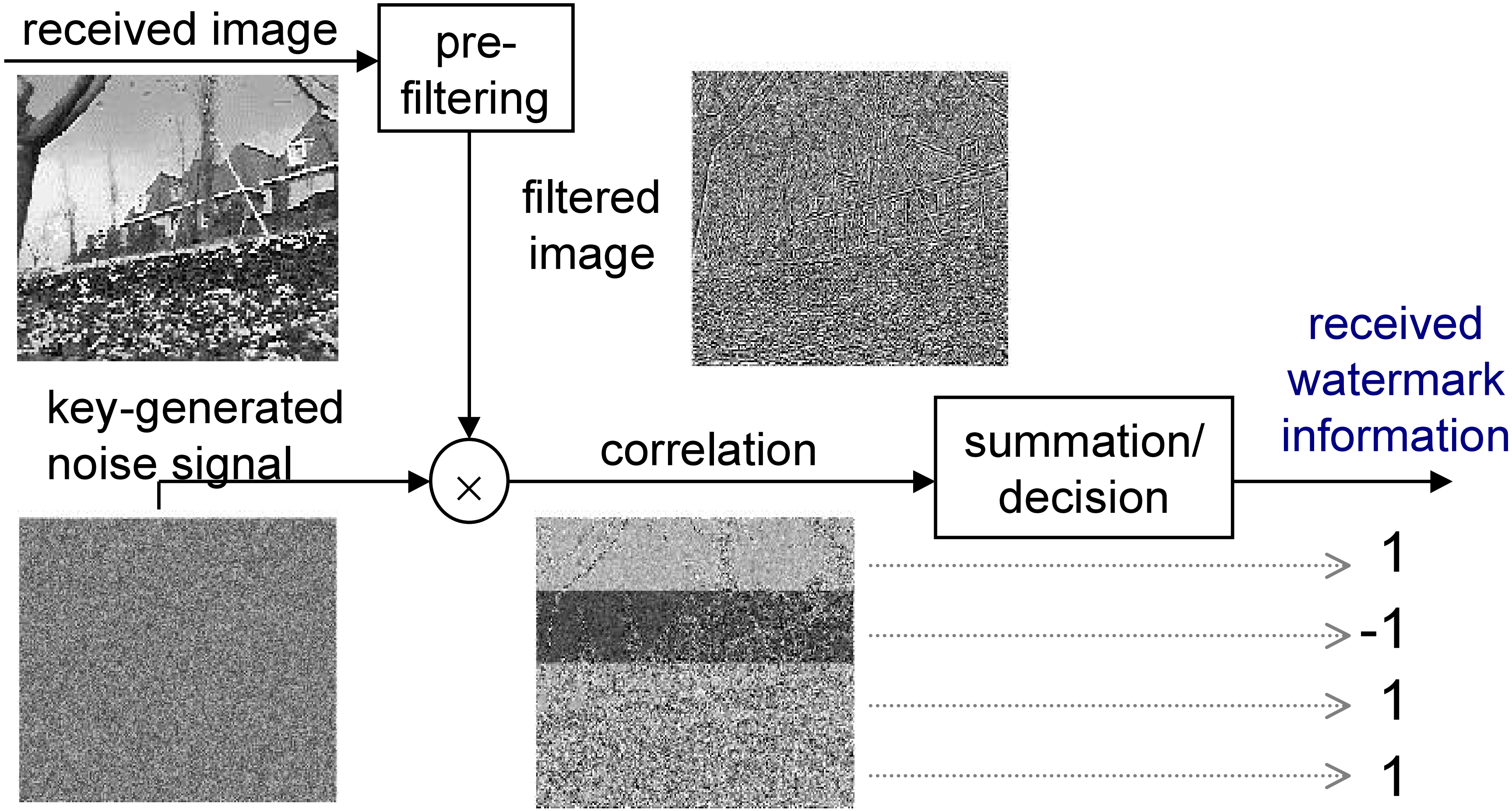
- Correlation detection is optimal for <u>white</u> noise
- For colored noise, use pre-whitening filter $h[n]$
  [Hancock, Wintz 1966], [Depovere *et al*. 1998], [Kalker, Janssen 1999]



Example: Watermark Embedding



Example: Watermark Retrieval



[Bender, Gruhl, Morimoto 1994]

Early Example: "Patchwork" Algorithm

- 2 disjoint sets, A and B, of n pixels each
  - pixels in each set ("patch") chosen randomly
  - assumption:
    $$S = \sum_i \left(A_i - B_i\right) \approx 0$$
  - embedding: $A'_i \leftarrow A_i + 1$, $B'_i \leftarrow B_i - 1$
    $$S' = \sum_i \left(A'_i - B'_i\right) \approx 2n$$
  - detection: if $S' \approx 2n$, watermark present
- Like spread-spectrum watermarking
  - communicate information via many small changes that are randomly selected

[Cox, Kilian, Leighton, Shamoon 1995-1997]

Popular Example: NEC Scheme

- Heuristic claim
  - watermark should be embedded in the "perceptually significant frequency components" for best robustness
- Embedding
  - N watermark samples w_i ~N$(0,1)$; e.g., $N = 1000$
  - embed in the N largest-amplitude DCT coefficients (except DC coefficient) x_i

$$y_i = x_i(1 + \alpha w_i)$$

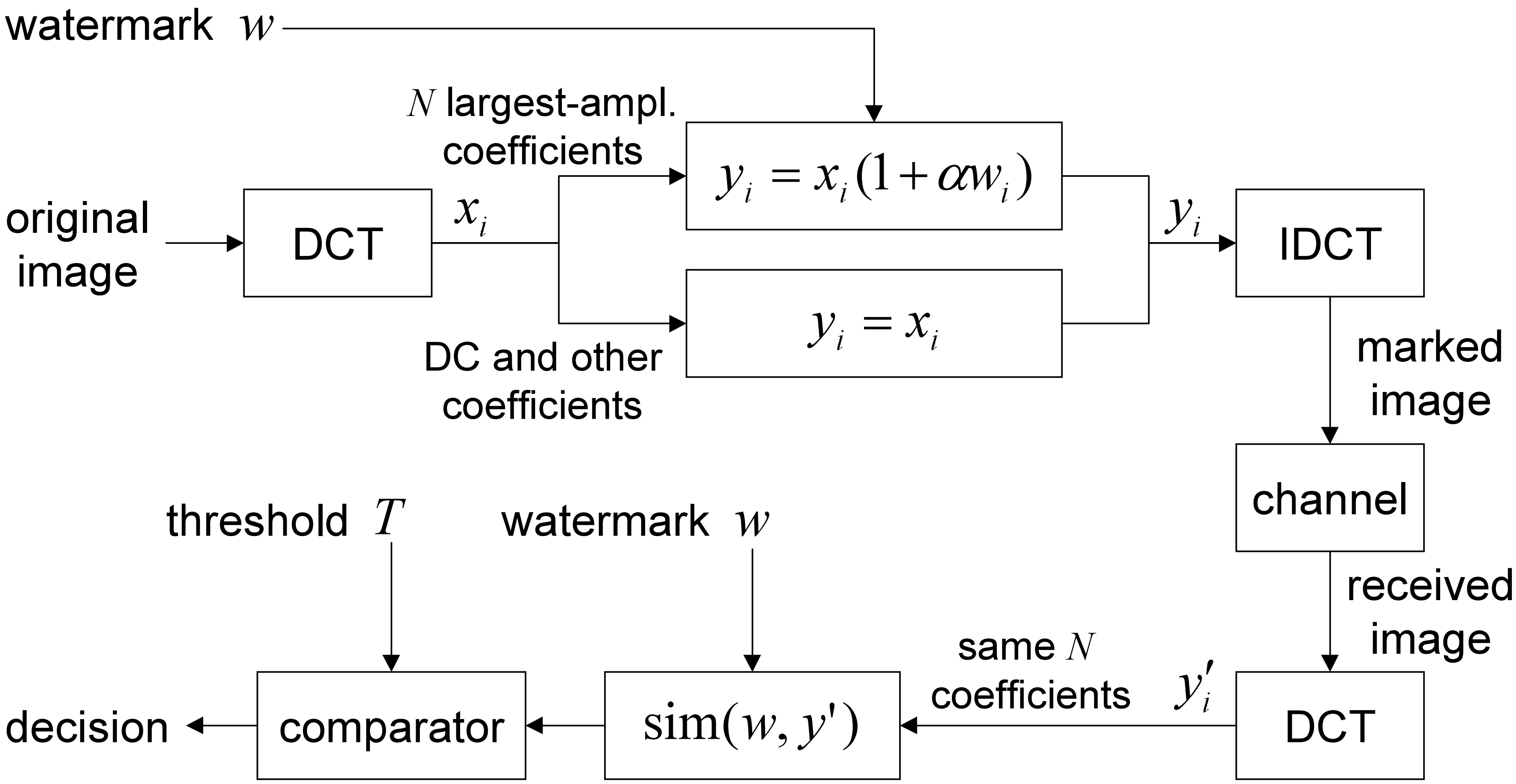
- Detection
  - extract the same N DCT coefficients y_i'
  - compute the similarity (normalized correlation) between y_i' and w_i

$$\text{sim}(w, y') = \frac{\langle w, y' \rangle}{\sqrt{\langle y', y' \rangle}}$$

  - watermark w is present if $\text{sim}(y', w) > T$

[Cox, Kilian, Leighton, Shamoon 1995-1997]

Block Diagram of NEC Scheme



Possible Drawbacks of Spread Spectrum

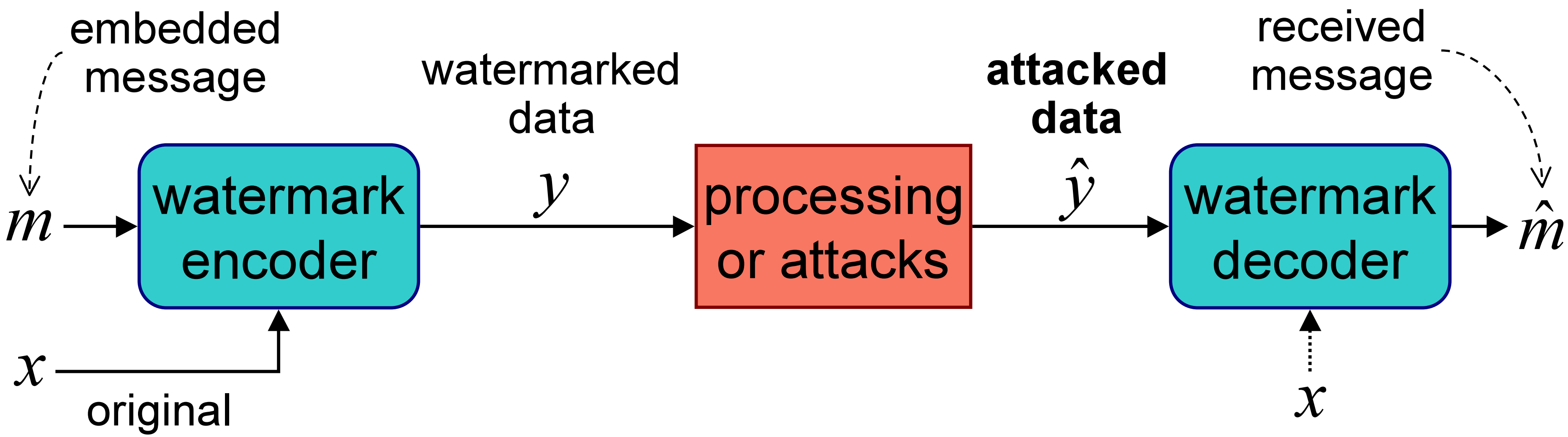
- Fails if synchronization is lost
  - autocorrelation property of spreading sequence
  - re-synchronization can be computationally expensive
- Watermark can be removed
  - knowledge of spreading sequence enables one to compute watermark signal and subtract it from the watermarked data
- Blind watermarking
  - imperceptibility means original data behaves like a powerful interferer
  - low communication rates

ATTACKS AND ROBUSTNESS

- Examples and classes of attacks
- Notion of robustness
- Kerckhoff’s principle

Definition of Attack

- Watermarked data will likely be processed
- Attack - any processing that may coincidentally or intentionally impair communication of the embedded information
- Treat attacks like a communications channel



Examples of Attacks

- format conversion
  - 4:3→16:9, frame rate
- lossy compression
  - JPEG, MPEG-2, MP3
- filtering, additive noise
- D/A+A/D
  - printing & scanning
  - CD→tape→CD
- geometric transformation
  - rotation, scaling, translation
  - cropping, composition
  - zoom, aspect ratio
- jitter
  - interchange of samples
  - line/frame holding/dropping
- histogram equalization
- time/space scaling
- collusion (multiple copies)
  - use several differently marked documents
- deadlock (protocol)
  - generate fake signals (watermark, original) that cannot be distinguished from true signals

[Hartung, Su, Girod 1999]

Classes of Attacks

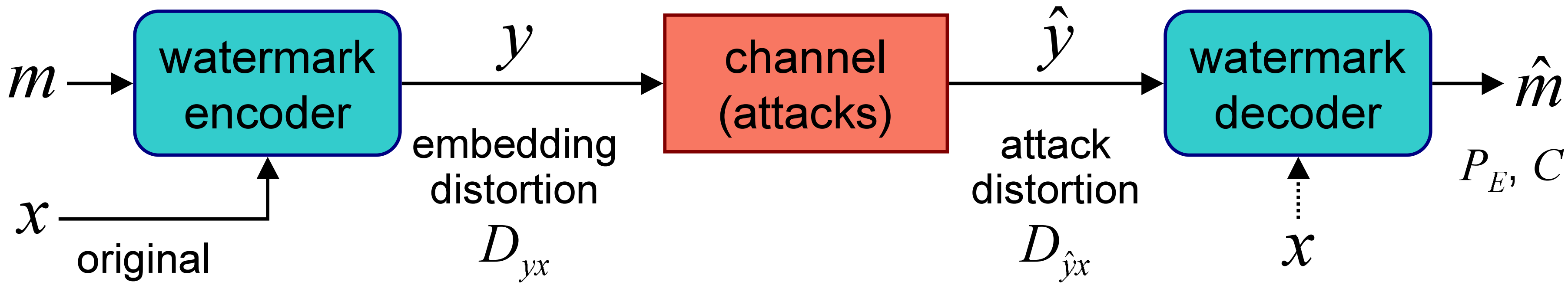
- Simple waveform processing - “brute-force”
  - impairs watermark and perhaps original data, too
  - linear filtering, additive compression, noise, quantization
- Detection-disabling - disrupt synchronization
  - geometric transformations (RST), cropping, shear, re-sampling, shuffling
  - watermark harder to locate
- Advanced jamming or removal - intentional attempt to impair/defeat watermark
  - watermark estimation
  - “optimum” attacks
  - collusion (multiple copies)
- Ambiguity/deadlock - exploit flaws in protocol
  - fake watermark or original
  - copy watermark signal

Notion of Robustness

- How well does a watermark resist an attack?
- Easy to define robustness
  - *"A watermark is robust if communication cannot be impaired without rendering the attacked data useless."*
- Hard to evaluate it
  - "When is communication impaired?"
  - "When is the attacked data useless?"

Evaluating Robustness

- "When is communication impaired?"
  - watermark-as-signal: no longer reliably detectable
  - watermark-as-information: no longer reliably decodable
  - measure P_E, C, etc.
- "When is the attacked data useless?"
  - multimedia: quantify "usefulness" by measuring distortion
  - also measure distortion after embedding



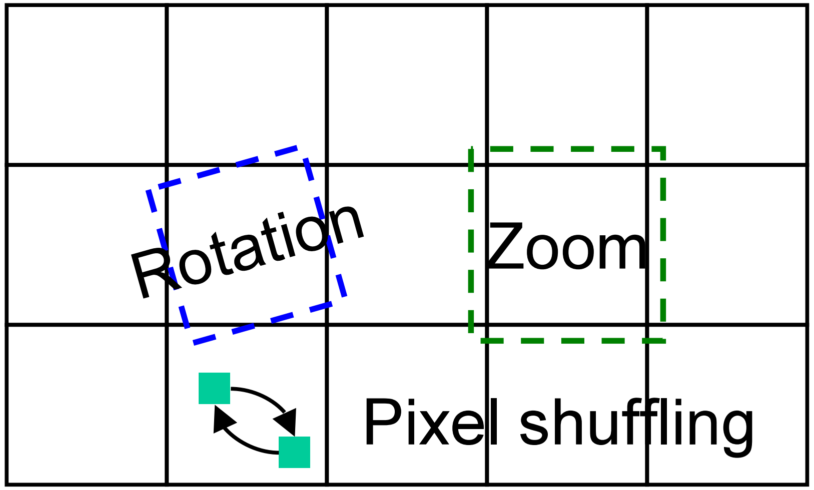
Attacks to be Discussed

- De-synchronization and re-synchronization
- Quantization and compression
- Watermark estimation
- Theoretically optimum attacks and defenses
- Collusion (multiple copies)
- Ambiguity & deadlock

DE- AND RE-SYNCHRONIZATION

Synchronization

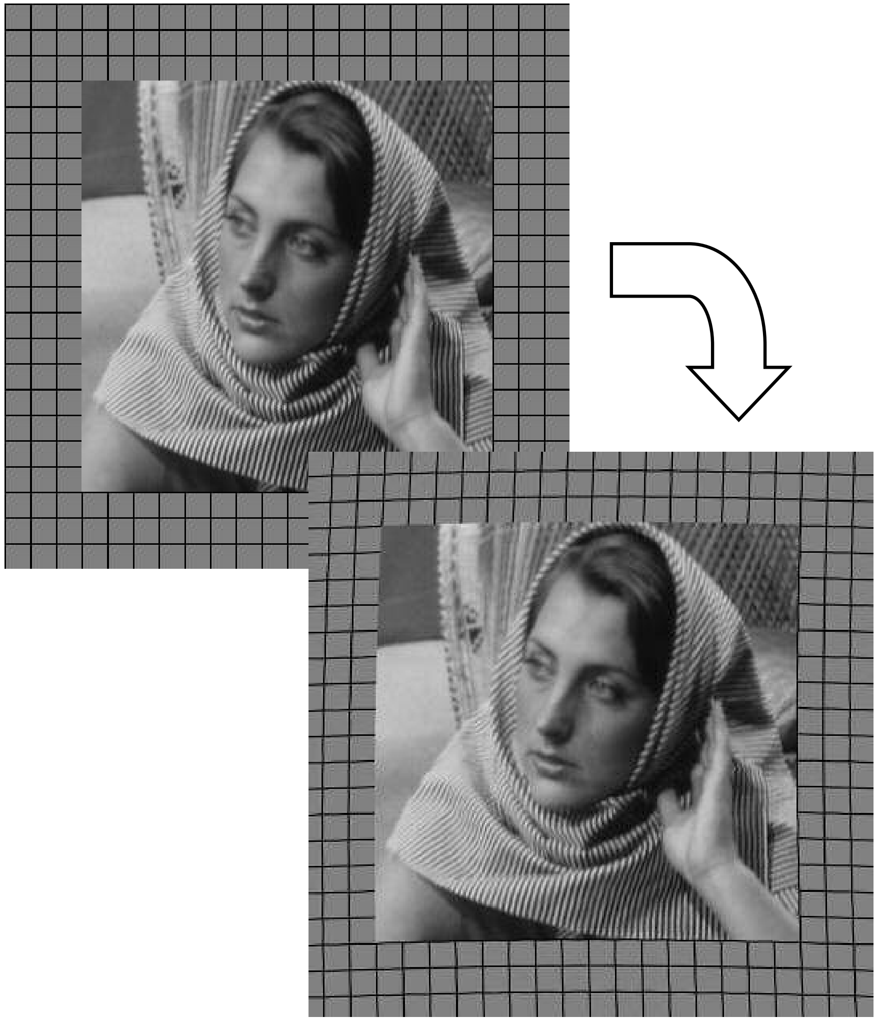
- Loss of synchronization
  - spread spectrum fails
  - defeats simple receivers
  - does not remove watermark signal, but…
  - makes watermark signal more difficult to locate



- Better receiver should be able to re-synchronize
- Open question: How to measure distortion?

[Petitcolas, Kuhn 1997-2000]

Example: StirMark



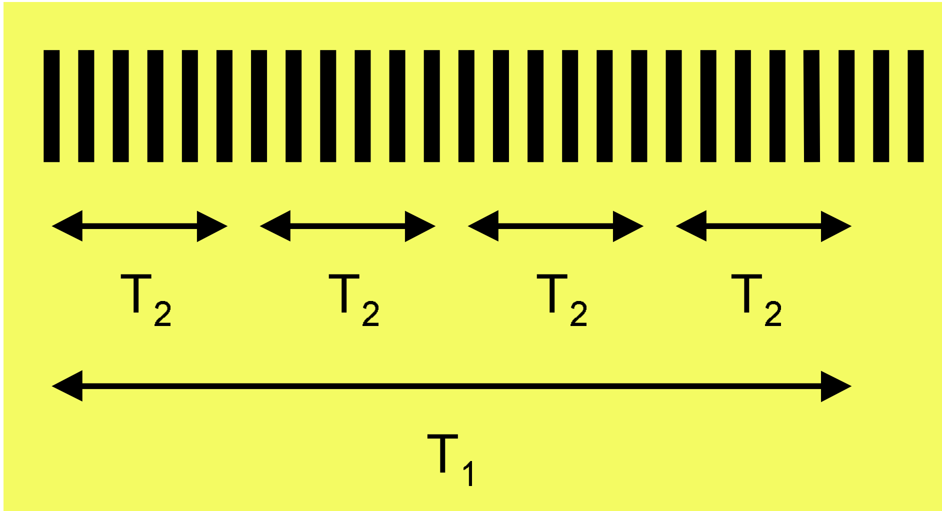
- Popular, free software
  - simulates printing & scanning
  - geometric distortion & JPEG (de-synchronization & compression)
  - easy to use and test
  - most features available elsewhere
- Does not use Kerckhoff's principle
  - does not target specific system weaknesses
  - suboptimal attack
  - false sense of security?

JAWS & Millennium

- Philips Video Watermarking for DVD-Video copy protection

Overview JAWS

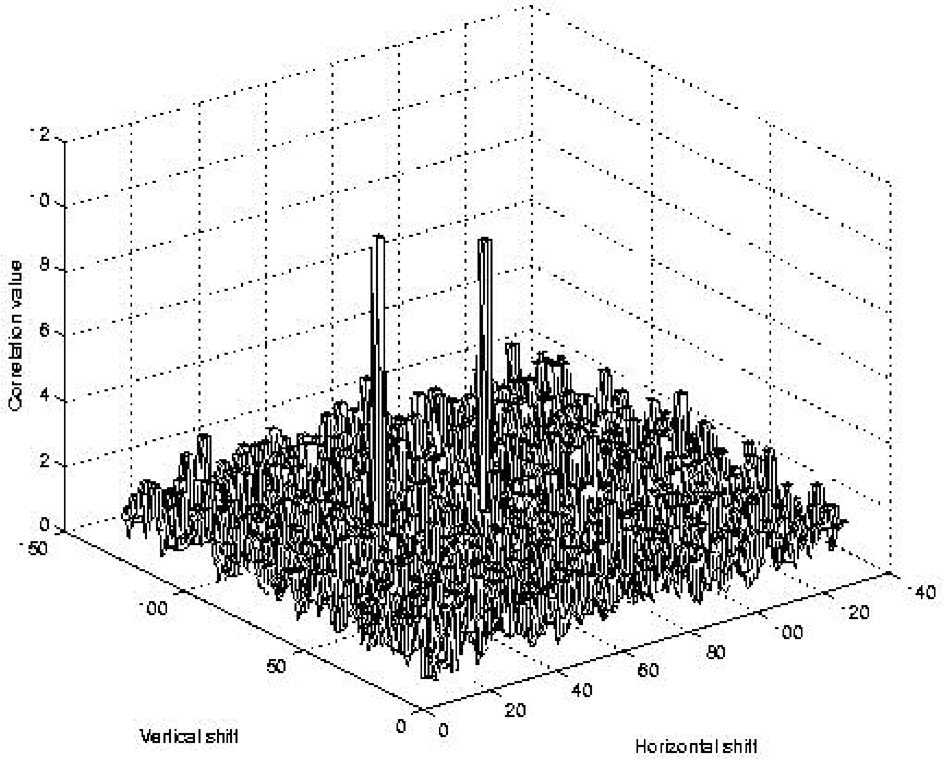
- JAWS = Just Another Watermarking System
- JAWS is a video watermarking system
- JAWS considers video as a sequence of still images
- JAWS marks chunks of T_1 consecutive frames with the same mark.



- JAWS detects on chunks of T_2 consecutive frames, $0<T_2<T_1$

Overview JAWS

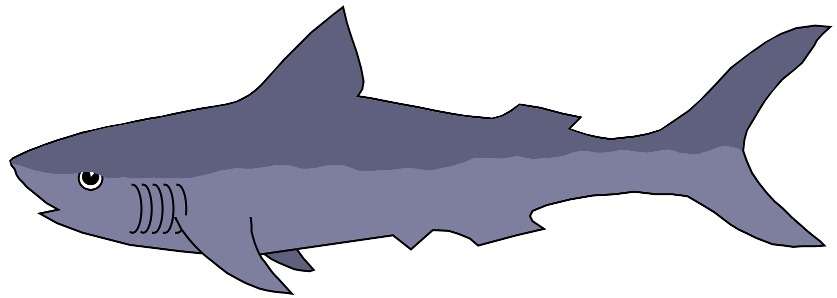
- JAWS embeds marks in the spatial domain
- JAWS uses pseudo-random noise sequences with translational symmetry (i.e. is a repetition of smaller tiles)
- JAWS embeds information (payload) in the *relative* position of embedded marks (not in presence/absence).



- JAWS is shift and cropping invariant

Overview JAWS

- Every JAWS detection yields
  - watermark present or not;
  - if present, payload is retrieved,
  - with an indicator of the reliability of detection and payload
- JAWS has successfully been tested in the DHSG of the CPTWG and the VIVA consortium

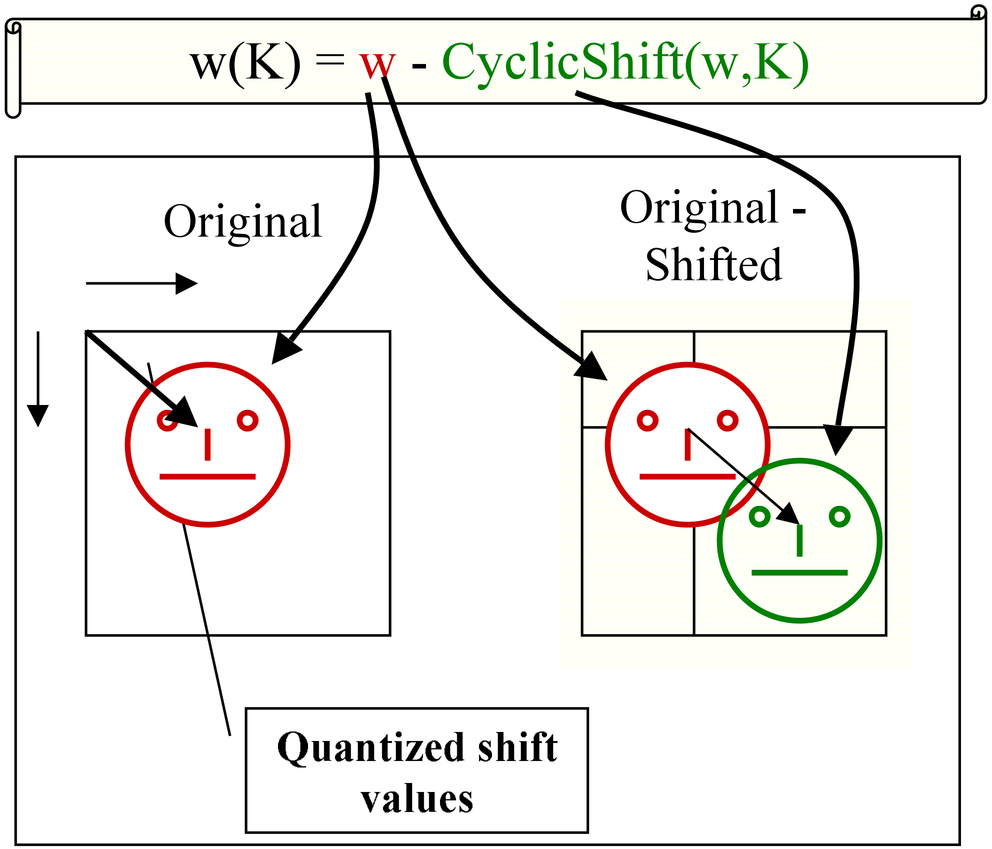


•JAWS is a registered trademark

•Philips is not allowed to use a shark symbol in connection with JAWS watermarking

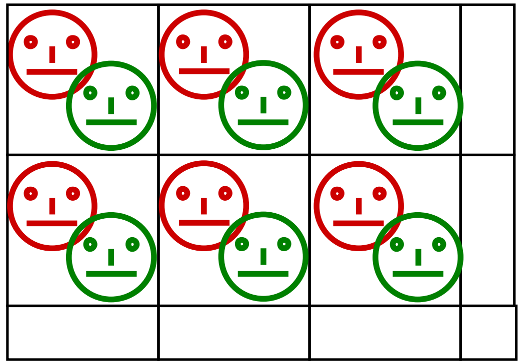
Ingredients

- Random matrix w
  - universal secret
  - size 128 x 128
  - i.i.d. from N(0,1)
- Payload K
  - 4 + 4 = 8 bits
- Payload secret w(K)
  - size 128 x 128
  - i.i.d. from N(0,1)



Embedding

- Video is seen as sequence of stills
  - every frame watermarked in identical manner
- w(K) is repeated to size of video frame
  - truncation if necessary
  - tiling
  - W(K)



Covering of a video frame by basic blocks

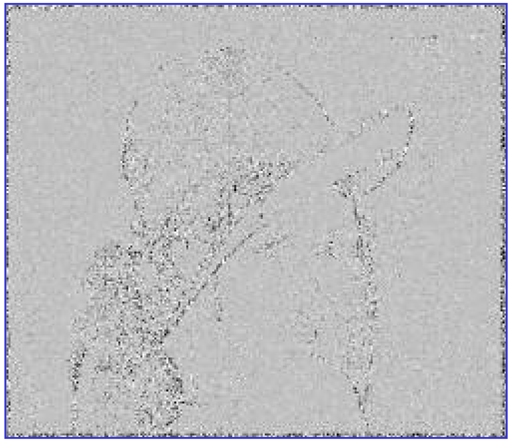
Local Depth

- Embedding rule

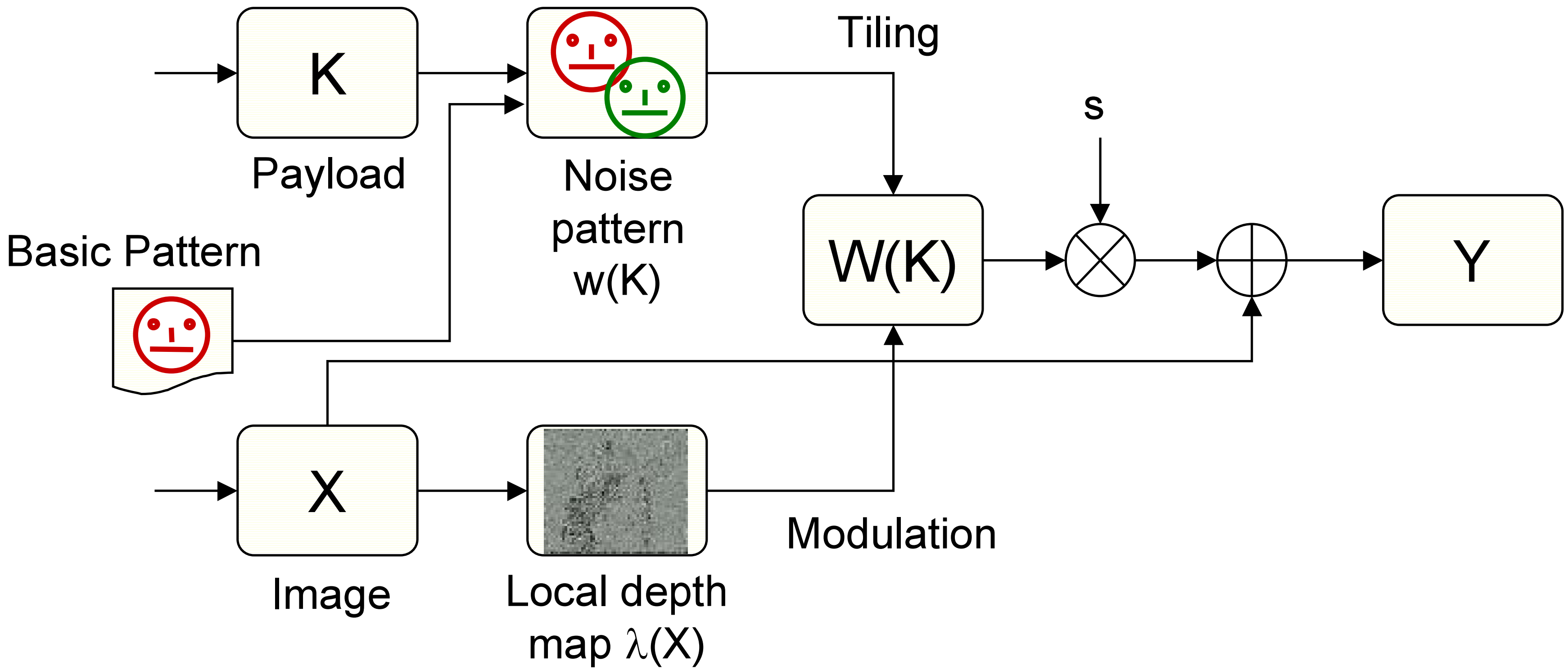
  $Y = X + s\,\lambda(X)\,W(K)$

- Embedding depth s
  - controls the reliability of detection
  - frame dependent
  - computable from required reliability and visibility

- Local embedding depth $\lambda(X)$
  - spatial masking
  - $\text{mean}(\lambda(X)) = 1$
  - small in non-textured and low luminance areas
  - large in textured and high



Embedding Overview



Detection

- Detection is correlation

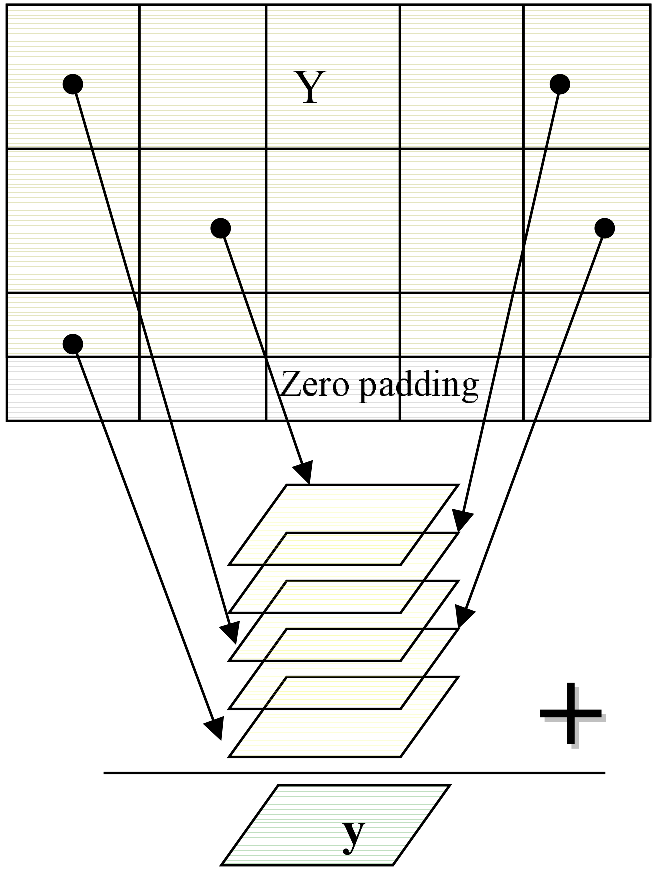
  $d = \langle Y,W \rangle = \langle X,W \rangle + s \langle W,W \rangle$

- Robustness increased by
  - accumulation in time (T_2)
  - *matched filtering*
- Synchronization a priori not known
  - search over 128 x 128 possibilities
  - efficient implementation through
    - folding
    - FFT

Folding

- Efficient implementation of correlation by folding (exploitation of structure of W)

  – d = <y,w>
  – y = fold(Y)
  – y of size 128 x 128



Synchronization

- Detection when synchronization is unknown
  - A priori exhaustive search is needed

$$d_k = <\mathrm{CyclicShift}(w,k),y>$$

  - k ranges over [128 x 128]
  - computationally infeasible
  - Efficient computation of d_k with Fast Fourier Transform

$$[d_k] = \mathrm{IFFT}(\mathrm{FFT}(y) * \mathrm{conj}(\mathrm{FFT}(w)))$$

SPOMF

SPOMF

- Matched filtering in the Fourier domain
  - Matched filtering can be done in the Fourier domain
    - no costly spatial filtering
  - Matched filtering can be taken to the extreme
    - “super whitening”
    - Discard magnitude information from FFT(y)

Phase(FFT(y)) = FFT(y) / Abs(FFT(y))

  - Extra detection boost by “whitening” FFT(w)

SPOMF (cont.)

- SPOMF = *Symmetrical Phase-Only Matched Filtering*

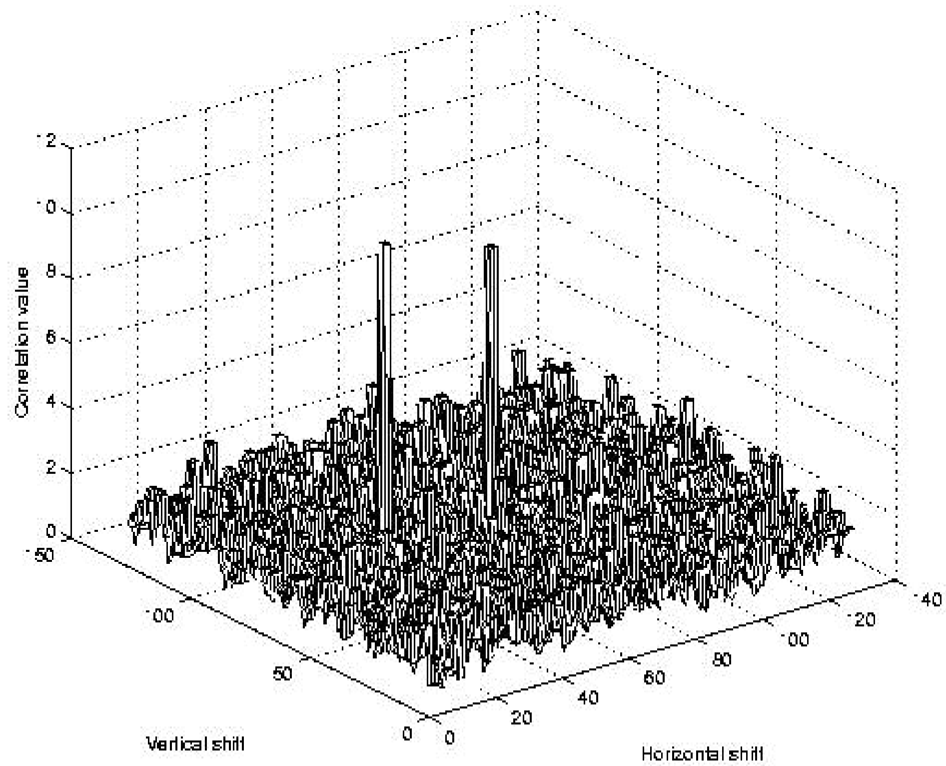
$$[d_k] = \mathrm{IFFT}(\mathrm{Phase}(\mathrm{FFT}(y)) * \mathrm{conj}(\mathrm{Phase}(\mathrm{FFT}(w))))$$

Stored in ROM
or better,
computed on the fly.

Payload

- Only one SPOMF operation needed
  - SPOMF(y,w) yields two peaks
    - One positive peak at position p
    - One negative peak at position q
    - Payload K retrieved by subtraction

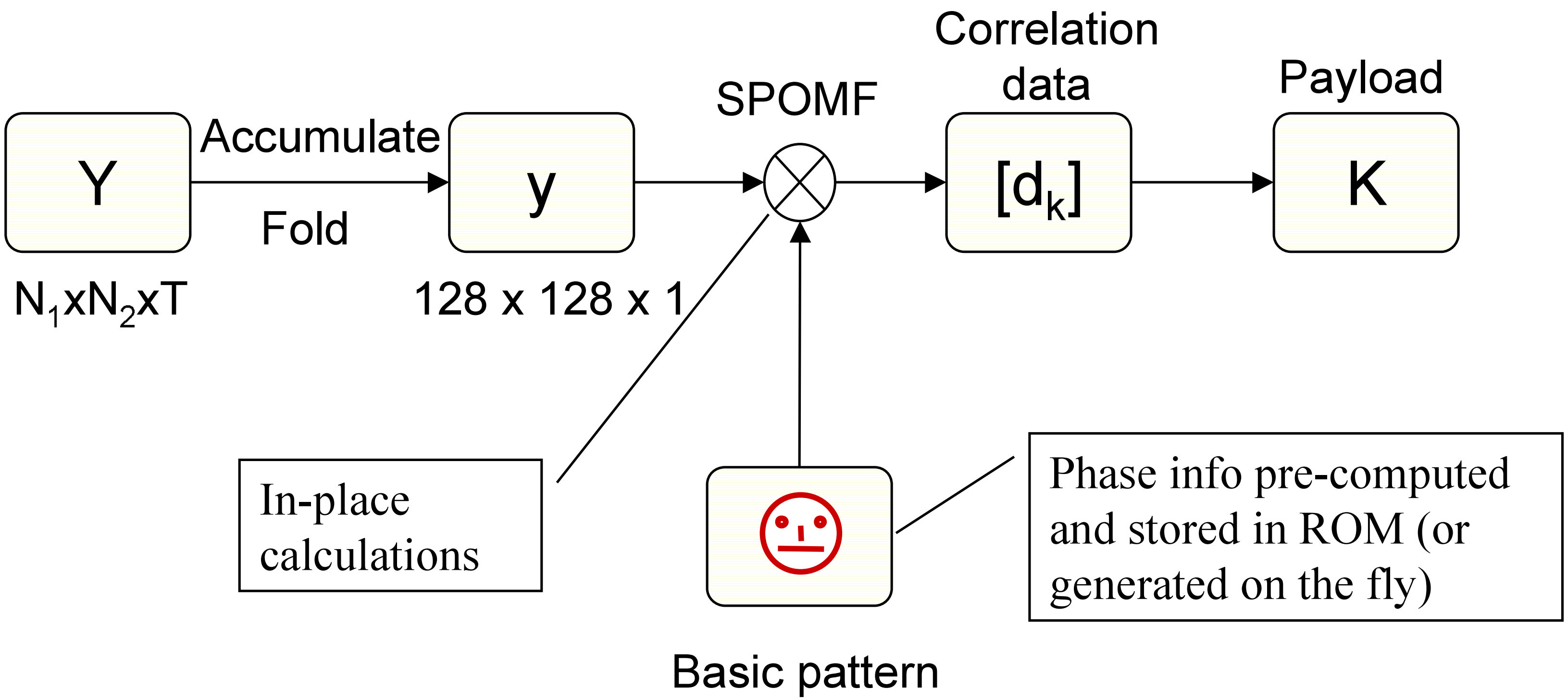
      K = q - p
  - Invariance for translations



False Positive Rate

- False positive rate with SPOMF
  - The matrix $[d_k]$ can be seen as a set of correlations of the watermark w with a large number of images.
  - The standard deviation $Std(d)$ can be estimated from this matrix.
  - If Y is watermarked, $[d_k]$ will contain 2 large values D_i.
  - The reliability of these peaks an directly be calculated from the quotient D_i / $Std(d)$.
  - For reliable detection, this ratio needs to at least 5.

Detection Overview



Millennium System Aspects

- Location of the watermark detector
- Copy Generation Control

Goal

- Goal: a copy protection system for DVD video
  - enforcing the mantra
    "keep honest people honest"
  - based upon digital watermarks,
  - robust to common processing
    - MPEG encoding, letter-boxing, ...
  - implementing 4 copy protection states,
  - not affecting the content quality,
  - allowing an efficient implementation,

Basic Philosophy

- Watermarking is only a part (though an essential one) of the DVD copy protection system.
- Watermark embedding is a delicate issue. Watermark embedding should only be done in a professional environment as not to compromise the quality of the content.
- Watermark detection should be possible in all video formats. Base-band detection, being the common denominator of all video formats, is therefore a required feature.
- Watermark detection preferably only occurs where base-band video is available.
- Watermark detection does not significantly increase the complexity of the hardware/software module in which it is embedded.
- The copy protection system should be scalable and extendible.
- The total copy protection system needs to implement copy generation control.

System Issues

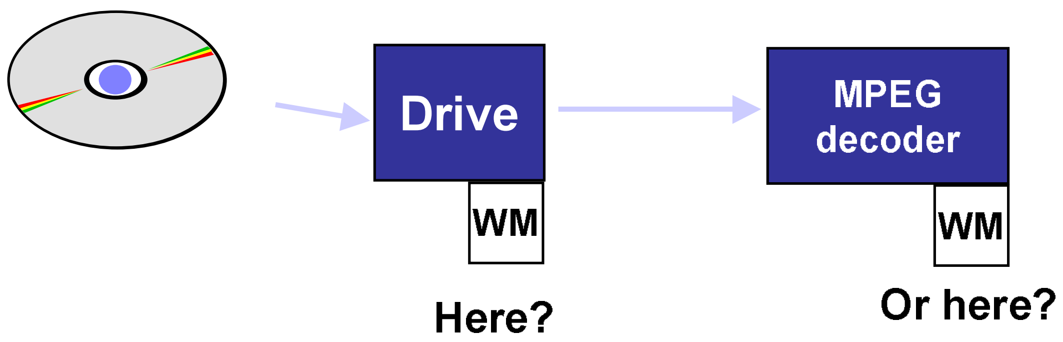
- System parts
  - JAWS watermarking
  - .....
  - .....
- Issues
  - location of the watermark detector
    - Watermark Detector at the application
  - copy generation control through remarking or not
    - Copy Generation Control through tickets

Location of watermark detector

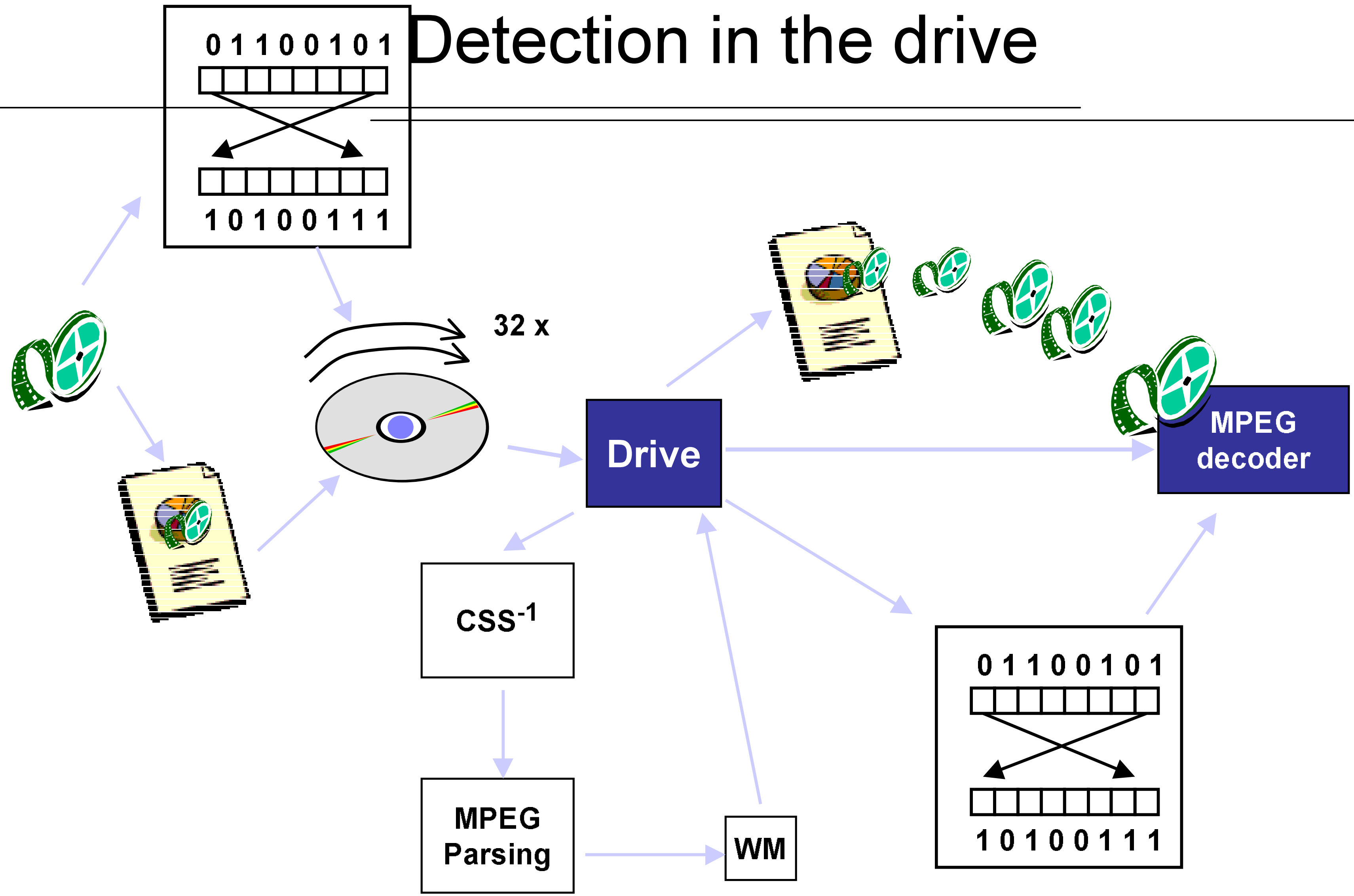
- Two options
  - detector in drive
  - detector near application (MPEG decoder)

Location of the watermark detector

- Two options
  - detector in drive
  - detector near application (MPEG decoder)



Detection in the drive



Detection in the drive

- Advantages
  - copy protected data will leave drive only if allowed
  - works with non-compliant decoders and STBs
  - drive-to-drive copying included
- Disadvantages
  - no opportunity to share resources
    - <CSS descrambling>, MPEG parsing
  - Detection in the drive has to handle all read methods
    - non-sequential, 32x speed

Detection in the drive

- Disadvantages (cont.)
  - Detection in the drive is not extendable
    - mJPEG, AVI, Wavelets, QuickTime, (MPEG) Audio, ...
  - Detection in the drive allows simple attacks
    - bit-inversion, wrappers, ...

Detection at the application

- Advantages
  - sharing of resources
  - extendable
  - simple attacks need non-compliant decoders
  - exploitation of crypto infrastructure
- Disadvantages
  - no digital links between drive and non-compliant application allowed
  - disk-to-disk copies need to be mediated

Detection at the application