Tatouage de signal et sécurité

30/11/2005

Stanislas Francfort

Diffusion Libre

D1 - 07/12/05

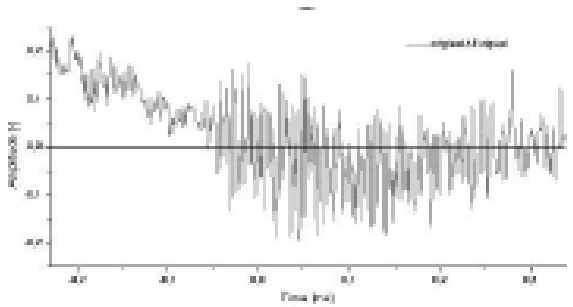
France Télécom
Recherche & Développement

Principes

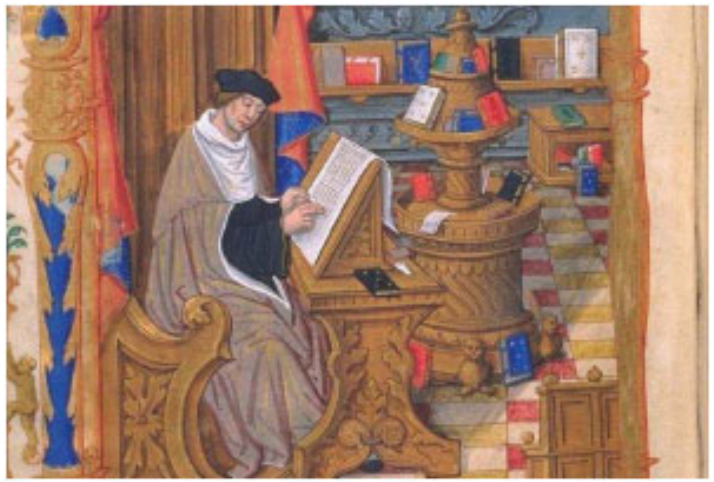
Tatouage - watermark

- **Un tatouage (ou watermak) est une insertion d'information (marque) dans un signal**

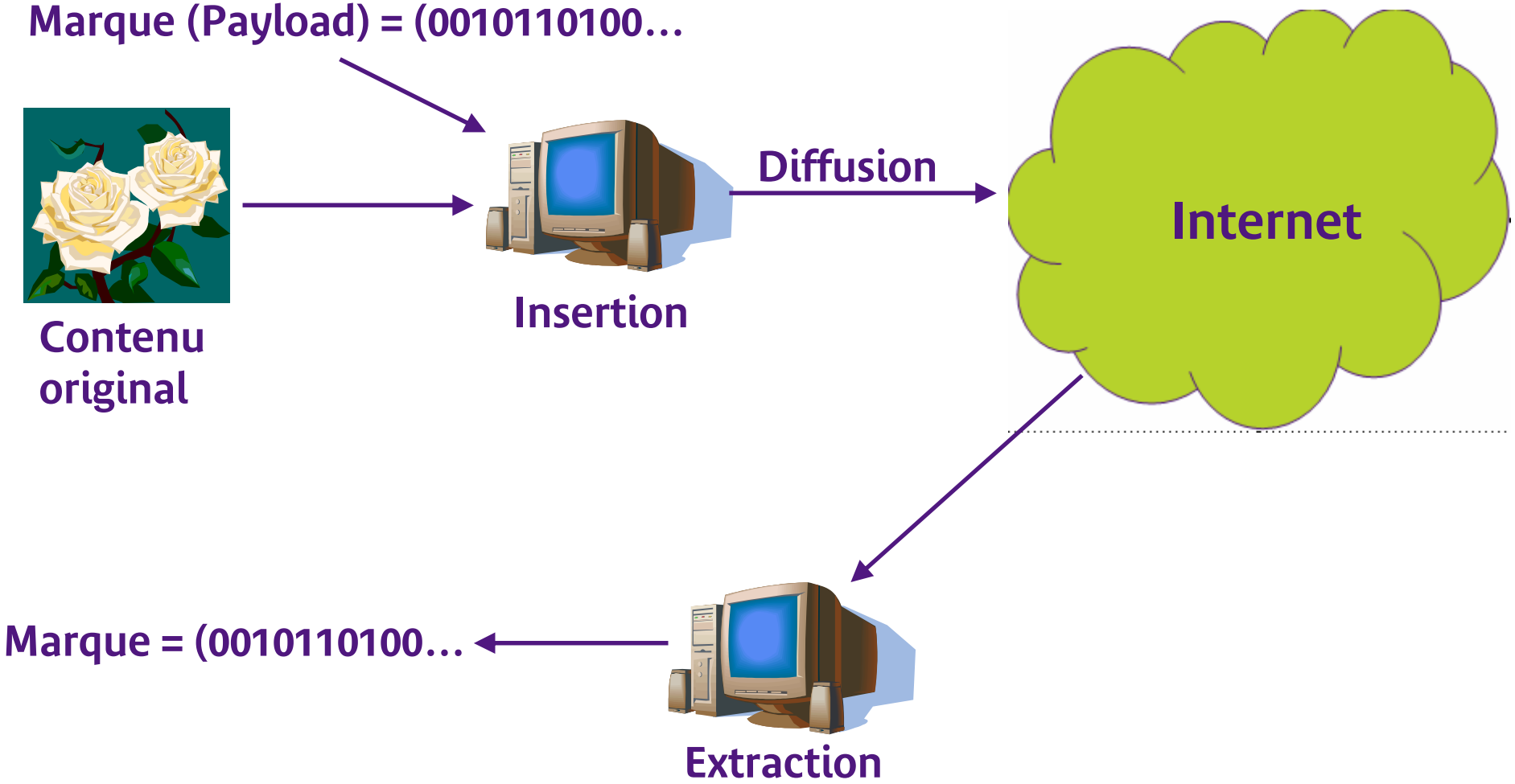
- **La marque devra être insérée directement dans le signal**
  - Elle doit rester présente quelque soit la représentation de ce signal



Une idée ancienne



Principe



Propriétés du watermark

- Rapidité d'insertion
- Rapidité d'extraction
- Taille du payload
- Watermark informé/aveugle
- Payload connu ou inconnu lors de la détection
- Faux positifs/faux négatifs
- Robustesse
- Sécurité

Imperceptibilité

Un tatouage doit être imperceptible

- Ceci n'est pas un tatouage :

La variation locale d'énergie (activité) de la marque devra être plus faible celle du contenu

- 4% du contraste pour une image
- 40dB pour un son

Imperceptibilité

Contenu original

Après masquage

Contenu watermarké

Robustesse

- **Un tatouage doit être difficile à effacer, il doit résister :**

- **au transcodage**
- **aux filtres classiques**
  - Ajout/suppression de bruit
  - Flou
  - Suppression de pixels/trames
- **aux transformations numérique/analogique/numérique**
- **à la compression avec perte (mp3, jpg, DivX…)**
- **Aux transformations géométriques**

Payload vs robustesse

Peu robuste

Robuste

Grande taille de Payload

Petit Payload (1 bit)

Sécurité

- **La sécurité d'un watermark est sa capacité à résister à des attaques hostiles, spécialement conçues pour mettre en échec le schéma de watermark**

- **Types d'attaques**
  - Insertion non autorisée
  - Détection non autorisée
  - Suppression non autorisée

Applications

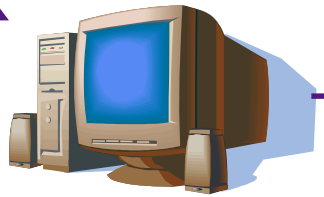
Utilisations

- Contrôle de Broadcast
- Traçage de transaction
- Copie contrôle
- Authentification
- Contrôle de périphérique
- Droit d'auteur

Contrôle de Broadcast

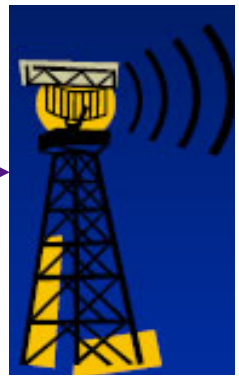
Marque (Payload) = (0010110100…

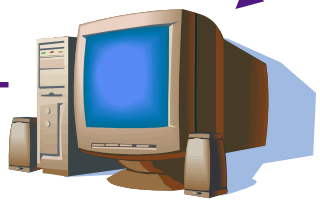
Contenu original



Insertion

Diffusion





Extraction

Le contenu a été Broadcasté

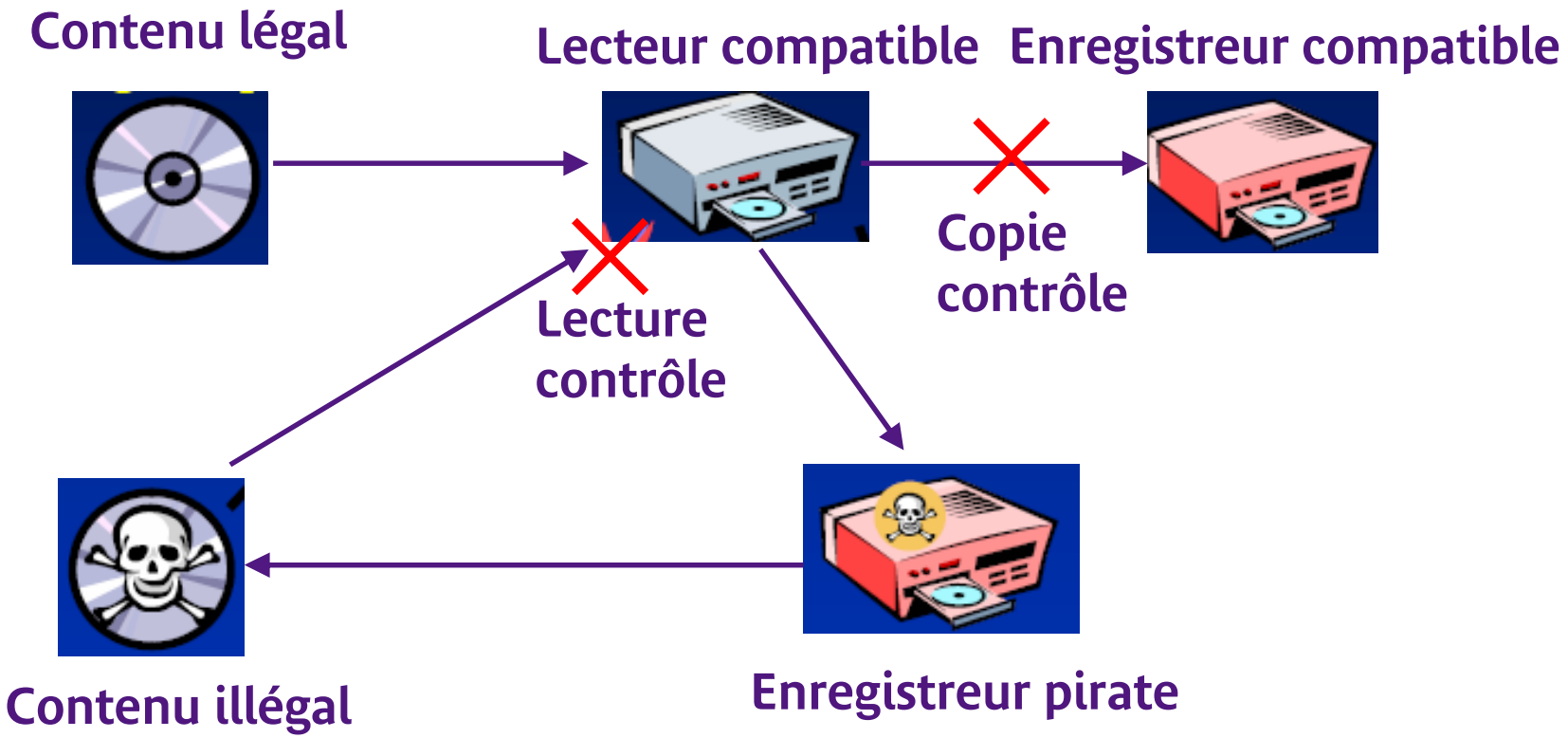
Diffusion Libre

Contrôle de Broadcast

- Contrôle si le contenu a été transmis sur les canaux de Broadcast (télévision, radio) comme convenu contractuellement

- Vérifie si la publicité est bien Broadcastée (scandale de 1997 au Japon)
- Vérifie le paiement des royalties (les artistes sont payés en fonction du temps de retransmission de leurs œuvres)
- Repérer les fraudes

Copie contrôle

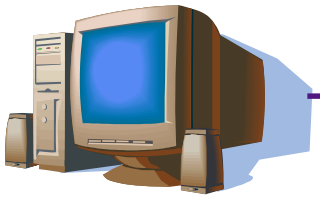


Copie contrôle

- **Insertion d'une information d'identification "copie contrôle" (ou "lecture contrôle")**
- **Le tatouage indique si le contenu peut être copié ou non**
  - Copie contrôle : les périphériques d'enregistrement contenant un détecteur refusent d'enregistrer les contenus ayant un copyright
  - Lecture contrôle : les players contenant un détecteur refusent de jouer des contenus piratés
- **Choisi pour les DVD-audio (société Verance)**

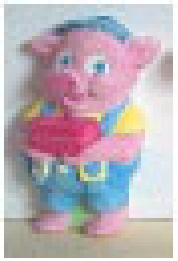
Traçage de transaction - Fingerprint

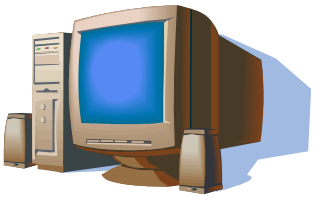
Contenu original



Marquage "naf-naf"

Diffusion

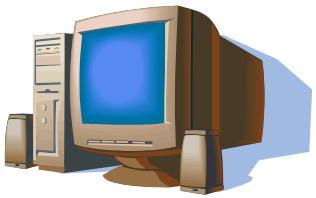




Marquage "Olrik"

Diffusion

Utilisation illégale



Détection : Olrik est Coupable !

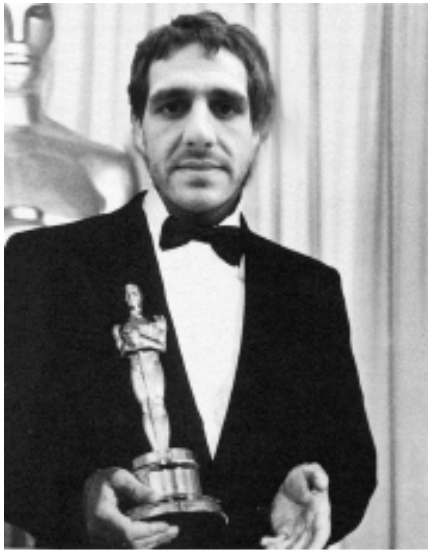
Diffusion Libre

Fingerprint

- Insertion d'une marque identifiant l'utilisateur légitime (fingerprint) dans le contenu
- Si l'utilisateur diffuse son contenu illégalement (Peer to Peer), la lecture de la marque permet de l'identifier comme pirate

Authentification

Authentification



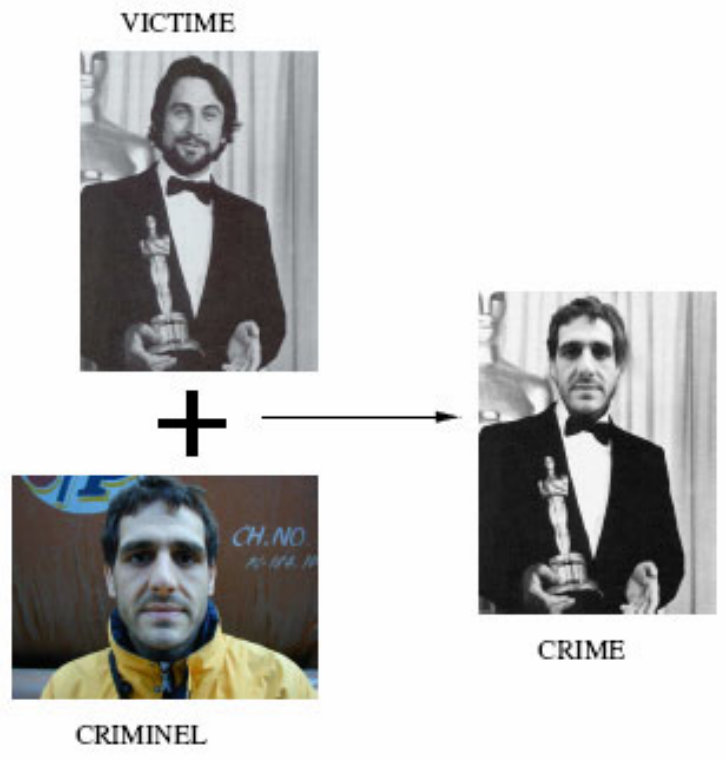
Lequel est le contenu original ?

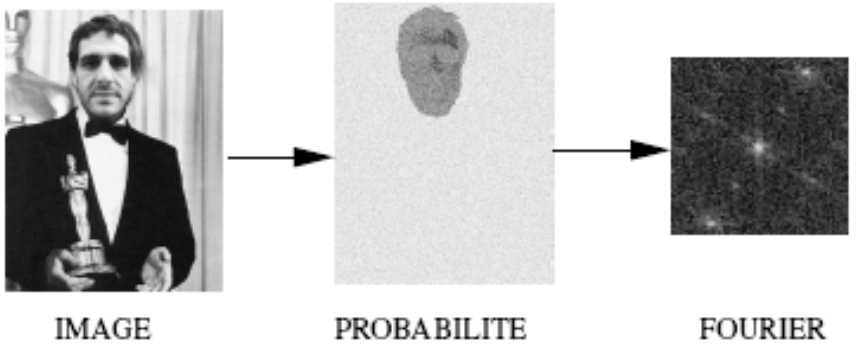
- Robert de Niro s'et affranchi de tout risque de manipulation de sa photo en y plongeant un watermark "fragile"

Authentification

- Il est nécessaire d'introduire un watermark "fragile" dans le contenu avant diffusion
- Une manipulation du contenu détruit le watermark fragile
- Utilisation complémentaire de l'intégrité cryptographique

Authentification





Manipulation

Détection

Applications et robustesse

- **La robustesse requise dépend de l'application visée**

- **Contrôle de Broadcast**
  - Doit être robuste aux transformations usuelles (compression avec perte d'information, perte de parties de contenu) ; mais pas rotations, attaques…
- **Authentification**
  - Robuste au transcodage, mais à aucune autre manipulation (watermark fragile)
- **Fingerprint**
  - Robuste à tout !

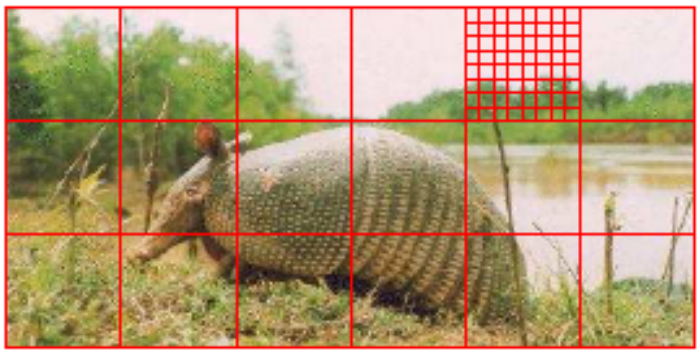
Applications et faux positif/négatif

- **Le coût des "faux" dépend de l'application**
- **Contrôle de Broadcast**
  - Coût élevé si faux négatif
- **Copie contrôle**
  - Coût élevé si faux positif
- **Fingerprint**
  - Coût peut être très élevé si erreur dans le Payload
- **Le seuil de décision dépend de l'application**

Watermark

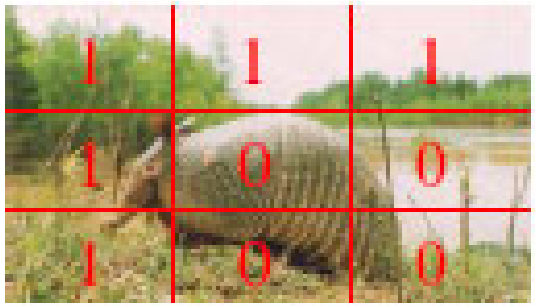
Insertion du pattern w

Insertion par bloc



$(1,1,1,1,0,0,1,0,0)$

1	1	1
1	0	0
1	0	0

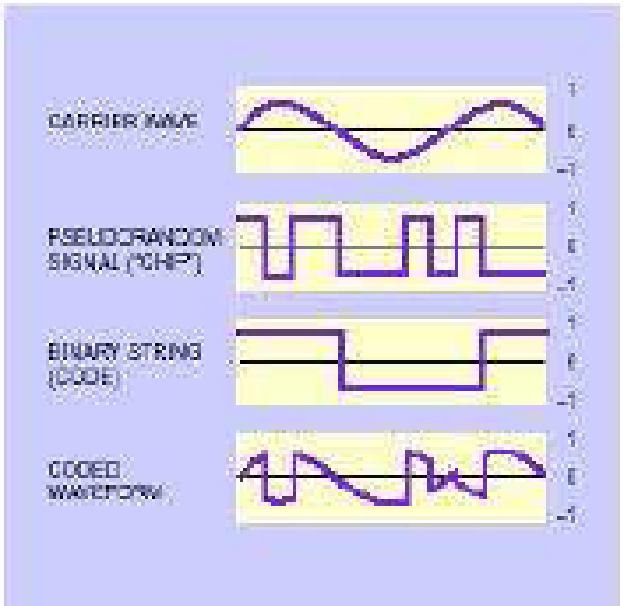


Diffusion Libre

Insertion du pattern w

Insertion par étalement de spectre

- Le contenu est considéré comme une onde porteuse
- Le Payload est masqué par une séquence pseudo-aléatoire
- La séquence obtenue module la porteuse



Une insertion simple de watermark

- **Soit…**
  - Un pattern de watermark w
  - Une image c_0
  - Un coefficient α

- **Calcul de l'image watermarkée :**

$$c_w = c_0 + \alpha w$$

Détection informée

- **Soit…**
  - Une image c éventuellement marquée
  - L'image originale c_0

- **Soustraire l'originale pour obtenir le pattern (si il est présent)**

$$w_n = c - c_0 \quad (\approx \alpha w \quad \text{si le watermark est présent)}$$

Test de corrélation linéaire

- Utiliser la corrélation linéaire pour déterminer si

$$wn \cong \alpha w$$

- La corrélation linéaire est définie comme

$$z_{lc}(w_n, w) = \frac{1}{N} w_n . w = \frac{1}{N} \sum_{x,y} w_n[x,y] w[x,y]$$

- Si $c = c_0 + n$ Alors $z_{lc}(w_n, w) \cong 0$
- Si $c = c_0 + \alpha w + n$ Alors $z_{lc}(w_n, w) \cong \alpha z_{lc}(w, w)$

Détection aveugle

- Si w est choisi tel que $z_{lc}(c_0, w)$ est probablement proche de zéro, alors $z_{lc}(c, w) \cong z_{lc}(w_n, w)$

- Il n'est alors pas nécessaire de soustraire c_0 avant de calculer la corrélation linéaire

- Si le pattern est un bruit blanc, alors il aura tendance a avoir une corrélation de faible magnitude avec une image

Seuil de détection

- Afin de déterminer si il y a présence du watermark, il faut comparer $Z_{lc}(c,w)$ à un seuil
- Si $Z_{lc}(c,w) >$ seuil, alors il y a présence de la marque
- Si $Z_{lc}(c,w) <$ seuil, alors la détection a échoué

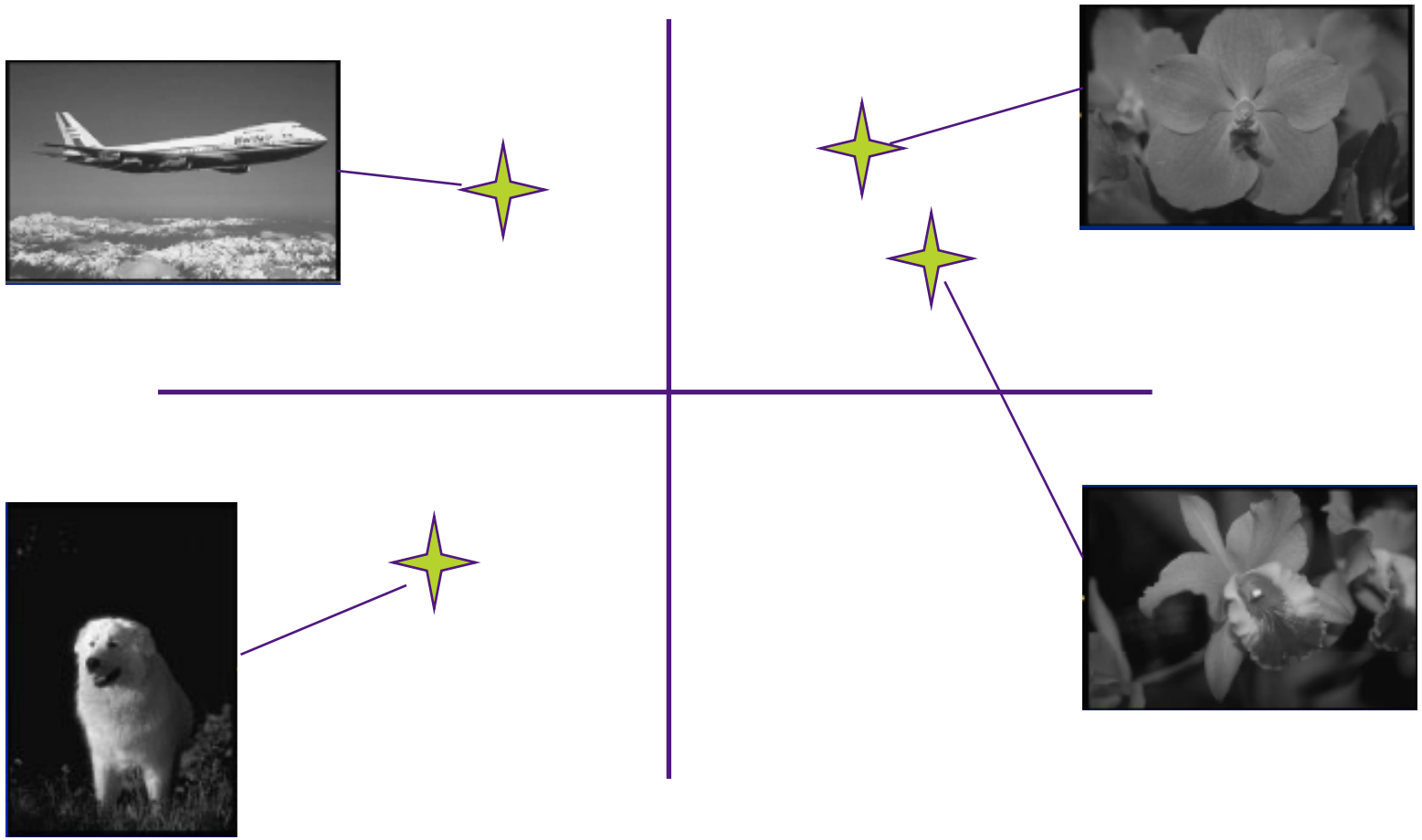
Interprétation géométrique

- L'espace des contenus est un espace de dimension élevée dans lequel chaque point correspond à un contenu

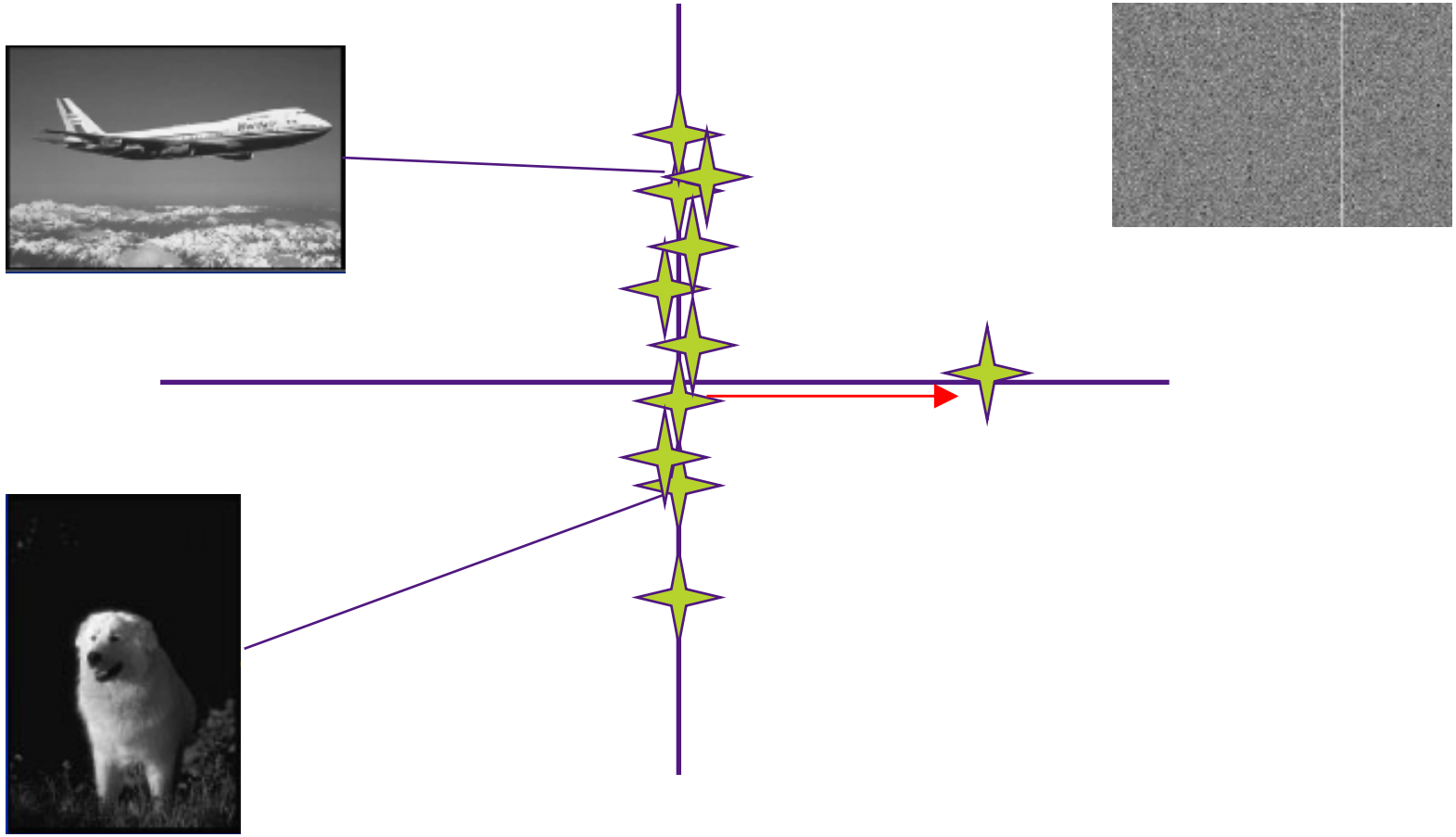
- Image en niveau de gris 256*256
  -> 65536 dimensions (une par pixel)

- 5 secondes de clip mono samplés à 44,1 kHz
  -> 220500 dimensions (une par sample)

Représentation de l'espace des contenus



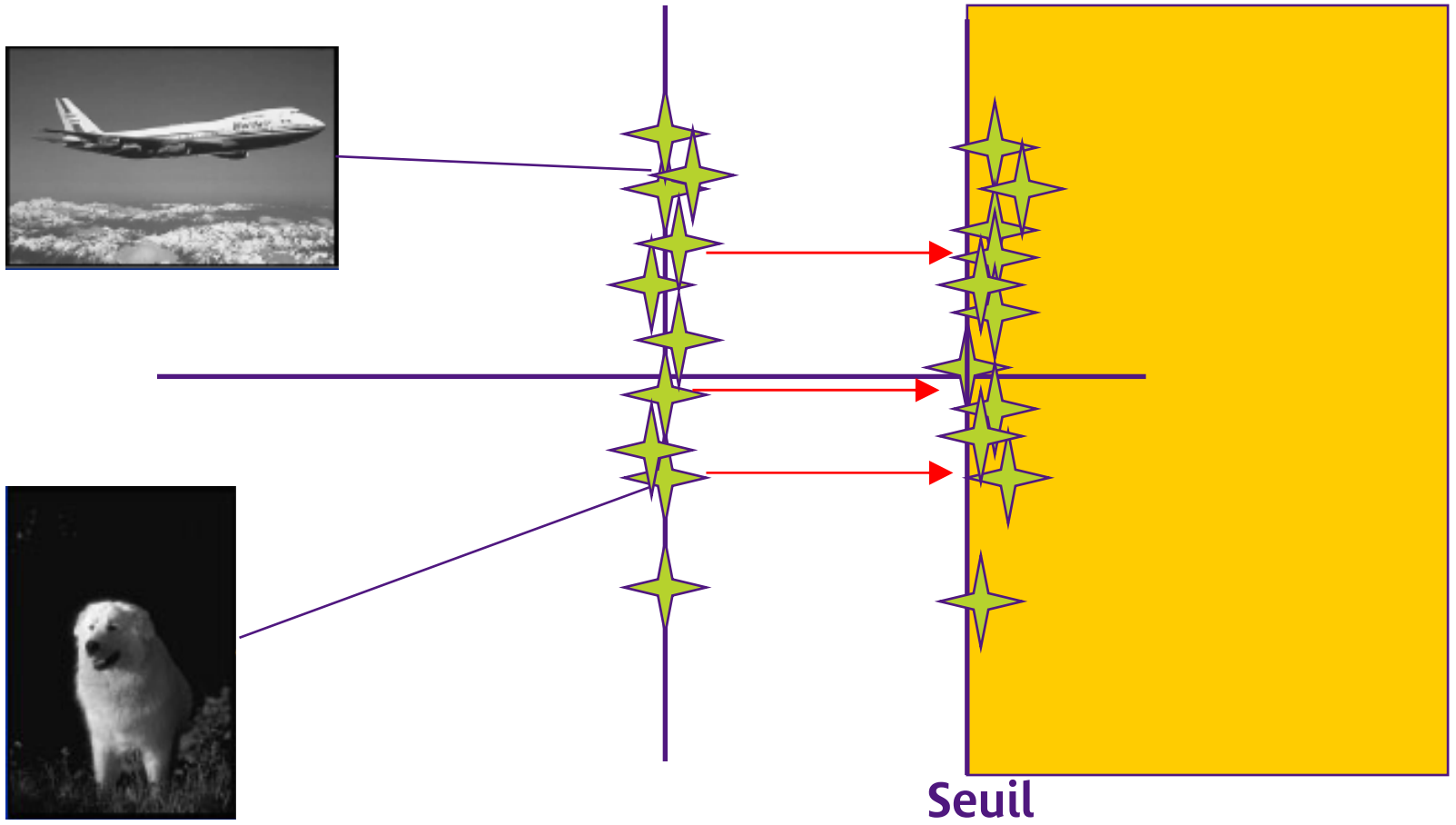
Watermark dans l'espace des contenus



Interprétation géométrique de $Z_{lc}()$

- $Z_{lc}(c,w)$ est simplement le produit scalair de c et w divisé par N
- Le produit scalair de c et w est le cosinus de leur angle multiplié par leurs longueurs respectives
- Si $|w| = 1$, alors $Z_{lc}(c,w)$ est la projection de c sur la direction w
- Comparer $Z_{lc}(c,w)$à un seuil revient à détecter une région ayant une frontière plane

Watermark dans l'espace des contenus



Attaques et contre-mesures

Changer le contraste d'une image

- Changer le contraste revient à multiplier l'image par un scalaire
- $c_{w_n} = \nu c_n$, où ν est un scalaire
- $z_{lc}(c_{w_n}, w) = \nu z_{lc}(c_w, w)$
- Si $\nu < 1$, alors la valeur de détection peut être plus petite que le seuil (même en présence de watermark)

Corrélation normalisée

- **Normalisation de la corrélation**

$$z_{nc}(c,w) = \frac{c.w}{|c||w|}$$

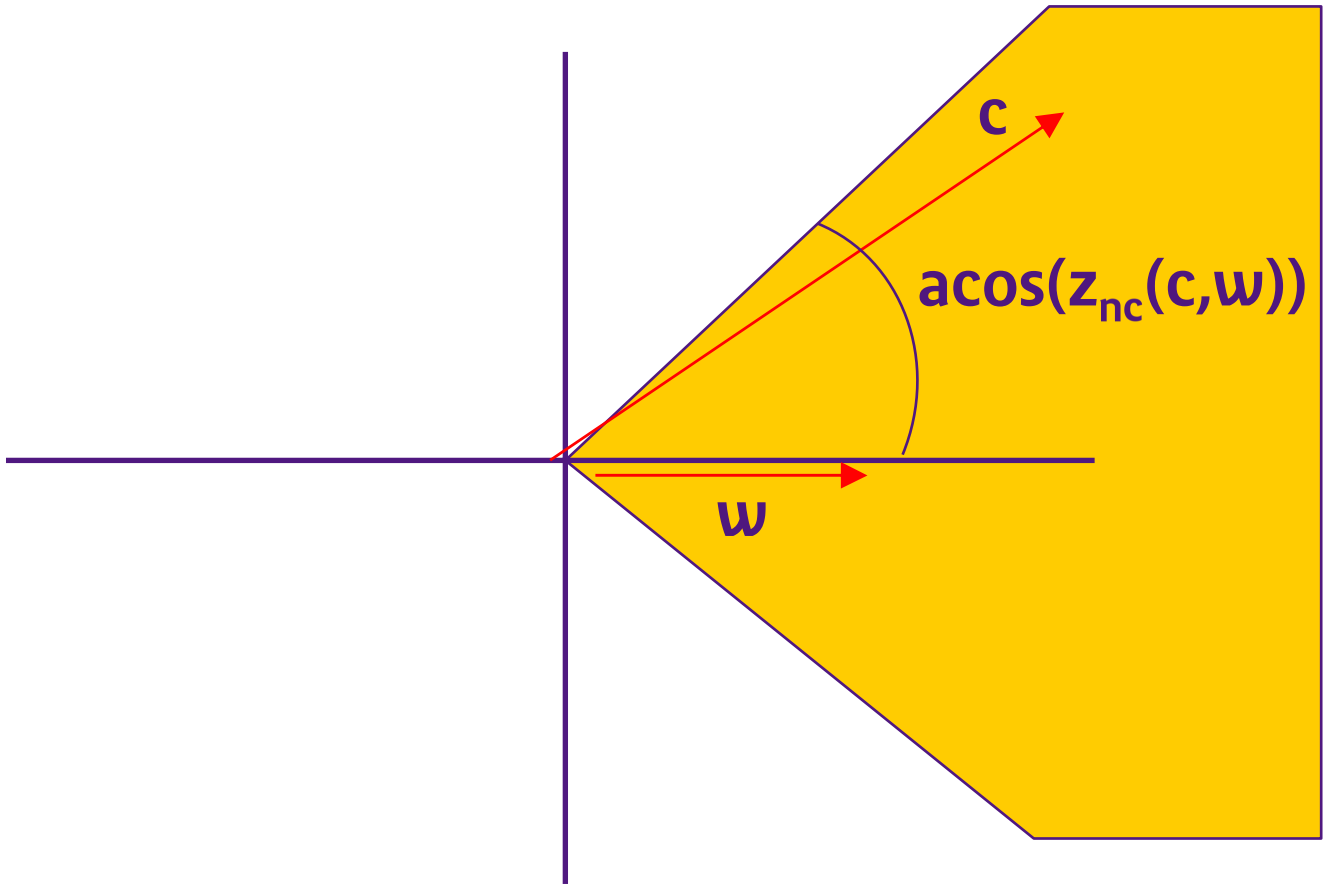
- **Le changement de contraste n'a alors aucun effet sur** $z_{nc}(c,w)$

$$z_{nc}(\nu c,w) = \frac{\nu c.w}{\nu|c||w|} = z_{nc}(c,w)$$

Interprétation géométrique de $z_{nc}()$

- La corrélation $z_{nc}(c,w)$ est simplement le cosinus de l'angle entre c et w
- Comparer $z_{nc}(c,w)$ à un seuil revient à comparer un angle à un angle limite
- Résultat : la frontière de la région de détection est un cône

Watermark dans l'espace des contenus



Attaque par translation

- **… qu'arrive-t-il si l'image subit une petite translation spatiale ?**
- **La valeur de détection va dépendre de la fonction d'auto corrélation**
- **Un bruit blanc a une auto corrélation (par rapport à son translaté) proche de zéro**
- **-> le watermark a peu de chance d'être détecté !**
  - Nous verrons plus loin comment résister à la translation

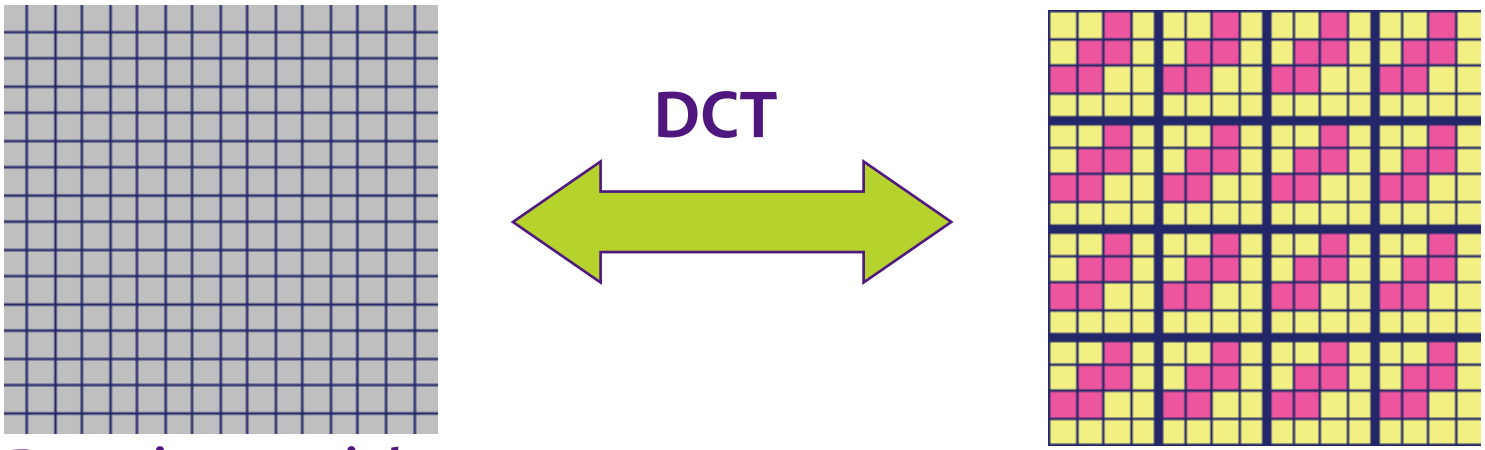
Attaque par filtre

- **Ces attaques considèrent le watermark comme un bruit additif sur le signal**
- **Suppression du watermark = débruitage de signal**
- **Filtres de débruitage :**
  - Moyen, passe-bas, stop-bande
  - Median
  - Dans les domaines : spatial, fréquentiel, d'ondelettes
  - wiener

Débruitage

Dans le domaine fréquentiel : passe-bas, stop-bande

$$\hat{c} = F^{-1}(F(h).F(c))$$



Domaine spatial

Débruitage

Image originale

Moyenne 9*9

Median 9*9

Flou Gaussien

Débruitage

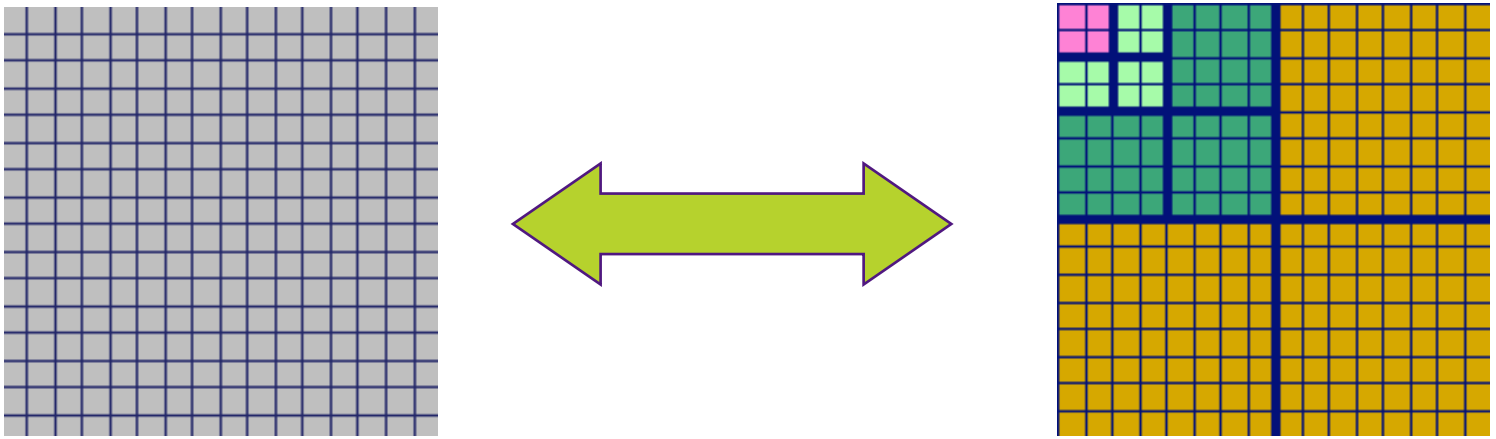
- Il est possible de viser des zones précises, si l'attaquant connaît le domaine d'insertion du watermark (ou si il arrive à le déterminer)

- Un filtre sous-bande de premier niveau affecte les bandes correspondantes du contenu

$$\hat{c} = [c_{LL1}.h_{LL1}, c_{LH1}.h_{LH1}, c_{HL1}.h_{HL1}, c_{HH1}.h_{HH1},]$$

Débruitage

- **Dans le domaine d'ondelettes : passe-bas, stop-bande**



Domaine spatial

Débruitage – Contre attaque

- Utilisation de la redondance
- Utilisation de codes correcteurs d'erreurs
- Insertion de la marque dans différentes fréquences/échelles

Attaque par ajout de bruit

- Ces attaques ne sont pas efficaces, la plupart des watermarks y résistent
- Contre mesures : codes correcteurs d'erreurs et utilisation de la corrélation

Originale

20 dB SNR

0 dB SNR

Attaques géométriques

- **Objectif : désynchroniser le média pour leurrer la procédure de détection de la marque**
  - Il est plus difficile de trouver la marque si on ne sait pas où chercher !
  - Ces attaques ne suppriment pas la marque

- **Désynchronisation :**
  - Translation, rotation, changement d'échelle
  - Combinaison (linéaire ou non linéaire, locale ou globale)
  - Effacement/échange de trames (son, vidéo)

Attaques géométriques

- **Attaque globale (ou stationnaire)** $\hat{c}(i,j) = c(T(i,j))$

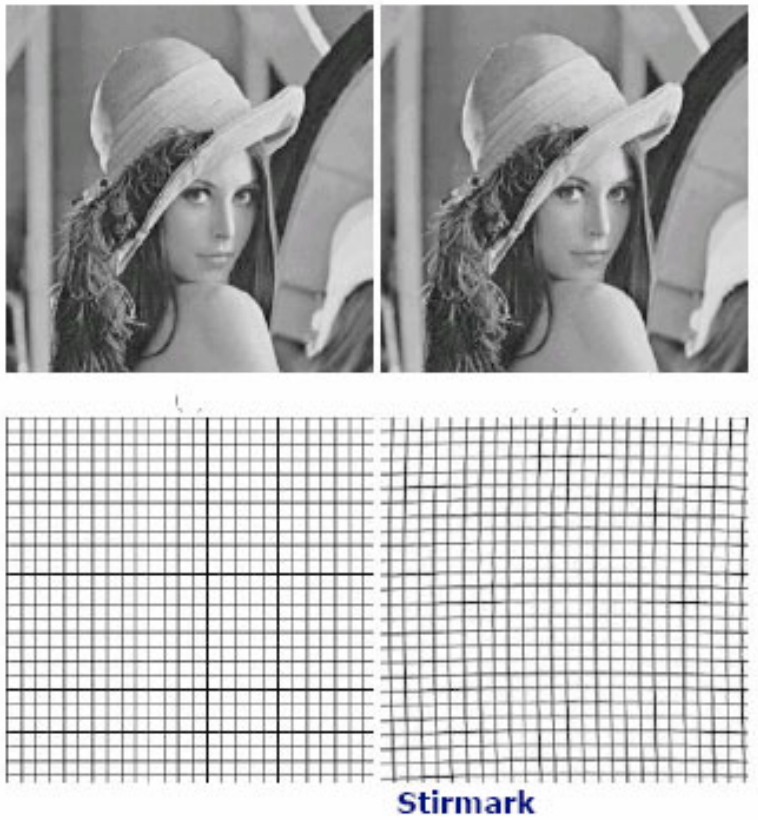
- **Rotation, translation, changement d'échelle**

$$T(i,j) = \begin{bmatrix} kcos(\theta) & ksin(\theta) & t_x \\ -ksin(\theta) & kcos(\theta) & t_y \end{bmatrix} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix}$$

- **Décalage circulaire**

$$\hat{c}(i,j) = c(i + \Delta_x \mod M, j + \Delta_y \mod N)$$

Attaques géométriques locales



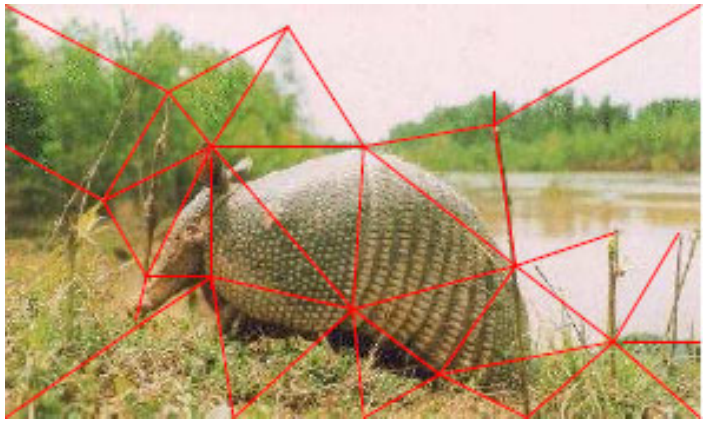
Résistance aux transformations

- **Plusieurs stratégies afin de rendre un watermark robuste aux transformations**
  - Invariance
  - Inversion de transformation au niveau du détecteur
  - Insertion du watermark dans des zones significativement perceptibles du contenu

Invariance

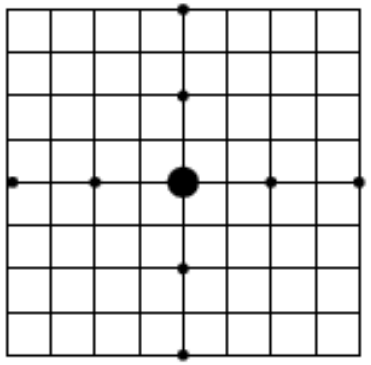
- La transformation de Fourier est invariante par translation
- La transformation de Fourier-Mellin est invariante par rotation, zoom, translation

- Insertion du pattern selon des points "remarquables" du contenu



Inversion de transformation

- Un pattern auto corrélé est additionné au contenu
- Le pattern est une séquence d'étalement de spectre de moyenne nulle
- Le plongement a lieu dans le domaine DCT
  => meilleur résistance



- Fonction d'auto corrélation judicieusement choisie
- Processus de détection du pattern récursif

Attaque cryptographique

- **Contre-mesure : utiliser une cryptographie solide !**

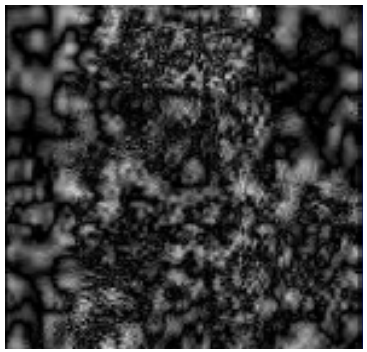
Attaque par estimation

- **La marque ou le contenu est estimé**
  - Sans la connaissance des secrets
  - Avec de l'information sur
    - La marque
    - Le contenu
- **Peut être utilisé quand une grande quantité de données est disponible**
  - Avec de nombreux contenus watermarkés, ou un contenu de grande taille
- **Débruitage optimal, compression parfaite, réduction de la taille de l'espace des secrets**

Attaque par estimation

Contenu watermarké

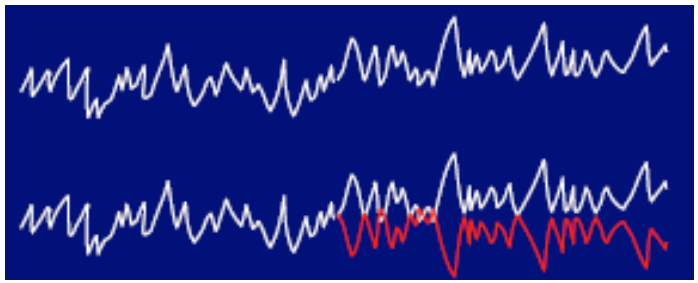
Estimation du watermark



Estimation de l'original

Attaque par estimation

- **Avec une estimation de la marque, un attaquant peut remoduler :**
  - Soustraire le watermark estimé du contenu marqué (modulation opposée)
  - Le but est d'abaisser la valeur de détection en dessous du seuil, grâce à des corrélations négatives



Attaque par estimation, contre-mesure

- **Afin de prendre en défaut une estimation basée sur l'approximation des moindre carrés de plusieurs contenus**

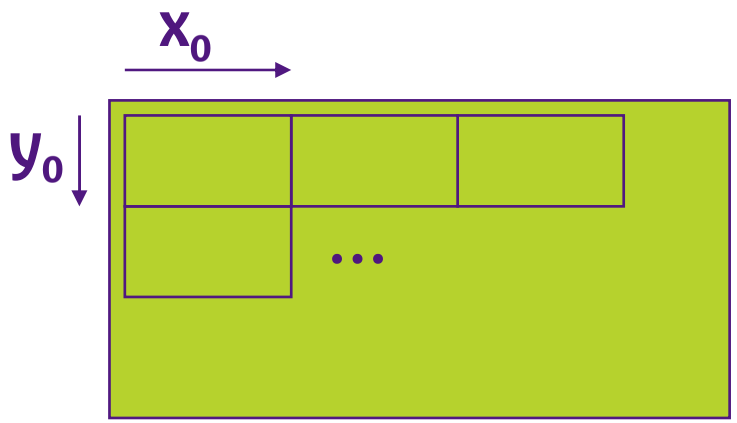
- **=> Insérer un watermark dont la puissance spectrale est proportionnelle à la puissance spectrale du contenu**

Estimation - exemple

- **Watermark par dissimulation d'écho (echo hiding)**
  - $C_w(t) = c(t) + \alpha c(t+\delta)$
  - Ce watermark est aisément détectable
  - De plus, il est possible de l'estimer puis de l'effacer directement
- **Challenge SDMI (2000) a subi des attaques par estimation de ce type**

Estimation - exemple

- **Si le pattern d'un watermark détecté par corrélation est répété pour augmenter la redondance :**



- **ArgMax (δ_x, δ_y) $Z_{nc}(c_w(x,y),c_w(x+\delta_x,y+\delta_y) = (x_0,y_0)$**
  - L'attaquant obtient de l'information utile pour l'estimation

Attaque de protocole

- En analysant le protocole
- en tirant partie de l'algorithme de watermark utilisé

- Il est parfois possible de concevoir une attaque prenant en défaut le protocole d'utilisation du watermark (et non le watermark lui-même)

- Un grand nombre d'attaques de protocole existe dans les schémas d'utilisation du watermark pour la preuve de copyright

La marque inversible

Attaque d'inversion

- Si Olrik peut inverser le moteur de tatouage

- Alors $f_{W_O}(f^{-1}_{W_O}(f_{W_B}(C))) = f_{W_B}(C)$
- Qui a marqué $f_{W_B}(C)$???
- Blake peut garder le contenu original pour se prélunir d'une telle attaque
- Il peut aussi utiliser un tatouage non inversible

La marque ambiguë

Attaque d'ambiguïté

- Olrik peut parfois trouver une marque $\hat{W}$ présente dans $f_{W_B}(C)$ et dans C en analysant l'algorithme de tatouage

C, W_B

$f_{W_B}(C)$

$f^{-1}_{\hat{W}}(f_{W_B}(C)), \hat{W}$

Exemple si l'algorithme de détection est une corrélation :

- détection par corrélation si $\langle W, f_W(C) \rangle > seuil$ alors $f_W(C)$ est déclaré marqué avec W
- $\hat{W} = \mathcal{F}_{haut}(f_{W_B}(C))$ un filtre passe-haut
- alors $\langle \hat{W}, C \rangle > seuil$

Codes de Costa

Codage informé

- **Un payload significativement plus grand peut être inséré si la construction du pattern est dépendant du contenu dans lequel il sera inséré**

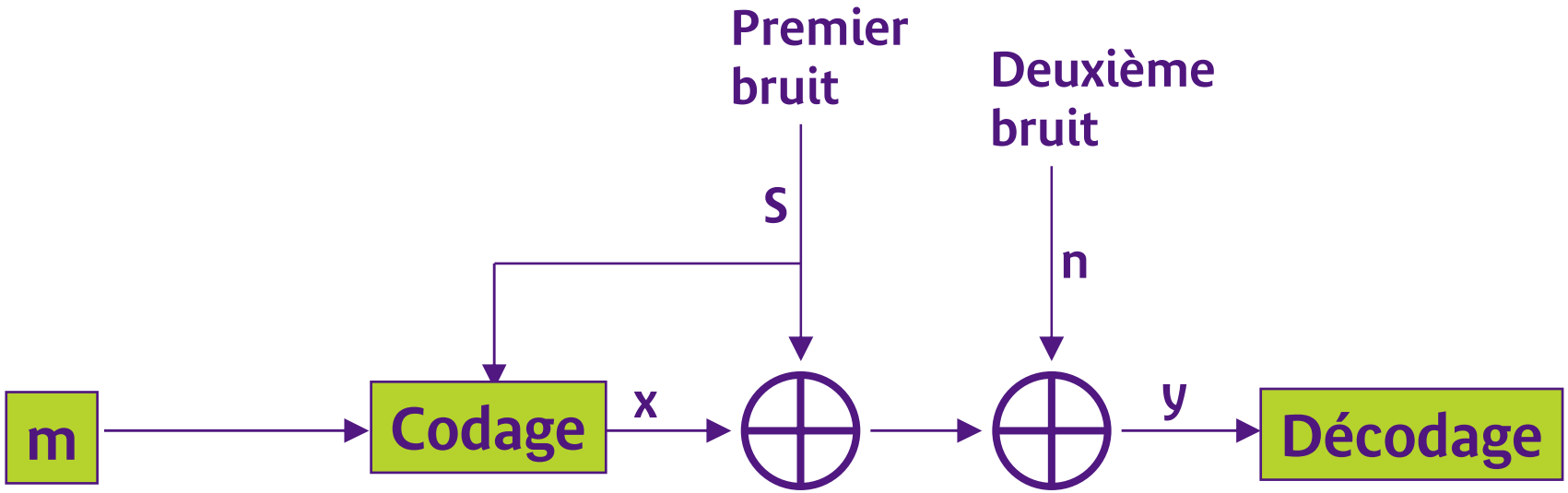
Codage informé

- "Writing on dirty paper" (écrire sur un papier sale – problème étudié par M. Costa)
- Dirty Paper codes
- Application de ces codes au watermarking

Writing on dirty paper

- **M. Costa a étudié le problème du "dirty paper channel"**
  - Débuter par un "papier sale" aux taches normalement distribuées (distribution normale)
  - Écrire un message en utilisant une quantité limitée d'encre
  - Envoyer le message qui sera de nouveau taché (selon une distribution normale)
  - Le décodeur ne peut pas distinguer l'encre des taches
  - Quelle quantité d'information peut-on envoyer ?

Writing on dirty paper



X est limité par une contrainte de puissance

$$\sum_i x[i]^2 \leq p$$

Dirty paper codes

- **Le premier bruit n'a pas d'effet sur la capacité du canal !**

Dirty paper codes

- **Idée de base :**
- **Dirty paper code : chaque mot du code (vecteur) est représenté par plusieurs mot de code alternatifs**
- **Parmi l'ensemble des mots de code qui représentent le message original, choisir celui qui est le plus proche du premier bruit**

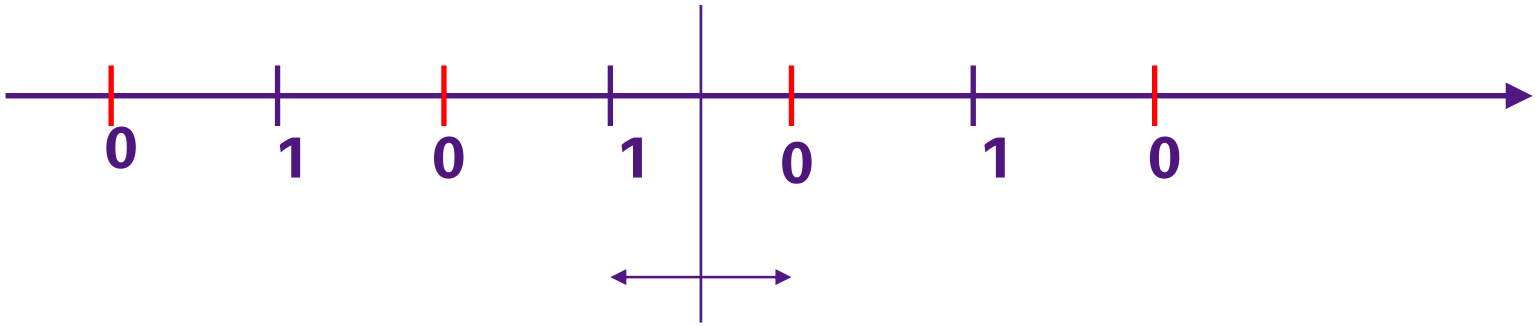
Costa code

- **En pratique :**

- **Le code de Costa est généré aléatoirement**
  - Il requière une recherche exhaustive pour les phases de codage et décodage
  - Efficace pour des petites tailles de payload
- **Lattice-code sont plus efficaces en pratique (et plus étudiés)**
  - Chen & Wornell ("Dither Index Modulation", "Quantization Index Modulation")
  - Egger, Su & Girod ("Scalar Costa Scheme")

Lattice code

- Chaque dimension dans l'espace encode un symbole (habituellement un Bit)
- Les Bits sont encodés en choisissant entre deux quantisateurs (selon un pas de quantisation)



Lattice code

- **En pratique, les "Lattice code" ont une capacité d'insertion bien plus élevée que les systèmes basés sur la corrélation**

- **Ils sont en général moins robuste que les systèmes basés sur la corrélation**
  - Les systèmes basés sur la corrélation ont en général un meilleur rapport Payload/Robustesse en présence de bruit élevé
  - Les "Lattice-code" sont susceptibles de changer la luminance de l'image ou le volume du son…

Treillis code

- Les "Treillis code" sont une alternative au "Lattice code"
- Ils ont été spécialement conçues pour résister aux modifications volumétriques (contrastes d'une image ou volume du son)

Attaque sur les codes de Costa

- **La connaissance (ou la découverte) du pas de quantisation permet à un attaquant**
  - D'effacer sa marque
  - De créer un nouveau pattern
  - Voire de créer un nouveau payload si l'attaquant arrive à savoir comment le pattern est construit à partir du payload
  - => utilisation nécessaire de cryptographie