ATTAQUE MALVEILLANTE D'IMAGES TATOUÉES BASÉE SUR L'AUTO-SIMILARITÉ[1]

Gabriella CSURKA, Jean-Luc DUGELAY, Caroline MALLAURAN, Jean-Pierre NGUYEN, Christian REY
Institut EURECOM, Département Communication Multimédia
2229 route des Crêtes B.P. 193, Sophia Antipolis, FRANCE
http://www.eurecom.fr/~image
jean-luc.dugelay@eurecom.fr

Résumé

Le tatouage d'images consiste à cacher de manière imperceptible et robuste une information dans une image, de manière à pouvoir extraire cette information, même si l'image a subi une attaque bien ou malveillante. Afin d'évaluer l'efficacité d'un algorithme de tatouage, il est important de tester sa robustesse par rapport à un ensemble de manipulations photométriques et géométriques classiques, compressions, mais également d'attaques malveillantes que l'image tatouée risque de subir. En conséquence, il est important de développer certaines attaques permettant de tester et donc d'améliorer les algorithmes de tatouage. Dans ce sens, l'objectif de ce papier est de proposer un algorithme d'attaque malveillante d'images tatouées en se basant sur la propriété d'auto-similarité des images.

Mots Clef

Tatouage d'images, évaluation, auto-similarité, attaque malveillante

1 Introduction

Le tatouage d'images consiste à cacher un filigrane digital imperceptible contenant un message dans une image de manière à pouvoir extraire ce filigrane (message) même si l'image a subi certaines manipulations bien ou malveillantes [3]. Depuis ces dernières années, beaucoup d'algorithmes de tatouage en images fixes ont été proposés. Certains algorithmes travaillent directement dans le domaine spatial, mais la plupart cachent le filigrane via un domaine transformé (la transformée discrète en cosinus, la transformée discrète de Fourier, les ondelettes ou les fractales).

Afin de pouvoir comparer ces systèmes de tatouage, il est nécessaire de tester leur résistance par rapport à des manipulations photométriques et géométriques classiques, compressions, mais également à des attaques malveillantes effectuées sur un même ensemble d'images de tests représentatives. Parmi de tels logiciels d'évaluation, on peut mentionner le logiciel Stirmark [4], qui propose non seulement une panoplie de manipulations géométriques et photométriques mais aussi l'attaque malveillante Stirmark, consistant en une succession de distorsions géométriques aléatoires appliquées localement à plusieurs endroits dans l'image. Immédiatement, cette attaque a mis en défaut la quasi totalité des tatoueurs. Depuis, certains tatoueurs ont réussi à améliorer leurs performances afin de résister à cette attaque.

Au sein de la communauté «watermarking», il existe depuis le départ, une sorte de compétition entre les «watermarkers» d'une part et les «crackers» d'autre part. Cependant, les recherches des «crackers» sont utiles aux recherches des «watermarkers». En effet, il est important de développer certaines attaques permettant d'évaluer et donc d'améliorer les algorithmes de tatouage. Parmi ces attaques malveillantes, nous pouvons distinguer celles qui perturbent l'image de telle sorte que, même si la marque reste présente dans l'image tatouée, le récupérateur de marque ne sait pas l'extraire sans avoir recours à l'image originale et celles qui «lessivent» la marque dans l'image.

Notre objectif est donc de définir, valider et tester un nouvel algorithme d'attaque malveillante basée sur les auto-similarités incluses dans les images. L'attaque optimale souhaitée ferait en sorte qu'avec une distorsion minimale de l'image et tout en conservant une performance comparable à celle de Stirmark, mais sans ajouter de distorsions géométriques, le récupérateur de marque soit suffisamment perturbé pour ne plus pouvoir extraire la marque correctement. Contrairement à Stirmark, il est ici toujours possible de calculer une erreur 'pixel' à 'pixel' entre les images tatouées obtenues avant et après attaque, et de rapprocher cette erreur avec celle introduite par le marquage (i.e. différence entre image originale et tatouée).

2 Méthode proposée

La principale caractéristique de l'approche proposée est l'exploitation de la notion d'auto-similarité présente dans les images. Les auto-similarités dans une image peuvent être considérées comme un type particulier de redondances. En effet, au lieu de rechercher la corrélation

[1] Ce travail a été, en partie, réalisé dans le cadre du Projet Européen - IST-1999-10987, CERTIMARK - *Certification for watermarking technique* (http://www.certimark.org).

entre les pixels adjacents, on s'intéresse ici à des corrélations entre des parties plus ou moins espacées dans l'image. L'idée des auto-similarités dans les images a été exploitée avec succès pour la compression fractale [2].

Au niveau du codage fractal, deux approches ont été développées : une première approche dans le domaine spatial [2] et une seconde dans le domaine transformé [1]. De ce fait, l'attaque proposée présente plusieurs déclinaisons possibles liées au domaine dans lequel on désire attaquer. Etant donné que certains algorithmes de tatouage travaillent dans le domaine spatial et que d'autres tatouent dans le domaine transformé, il semble intéressant de travailler sur les deux plans.

2.1 L'attaque spatiale

Dans le domaine spatial, l'image initiale est balayée bloc par bloc avec un recouvrement éventuel. Ces blocs sont appelés *Range block* (bloc **R**) de dimension donnée. Chaque bloc $\mathbf{R_i}$ est ensuite mis en correspondance avec un autre bloc transformé $\mathbf{D_j}$ lui « ressemblant » (modulo des ajustements photométriques et géométriques) au sens d'une mesure d'erreur *RMS* (Root Mean Squared) définie par :

$$RMS(f,g) = \frac{1}{n}\sqrt{\sum_{x=1}^{n}\sum_{y=1}^{n}\left[f(x,y)-g(x,y)\right]^2}$$

Le bloc $\mathbf{D_j}$, appelé *Domain block,* est recherché à travers une librairie composée de **Q** blocs appartenant à l'image. Les **Q** blocs ne forment pas nécessairement une partition de l'image. Chaque bloc $\mathbf{Q_i}$ est ramené à l'échelle de manière à être de même taille que $\mathbf{R_i}$ (si leurs tailles ne sont pas les mêmes). Il subit ensuite une transformation géométrique $T_\mathbf{k}$ parmi un ensemble de transformations prédéfinies (identité, 4 réflexions et 3 rotations de k*90°). Pour chaque bloc $\mathbf{Q_i}$ transformé ($T_\mathbf{k}$ ($\mathbf{Q_i}$)), la contraction photométrique (scaling **s**) et le décalage (offset **o**) sont calculés en minimisant l'erreur entre ce bloc $g=T_\mathbf{k}$ ($\mathbf{Q_i}$) et le bloc $f=\mathbf{R_i}$ par la méthode des Moindres Carrés :

$$R = \sum_{x=1}^{n}\sum_{y=1}^{n}\left(\mathbf{s}.g(x,y)+\mathbf{o}-f(x,y)\right)^2$$

Finalement, le bloc $\mathbf{D_i}$ mis en correspondance avec $\mathbf{R_i}$ est le bloc $\mathbf{s}\times T_\mathbf{k}(\mathbf{Q_i})+\mathbf{o}$ pour lequel la distance RMS est minimale.

Puisque le bloc $\mathbf{R_i}$ et le bloc $\mathbf{D_i}$ sont similaires, nous pouvons remplacer $\mathbf{R_i}$ par $\mathbf{D_i}$. Ainsi, le contenu de l'image va peu ou ne pas changer, mais les informations concernant le tatouage seront dispersées dans l'image et donc le décodeur sera incapable de retrouver les informations aux endroits prévus. L'inconvénient de cette approche est que tous les blocs n'ont pas de correspondants qui soient suffisamment similaires pour maintenir une qualité d'image acceptable (voir résultats expérimentaux).

2.2 L'attaque fréquentielle

L'approche via le domaine fréquentiel est inspirée du codage fractal dans le domaine transformé [1]. L'idée de base est de chercher pour la DCT (Transformée Discrète en Cosinus) du bloc $\mathbf{R_i}$ un bloc $\mathbf{D_i}$ transformé DCT. Mais, puisque les coefficients n'ont pas la même importance, le calcul global d'un «scaling **s**» et d'un «offset **o**» par bloc a peu de sens. Nous avons donc essayé d'utiliser plusieurs **s** et **o** en regroupant les coefficients selon les différents niveaux de fréquences. Cependant, nous avons rencontré une autre difficulté qui était de définir une mesure de ressemblance adéquate dans le domaine fréquentiel car une simple *RMS* ne tient pas compte des disparités entre les coefficients DCT. Une solution envisageable est d'introduire une forme de pondération ou bien d'utiliser des mesures plus complexes telle que la mesure Watson [6] qui est une mesure d'erreur agissant directement dans le domaine DCT.

Cependant, nous n'avons pas poursuivi nos investigations dans cette direction car nous avons choisi de développer une approche hybride «*spatio-fréquentielle*».

2.3 L'attaque spatio-fréquentielle

L'idée de base est de rechercher d'abord des blocs similaires dans le domaine spatial comme décrit pour «*l'attaque spatiale*», mais ensuite de transformer par la transformée discrète en cosinus les blocs $\mathbf{R_i}$ et $\mathbf{D_i}$ mis en correspondance dans le domaine direct. Afin de garder une meilleure qualité d'image, le bloc $\mathbf{R_i}$ conservera les *N* premiers coefficients DCT selon un parcours en zigzag (voir Figure 1). Les autres coefficients du bloc DCT($\mathbf{R_i}$) seront substitués par ceux du bloc DCT($\mathbf{D_i}$). Suite au calcul de la transformée discrète inverse en cosinus du bloc obtenu après les modifications des coefficients, ce dernier sera intégré dans l'image de départ pour remplacer le bloc $\mathbf{R_i}$.

1	**2**	**6**	**7**	**15**	**16**	**28**	**29**
3	**5**	**8**	**14**	**17**	**27**	**30**	**43**
4	**9**	**13**	**18**	**26**	**31**	**42**	**44**
10	**12**	**19**	**25**	**32**	**41**	**45**	**54**
11	**20**	**24**	**33**	**40**	**46**	**53**	**55**
21	**23**	**34**	**39**	**47**	**52**	**56**	**61**
22	**35**	**38**	**48**	**51**	**57**	**60**	**62**
36	**37**	**49**	**50**	**58**	**59**	**63**	**64**

Figure 1. *Parcours diagonal en zigzag dans un bloc de taille 8x8.*

Le compromis entre la qualité de l'image et l'efficacité de l'attaque est définie par le choix de *N*. Plus grand est *N* plus la qualité de l'image est préservée et inversement en diminuant *N* l'attaque devient plus efficace mais la qualité d'image diminue.

De plus, les tests menés ont montré qu'un *N* global n'était pas satisfaisant. Pour cette raison, le choix de *N* s'effectue localement en fonction de l'erreur entre $\mathbf{R_i}$ et $\mathbf{D_i}$ d'une part, et du contenu du bloc $\mathbf{R_i}$ d'autre part (zone uniforme, texturée, ou incluant des contours).

Finalement, afin d'éviter les effets blocs, les range blocs sont choisis avec un recouvrement et la substitution est effectuée avec un masque donné (dans notre cas, un cercle inscrit dans le bloc) ; c'est-à-dire que seule une partie du bloc définie par le masque est remplacée.

3 Résultats expérimentaux

Pour effectuer nos tests, nous avons utilisé plusieurs images de tailles différentes, plus ou moins texturées, souvent utilisées pour tester des tatoueurs [7]. Ces images sont présentées dans la Figure 2.

Figure 2. *Les images originales utilisées et leurs tailles* : *Baboon (512×512), Bear (394×600), Skyline_arch (400×594), Lena (512×512), Newyork (842×571)*

Nous avons évalué notre attaque en marquant les images avec comme tatoueur de référence Digimarc, qui reste, à l'heure actuelle, un des tatoueurs le plus utilisé.

Dans un premier temps, nous avons testé l'attaque «*spatiale*», c'est-à-dire l'attaque pour laquelle nous remplaçons chaque bloc $\mathbf{R_i}$ par le bloc $\mathbf{D_i}$. Les Figures 3 et 4 montrent les images Lena et Baboon tatouées et leurs correspondantes marquées et attaquées. Nous pouvons constater les dégradations sur les images attaquées. Par contre si nous appliquons l'attaque «*spatio-fréquentielle*», la qualité des images est préservée comme nous pouvons constater sur la Figure 5. La Figure 6. montre d'autres images marquées et attaquées avec l'attaque «*spatio-fréquentielle*». Dans les trois cas, comme pour Lena et Baboon, le marqueur Digimarc n'a retrouvé aucun filigrane après notre attaque.

Figure 3. *L'image Lena tatouée et l'image marquée, puis attaquée. Le PSNR entre les deux images est de 25.5dB.*

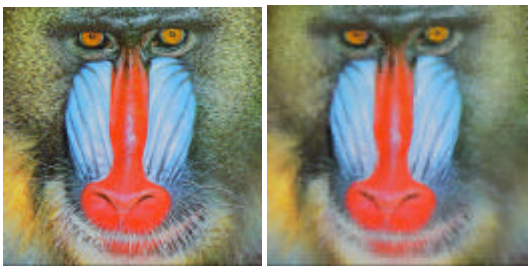

Figure 4. *L'image Lena tatouée et l'image marquée, puis attaquée. Le PSNR entre les deux images est de 19.25dB.*

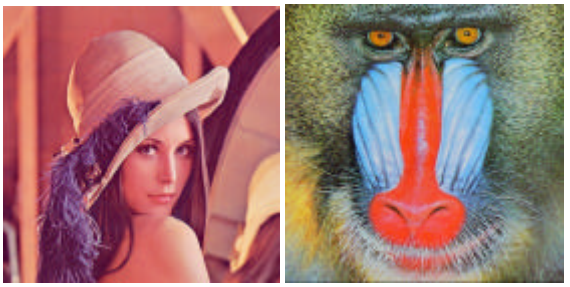

Figure 5. *Les images Lena et Baboon marquée, puis attaquées. Les PSNR entre les images tatouées et celles attaquées sont respectivement de 34.54dB et de 24.51dB.*

3.1 Analyse des résultats

Il est important de noter que par souci de ne pas perdre l'information par une simple compression JPEG, les tatoueurs récents insèrent le plus souvent les informations concernant le tatouage dans les fréquences moyennes. Il est donc important pour qu'une attaque soit efficace que les *N* coefficients qui ne seront pas remplacés soient entièrement dans les basses fréquences.

Figure 6. *Les images Skyline_arch, Newyork et Bear marquées et attaquées. Les PSNR entre les images tatouées et celles attaquées sont respectivement de 34.28dB, de 24.9dB et de 33.58dB.*

Mais comme nous l'avons dit précédemment, si *N* est trop petit, on diminue forcement la qualité d'image. Notre but était d'arriver à avoir une attaque efficace avec une distorsion équivalente à celle provoquée par le tatouage (≈38-40dB). Mais atteindre ce but n'est pas évident car les tatoueurs sont de plus en plus performants (grâce aussi à des attaques qui ont montré les faiblesses des anciens tatoueurs). En effet, les filigranes étant dépendants de l'image, il est difficile de les «effacer» ou même les «perturber» sans affecter les informations concernant l'image.

Finalement, il faut noter que les valeurs numériques (i.e. PSNR) mentionnées pour donner une indication sur la qualité des images ne sont pas très significatives. En effet, il est bien connu que le PSNR comme mesure de qualité n'est pas bien adapté (les images dans la Figure 7 en sont des bons exemples) et des mesures plus proches du système visuel humain (SVH) sont nécessaires pour mieux évaluer la distorsion introduite par l'attaque. Même si le PSNR est encore largement utilisé, des nouvelles mesures basées sur le SVH ont été proposées parmi lesquelles nous pouvons mentionner celle de Watson [6] ou Saadane et. al. [5].

4 Conclusion

Dans ce papier, nous avons présenté une attaque malveillante basée sur les auto-similarités dans les images. Une première déclinaison de cette attaque opère dans le domaine spatial, et une seconde dans le domaine fréquentiel (DCT). Cependant, afin d'avoir une attaque simple, efficace tout en préservant au mieux la qualité des images, nous avons proposé une attaque « *spatio-fréquentielle* » où la recherche des bloc similaires s'effectue dans le domaine spatial, mais la desynchronisation dans le domaine fréquentiel. L'attaque a été testée avec succès sur le tatoueur Digimarc.

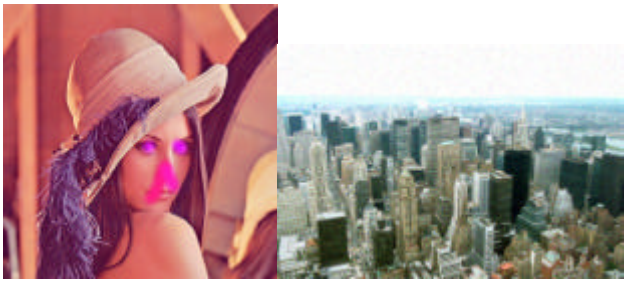

Figure 7. *L'image Lena marquée sur laquelle on a ajouté des tâches visibles et gênantes et l'image Newyork marquée sur laquelle nous avons ajouté des bruits gaussiens visibles. Les PSNR entre les images marquées et ces images manipulés sont plus grand (35.32dB pour Lena et 25.3dB pour Newyork) que dans le cas de notre attaque (34.54dB et 24.9dB) malgré le fait qu'il soit clair que visuellement nos images attaquées sont de qualités supérieures.*

Références

[1] Barthel (K-U), Schüttemeyer (J.), Noll (P.), « A new image coding technique unifying fractal and transform coding », IEE on Image Processing, Austin Texas, 13-16 November 1994.

[2] Fisher (Y.), « Fractal Image Compression – Theory and Application », Springer-Verlag, New-York, 1994.

[3] Katzenbeisser (S.), Petitcolas (F. A.P.), « Information Hiding – Techniques for Steganography and Digital Watermarking », Artech House, Boston-London, 2000.

[4] Kuhn (M. G.), Petitcolas (F. A.P.), Stirmark, 1997 : http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/

[5] A. Saadane, N. Bekkat, D. Barda, « On the masking effects in a perceptually based image quality metric », Advances in the theory of computation and computational matematics book series, Vol. Imaging and Vision Systems, 2001.

[6] A. B. Watson. DCT quantization matrices visually optimized for individual images. Proceedings of SPIE : Human vision, Visual Processing and Digital Display IV, Vol. 1913, pp 202-216, 1993.

[7] Base d'image : http://www.cl.cam.ac.uk/~fapp2/watermarking/benchmark/image_database.html