Techniques sûres de tatouage pour l'image

B. Mathon[1] F. Cayre[1] P. Bas[1]

[1] GIPSA-Lab Dépt. Images et Signal, UMR CNRS 5216

961 rue de la Houille Blanche
Domaine universitaire - BP 46
F - 38402 Saint Martin d'Hères cedex

{benjamin.mathon, cayre, patrick.bas}@gipsa-lab.inpg.fr

Résumé

Cet article présente les premiers résultats obtenus en appliquant des techniques prouvées sûres pour le tatouage des images. Deux techniques sûres, le tatouage naturel (NW) et le tatouage circulaire (CW), sont comparées à deux techniques classiques : l'étalement de spectre (SS) et l'étalement de spectre amélioré (ISS). Les résultats sont analysés tant sous l'angle de la robustesse et de la distorsion que de la sécurité. Les résultats portent sur des tests effectués sur une base de 2000 photos de vacances immontrables. Ils permettent d'identifier la classe de sécurité la plus haute atteignable en pratique en tatouage d'image. Cette classe est connue sous le nom de sécurité des clefs (key-security).

Mots clefs

Tatouage, sécurité, image

1 Introduction

On s'intéresse, d'après la classification établie dans [1] au cas où l'attaquant ne peut observer que des images tatouées (il ne connaît ni les messages insérés ni les images originales). Ce cadre d'analyse est appelé WOA (Watermark Only Attack). Toujours dans [1], les auteurs proposent des niveaux de sécurité pour les méthodes de tatouage non sûres : la sécurité en tatouage consiste à montrer l'impossibilité ou la possibilité d'estimer les mots de code utilisés pour coder l'information cachée. Récemment, deux techniques prouvées sûres ont été proposées dans le cadre WOA : le tatouage naturel (NW) et le tatouage circulaire (CW). Cet article expose les premiers résultats obtenus dans un cas pratique (l'image) en utilisant ces techniques. Toutes les techniques de tatouage abordées ici sont fondées sur l'étalement de spectre [2] [3]. Le lecteur est renvoyé aux références pour les détails concernant les techniques utilisées ici.

2 Techniques d'étalement de spectre

Formellement, le tatouage par étalement de spectre peut être modélisé comme suit : on cherche à insérer un message binaire de N_c bits $\mathbf{m} \in \{0,1\}^{N_c}$ dans un vecteur $\mathbf{x} \in \mathbb{R}^{N_v}$. Ce vecteur est censé capturer la plupart des informations perceptuelles sur le contenu à tatouer. Par exemple, $\mathbf{x}$ peut être issu d'une sélection de coefficients DCT par bloc, d'un ensemble de coefficients d'ondelettes, etc.

2.1 Construction du signal tatoué

Le signal tatoué est construit comme dans [1]. Le message $\mathbf{m}$ est codé en utilisant N_c porteuses $\mathbf{u}_i \in \mathbb{R}^{N_v}$ que l'on assimile à des mots de code. Ces porteuses sont issues d'un générateur pseudo-aléatoire initialisé avec une graine K qui fait office de clef secrète. Les porteuses $\mathbf{u}_i$ sont des vecteurs gaussiens de loi $\mathcal{N}(0,1)$ et forment une base orthogonale, i.e. $\forall i \neq j\ ,\ <\mathbf{u}_i|\mathbf{u}_j>=0$ où $<.|.>$ désigne le produit scalaire. La construction du signal de tatouage $\mathbf{w}$ nécessite une modulation $s : 0,1 \rightarrow \mathbb{R}$:

$$\mathbf{w} = \sum_{i=1}^{N_c} \mathbf{u}_i s(\mathbf{m}(i)). \tag{1}$$

On ajoute ensuite $\mathbf{w}$ à $\mathbf{x}$ pour former $\mathbf{y}$ le vecteur tatoué :

$$\mathbf{y} = \mathbf{x} + \mathbf{w}. \tag{2}$$

Le décodage du message tatoué produit une estimation $\hat{\mathbf{m}}$:

$$\hat{\mathbf{m}}(i) = \text{signe}(<\mathbf{y}'|\mathbf{u}_i>) \tag{3}$$

où $\mathbf{y}'$ est le vecteur tatoué $\mathbf{y}$, éventuellement attaqué.
On évalue la distorsion causée par le tatouage à l'aide du WCR (Watermark to Content Ratio) :

$$WCR = 10\log_{10}\left(\frac{\sigma^2_{\mathbf{w}}}{\sigma^2_{\mathbf{x}}}\right). \tag{4}$$

Formellement, le problème de la sécurité en tatouage dans le cadre WOA consiste à estimer les porteuses $\mathbf{u}_i$. Il n'est pas nécessaire de remonter jusqu'à la clef K pour porter une attaque de sécurité. L'hypothèse fondamentale est ici que les messages sont tirés indépendamment les uns des autres. En effet, si l'on note les différents vecteurs en colonnes, on obtient la formulation suivante pour N_o contenus tatoués :

$$\mathbf{Y} = \mathbf{X} + \mathbf{US}, \tag{5}$$

avec $\mathbf{Y}, \mathbf{X} \in \mathcal{M}_{N_v,N_o}(\mathbb{R})$, $\mathbf{U} \in \mathcal{M}_{N_v,N_c}(\mathbb{R})$, $\mathbf{S} \in \mathcal{M}_{N_c,N_o}(\mathbb{R})$.
On reconnaît donc un problème de séparation de sources en aveugle. Les porteuses $\mathbf{U}$ jouent le rôle de la matrice de mélange, et les sources $\mathbf{S}$ sont les modulations des messages. Si les messages sont supposés indépendants, alors les techniques d'analyse par composantes indépendantes (ICA) permettent d'estimer $\mathbf{U}$ et la technique de tatouage est alors réputée non sûre. Plus spécifiquement, l'ICA ne permet que d'estimer les directions des $\mathbf{u}_i$ et non leur sens ni leur ordre. Lever ces indéterminations fondamentales liées à la séparation de sources requerrait de connaître quelques messages, ce qui dépasse le cadre de l'étude proposée ici.

2.2 Modulations

L'étalement de spectre classique (SS) utilise une modulation simple (analogue de la modulation BPSK en communications numériques). Le paramètre γ permet de régler l'ampleur de la distorsion :

$$s_{SS}(\mathbf{m}(i)) = \gamma(-1)^{\mathbf{m}(i)} \tag{6}$$

Dans [3] les auteurs améliorent sensiblement les performances de l'étalement de spectre en jouant sur l'information adjacente connue à l'insertion :

$$s_{ISS}(\mathbf{m}(i)) = \alpha(-1)^{\mathbf{m}(i)} - \lambda\frac{<\mathbf{x}|\mathbf{u}_i>}{\|\mathbf{u}_i\|^2} \tag{7}$$

où α et λ sont des paramètres ajustés pour miniser la probabilité d'erreur et maximiser la robustesse.
Dans [4], une nouvelle modulation est proposée, dite de tatouage naturel. Elle a pour but de conserver la distribution des $<\mathbf{x}|\mathbf{u}_i>$ (réputés suivre une loi normale), moyennant un facteur d'échelle $\eta \geq 1$:

$$s_{NW}(\mathbf{m}(i)) = \left(\eta(-1)^{\mathbf{m}(i)}\frac{<\mathbf{x}|\mathbf{u}_i>}{|<\mathbf{x}|\mathbf{u}_i>|} - 1\right)\frac{<\mathbf{x}|\mathbf{u}_i>}{\|\mathbf{u}_i\|^2}. \tag{8}$$

Dans [5], une autre modulation, dite de tatouage circulaire, est proposée : elle a pour but d'améliorer la robustesse de la modulation précédente (en se fondant sur la modulation ISS) tout en garantissant une sécurité sur les clefs (cf. infra) :

$$s_{CW}(\mathbf{m}(i)) = \alpha(-1)^{\mathbf{m}(i)}\mathbf{d}(i) - \lambda\frac{<\mathbf{x}|\mathbf{u}_i>}{\|\mathbf{u}_i\|} \tag{9}$$

où α et λ sont les mêmes que pour ISS, et $\mathbf{d}$ est généré comme suit à partir d'un vecteur $\mathbf{g} \sim \mathcal{N}(0,1)$:

$$\mathbf{d}(i) = \frac{|\mathbf{g}(i)|}{\|\mathbf{g}\|}. \tag{10}$$

Les modulations SS et ISS sont connues pour n'être pas sûres : il est possible, en observant des images tatouées, et sous l'hypothèse de messages indépendants, d'estimer les mots de code (les porteuses $\mathbf{u}_i$). En revanche, la modulation CW ne permet que d'estimer le sous-espace privé vect$(\mathbf{u}_i)$. La modulation NW, quant à elle, permet, lorsque $\eta = 1$, de faire de la stéganographie (i.e. $D_{KL}(p(\mathbf{x})\|p(\mathbf{y})) = 0$ et il n'est donc pas possible de décider si un vecteur est tatoué ou pas). Lorsque $\eta > 1$, la modulation NW permet de décider si un contenu est tatoué et d'estimer le sous-espace privé.

3 Classes de sécurité

La définition des classes de sécurité utilisées en WOA sont tirées de [6]. On en rappelle brièvement la classification :
– stégo-sécurité : le pirate n'est pas capable de décider si un contenu est tatoué ou pas ;
– sous-espace-sécurité : le pirate ne peut pas exhiber le sous-espace privé dans lequel a lieu l'opération de tatouage ;
– clef-sécurité : le pirate peut exhiber le sous-espace privé, mais ne peut pas obtenir davantage d'information ;
– non-sécurité : le pirate peut estimer le secret avec lequel on a tatoué.

En pratique, la classe de sécurité des clefs marque la limite qui existe entre robustesse et sécurité : l'attaque optimale consistant à se focaliser sur le sous-espace privé afin de minimiser la distorsion d'attaque est encore possible, mais pas l'estimation plus précise du secret.
Lorsque $\eta = 1$, le tatouage naturel est stégo-sûr. Lorsque $\eta > 1$, il assure la sécurité des clefs. Le tatouage circulaire assure lui aussi la sécurité des clefs. Les modulations classiques SS et ISS ne sont pas sûres. Le but de ce papier est de montrer quelle classe de sécurité on peut obtenir en pratique dans le cas d'une attaque WOA.

4 Implantation

La plupart des méthodes théoriques de tatouage font l'hypothèse d'un contenu $\mathbf{x}$ distribué suivant une loi normale. Cette modélisation est incompatible avec la plupart des transformées connues : les coefficients DCT peuvent se modéliser à l'aide d'une loi laplacienne, les coefficients d'ondelette se prêtent mieux à un modélisation par une loi de Gauss généralisée, quand les pixels eux-mêmes peuvent se modéliser à l'aide d'une mixture de gaussiennes. Cette hypothèse de gaussianité de $\mathbf{x}$ sert principalement à s'affranchir de calculs pénibles pour évaluer la distorsion causée par le tatouage. Toutefois, pour rester au plus près de cette hypothèse, nous avons développé l'algorithme de tatouage présenté ci-après.

4.1 Espace de tatouage

Une image de taille (M, N) est transformée en ondelettes (schéma de lifting avec filtre 9/7 de Daubechies) avec 4 niveaux. On range dans un vecteur $\mathbf{x}_t \in \mathbb{R}^{N_t}$ les coefficients d'ondelette des 9 sous-bandes de plus hautes fréquences. Le vecteur $\mathbf{x}$ est obtenu comme suit :

$$\mathbf{x}(i) = \frac{2}{\sqrt{3N_t}}\sum_{j}^{N_t}\mathbf{x}_t(j)\mathbf{a}_i(j) \tag{11}$$

où les $\mathbf{a}_i$ sont des vecteurs pseudo-aléatoires dont les coefficients sont tirés suivant une loi uniforme centrée en zero et normalisés pour obtenir un vecteur de norme unité. On assure ainsi par le théorème central limite que les $\mathbf{x}$ suivent asymptotiquement une loi normale ($N_t >> 1$). Le tatouage est ensuite suivi de la rétro-projection adéquate permettant de retrouver les coefficients d'ondelette tatoués. Il est à remarquer que cette projection est quasi-orthogonale (plus rapide à appliquer qu'une transformation orthogonale).

4.2 Dimensionnement

Le but de ce papier est de montrer quelle classe de sécurité est atteignable en pratique dans le cas d'une attaque en WOA. Aux fins d'illustration, on cachera donc $N_c = 10$ bits dans chaque image. Les vecteurs $\mathbf{x}$ et $\mathbf{x_t}$ ont une longueur de $N_v = 256$ et $N_t = 258048$, respectivement.

Concernant le tatouage naturel, un autre problème de dimensionnement se pose : cette technique ne permet pas de tatouer lorsque η est trop petit. En particulier, la classe de sécurité où $\eta = 1$, la stégo-sécurité, ne peut être atteinte en pratique. L'explication est due à l'extrême fragilité de la méthode, qui produit une marque disparaissant à la quantification des pixels sur 8 bits pour produire l'image tatouée.

5 Résultats

Nous ne nous intéressons pas au problème de la resynchronisation, globale ou locale, après attaque géométrique. Seules seront évaluées la robustesse et la sécurité, sous contrainte d'un budget de distorsion pour l'insertion de la marque.

5.1 Distorsion

Pour fixer la distorsion, on a relié trivialement le WCR au PSNR par la relation suivante :

$$WCR = 10\log_{10}\left(\frac{255^2}{\sigma_{\mathbf{x}}^2}\right) + 10\log_{10}\left(\frac{MN}{N_v}\right) - PSNR. \quad (12)$$

Au passage, la quantification destinée à ramener l'image tatouée sur 8 bits introduit ici une petite variation, négligeable, sur la distorsion atteinte même dans le cas de la modulation SS (pourtant censée, au contraire des trois autres modulations, offrir une distorsion exacte et non une distorsion atteinte en espérance). On résume dans le Tab. 1 les résultats obtenus par notre algorithme pour notre ensemble d'images et pour chaque modulation, en fixant une distorsion cible de 45dB.

On constate que seule la modulation NW dépasse (de peu) le budget autorisé, en espérance. En outre, il s'agit de la modulation présentant la plus forte variation d'une image sur l'autre. Toutefois, la grande qualité (45dB) requise pour l'insertion de la marque autorise des écarts assez substantiels numériquement sans être visuellement perceptibles.

Modulation	$\mathbb{E}[PSNR](dB)$	$\sigma_{PSNR}(dB)$
SS	44.75	1.18e-1
ISS	44.76	2.17e-1
NW	45.19	1.89e0
CW	44.76	2.16e-1

Tableau 1 – *Distorsion causée par l'insertion de la marque, pour les quatre modulations. PSNR cible : 45dB.*

5.2 Robustesse

La robustesse a été évaluée à l'aide d'une attaque non optimale mais usuelle : une compression JPEG. Pour plus de clarté, nous n'avons pas représenté toute la plage de variation du facteur de qualité Q. Nous n'avons caché que $N_c = 10$ bits. Un tel nombre de bits cachés est déjà bien plus que suffisant pour les applications de protection de la copie, mais reste encore insuffisant pour les cadres usuels d'application de protection des droits (tout au moins tels qu'envisagés classiquement). Des expériences visant à estimer la robustesse pour un plus grand nombre de bits cachés sont envisagées. En l'état actuel des travaux, on comparera la robustesse des différentes modulations face à la compression JPEG sur la Fig. 1.

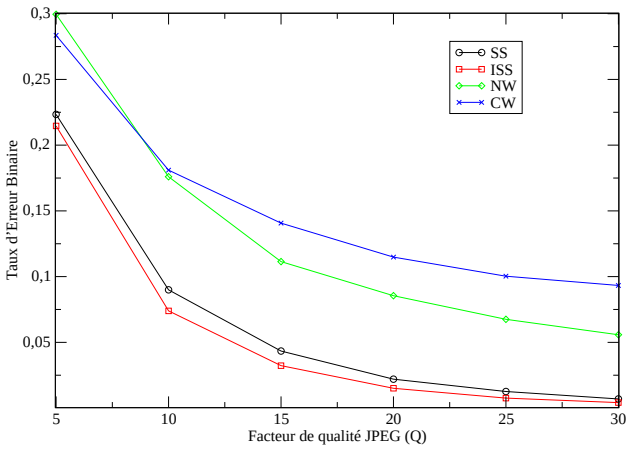

Figure 1 – *Robustesse des modulations proposées pour la compression JPEG*

La Fig. 1 permet de mesurer quel est en pratique le coût de la sécurité en terme de robustesse (celle-ci peut paraître élevée par rapport aux résultats théoriques, mais ceux-ci sont exprimés en fonction d'un ajout de bruit gaussien, ce qui n'est pas le genre de bruit produit par une compression JPEG). Rares toutefois sont les applications qui peuvent se permettre des trous de sécurité.

5.3 Sécurité

Dans le cadre d'attaque qui nous intéresse ici (WOA), le pirate n'a accès à aucun message. Il ne peut que faire l'hypothèse de messages tirés indépendamment les uns des autres.

En outre, les techniques de séparation de sources qu'il utilisera auront toutes les deux mêmes limitations :

1. le signe de la porteuse sera indéterminé (on n'estime que des directions, sans le sens) ;
2. l'ordre des porteuses sera aussi indéterminé.

Dans un cas réel d'attaque, et en raisonnant en intégrant le principe de Kerckhoffs [7], on doit supposer que le pirate dispose du code source de l'algorithme de tatouage qu'il tente d'attaquer. Il peut donc estimer un nombre d'observations nécessaires à une bonne estimation des porteuses. Il se servira de ce nombre d'observations pour dimensionner son attaque sur le secret qu'il cherche réellement à estimer. En conséquence des limitations ci-dessus, nous n'avons que la manière suivante de calculer le score S obtenu par le pirate dans l'estimation des porteuses :

$$S = \frac{1}{N_c} \sum_i \left(\max_j^1 |z(\mathbf{u}_j, \hat{\mathbf{u}}_i)| - \max_j^2 |z(\mathbf{u}_j, \hat{\mathbf{u}}_i)| \right) \quad (13)$$

où $\hat{\mathbf{u}}_i$ représentent les porteuses estimées par ICA [8], et $\max^{1,\text{resp. } 2}$ le premier (resp. second) maximum de la valeur absolue de la corrélation normalisée z entre chaque porteuse estimée $\mathbf{u}_i$ et chaque porteuse estimée $\hat{\mathbf{u}}_j$. Une telle mesue a déjà été utilisée avec succès [9].
En gardant à l'esprit que les $\mathbf{u}_{i \neq j}$ forment une base, il devient évident qu'une estimation correcte des porteuses produira un score S se rapprochant de 1 avec le nombre d'images observées. Au contraire, une mauvaise estimation des porteuses produira un score S plutôt petit, sans pouvoir évidemment être nul. D'autres comportements de S ne sont pas à exclure dans le cas d'une mauvaise estimation des porteuses, comme le montre la Fig. 2.

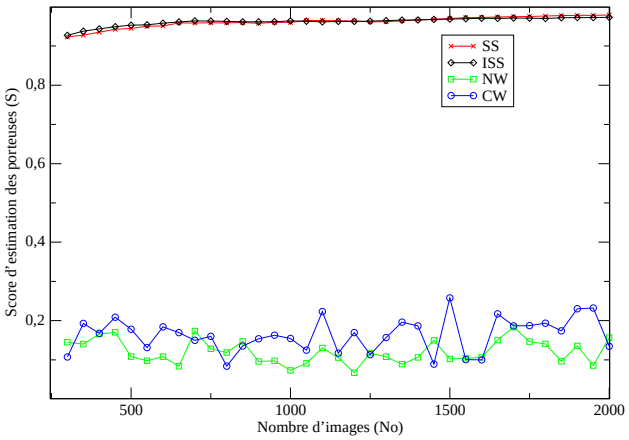

Figure 2 – *Estimation des porteuses pour les quatre modulations. Un score S proche de 1 indique que les porteuses sont esttimées avec une bonne précision.*

6 Conclusions

Nous avons présenté la première étude d'ampleur sur la sécurité des méthodes de tatouage d'image par étalement de spectre dans le cadre d'une attaque WOA. Nos expériences confirment le fait que les modulations classiques SS et ISS ne sont pas sûres. Elles confirment également que les modulations NW et CW présentent une sécurité utilisable pour les applications sensibles. Les faits marquants que nous avons établis dans ce papier sont les suivants :

– le tatouage naturel (NW) est inutilisable lorsque $\eta = 1$, et se révèle donc en pratique inutilisable pour les applications de stéganographie pure ;
– la robustesse des méthodes sûres en WOA (CW et NW) se révèle bien plus élevée en pratique que ne le laissaient présager les simulations numériques sur des vecteurs synthétiques.

Remerciements

Ce travail a été effectué dans le cadre des projets IST-2002-507932 ECRYPT, ANR-06-SETI-009 Nebbiano, RIAM Estivale et ARA TSAR.

Références

[1] François Cayre, Teddy Furon, and Caroline Fontaine. Watermarking security : Theory and practice. *IEEE Trans. Sig. Proc.*, 53(10) :3976–3987, October 2005.

[2] I.J. Cox, J. Killian, F.T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Trans. Im. Proc.*, 6(12) :1673–1687, December 1997.

[3] Henrique S. Malvar and Dinei Flôrencio. Improved spread spectrum : a new modulation technique for robust watermarking. *IEEE Trans. Sig. Proc.*, 53 :898–905, April 2003.

[4] Patrick Bas and François Cayre. Natural watermarking : a secure spread spectrum technique for woa. In *Proc. Information Hiding*, Alexandria, VA, July 2006.

[5] Patrick Bas and François Cayre. Achieving subspace or key security for woa using natural or circular watermarking. In *Proc. ACM Multimedia Security Workshop*, Geneva, September 2006.

[6] François Cayre and Patrick Bas. Kerckhoffs based embedding security classes. *IEEE Trans. Inf. For. Sec.*, 2007.

[7] Auguste Kerckhoffs. La cryptographie militaire. *Journal des Sciences militaires*, IX :5–38, January 1883.

[8] Aapo Hyvarinen. Fast and robust fixed-point algorithm for independent component analysis. *IEEE Trans. Neur. Net.*, 10(3) :626–634, 1999.

[9] Patrick Bas and Gwenäel Doërr. Practical security analysis of dirty paper trellis watermarking. In *Proc. Information Hiding*, St-Malo, June 2007.