Quantification adaptative pour la stéganalyse d'images texturées

Patrick BAS[1], Pierre CHAINAIS[1,2], Emmanuel ZIDEL-CAUFFET[1]

[1] LAGIS, UMR CNRS 8219, Ecole Centrale de Lille, 59651 Villeneuve d'Ascq Cedex

[2]INRIA-Lille Nord Europe, SequeL, 59651 Villeneuve d'Ascq

Patrick.Bas, Pierre.Chainais@ec-lille.fr, emmanuel.zidel@centraliens-lille.org

Résumé – Nous cherchons à améliorer les performances d'un schéma de stéganalyse (i.e. la détection de messages cachés) pour des images texturées. Le schéma de stéganographie étudié consiste à modifier certains pixels de l'image par une perturbation +/-1, et le schéma de stéganalyse utilise les caractéristiques construites à partir de la probabilité conditionnelle empirique de différences de 4 pixels voisins. Dans sa version originale, la stéganalyse n'est pas très efficace sur des images texturées et ce travail vise à explorer plusieurs techniques de quantification en utilisant d'abord un pas de quantification plus important puis une quantification adaptative scalaire ou vectorielle. Les cellules de la quantification adaptative sont générées en utilisant un K-means ou un K-means "équilibré" de manière à ce chaque cellule quantifie approximativement le même nombre d'échantillon. Nous obtenons un gain maximal de classification de 3.7% pour un pas de quantification uniforme de 3 sur $[-9, 9]$. En utilisant l'algorithme K-means équilibré sur le même intervalle, le gain par rapport à la version de base est de 9.3%.

Abstract – The goal of this work is to improve the performance of a steganalysis scheme for textured images. The steganographic scheme consists in modifying several pixel values of the image by +/- 1 and the steganalysis scheme uses co-occurence matrices of differences between neighbouring pixels. The original steganalysis scheme is not efficient for textured images in particular because the quantization is not optimal. We propose to improve the quantization process by either increasing the quantization step or using an adaptive scalar or vector quantization. The quantization cells are generated using K-means or a balanced version of K-means which fills each cell equally. We obtain a gain of 3.7% with a quantization step $\times 3$ on $[-9, 9]$ and the gain with respect to the baseline scheme is 9.3% with a balanced K-means on $[-9, 9]$.

1 Introduction

Nous cherchons dans ce travail à améliorer les performances d'un schéma de stéganalyse (i.e. la détection de la présence de messages cachés) pour des images texturées dont les variations importantes rendent la stéganalyse plus difficile que pour des images régulières. Par conséquent, nous limitons dans un premier temps cette étude à un schéma de stéganographie très simple appelée "LSB Matching" ou "LSB+/-1" qui consiste à modifier certains pixels de l'image par une perturbation ± 1 ; on ne modifie ainsi que les bits de poids faible, *Least Significant Bits*. Le dernier étage d'une méthode de stéganalyse est un classifieur *cover/stego* (*cover* désignant une image non modifiée, *stego* une image portant un message caché). La distinction entre images *cover* et *stego* étant subtile, en particulier lorsqu'il s'agit d'images texturées, une stéganalyse efficace s'appuie généralement sur un grand nombre de caractéristiques dont on estime des histogrammes joints ou matrices de cooccurence. Nous travaillons ici avec les caractéristiques SPAM [4] construites à partir de la probabilité conditionnelle empirique de différences entre pixels voisins. Les hypothèses sous-jacentes à cette approche sont que les images naturelles sont généralement localement régulières, la modification ponctuelle de certains pixels se traduisant alors par une fréquence "anormale" des différences d'intensité entre pixels voisins. Nous montrons que, dans sa version originale, ce schéma de stéganalyse n'est pas très efficace sur des images texturées. Notons que sur des images classiques telles que celles de BossBase [1], ce schéma obtient des performances de 100% avec un simple classifieur linéaire. Ce résultat est attendu car la distribution des gradients est beaucoup moins étalée pour les images régulières que pour des images texturées, i.e. plus désordonnées.

Tous les ingrédients d'une technique de stéganalyse sont potentiellement importants, depuis le choix des caractéristiques utilisées, ici les SPAM, jusqu'au classifieur *cover/stego*, par exemple une analyse linéaire discriminante (LDA). Nous choisissons de travailler avec un nombre de dimensions identique à celui de l'approche SPAM, sans changer ni le type de caractéristiques utilisées, ni le classifieur, typiquement une LDA. Le but de ce travail est d'identifier l'importance d'une étape de quantification des caractéristiques permettant d'estimer des matrices de coocurrences (histogrammes joints) sur un domaine plus large de valeurs des caractéristiques sans changer la taille des histogrammes. Notre souci est ici comme dans [3] de ne pas augmenter la dimension de l'espace de décision (= taille des histogrammes). Ce dernier point est crucial puisqu'il s'agit du premier niveau permettant une réduction de dimension. Rappelons qu'un histogramme joint de K caractéristiques quantifiées aur 2T+1 valeurs est un vecteur de taille $(2T+1)^K$.

Notre objectif est d'améliorer la quantification des différences de pixels initialement quantifiées avec un pas unitaire sur l'intervalle $[-3, 3]^3$. Nous étudions dans un premier temps

l'impact du choix d'un pas multiple du pas initial. Nous utilisons ensuite une autre stratégie consistant à effectuer directement une quantification vectorielle dans l'espace des données en utilisant une variante de l'algorithme k-means adaptée aux données discrètes. Cette quantification permet d'obtenir une quantification adaptée au modèle conjoint sous-jacent. Enfin, nous testons une quantification vectorielle équilibrée [6] assurant un compromis entre regroupement des points et taille des clusters de sorte que tous les clusters sont de taille similaire.

La partie 2 rappelle la définition des caractéristiques SPAM ainsi que la méthode de construction des matrices de cooccurence associées. Cette étude a nécessité la préparation d'une base de 3950 images texturées 512×512 appelée TEXTBASE et que nous décrivons dans la partie 3. La partie 4 détaille chacune des méthodes de quantification proposées. Nous présentons les résultats obtenus dans la partie 5 avant de conclure.

2 Les caractéristiques SPAM

Nous étudions la stéganalyse du schéma d'insertion stéganographique sur les bits de poids faible appelé "LSB Matching" ou "LSB+/-1". La stéganalyse de ce schéma s'est améliorée grâce à l'utilisation des caractéristiques SPAM [4]. Ces caractéristiques sont construites à partir d'une probabilité 3D conjointe $C(d_1, d_2, d_3)$, le triplet (d_1, d_2, d_3) représentant des différences de 3 couples de pixels adjacents calculées dans un sens de parcours choisi parmi les 8 possibles ($\leftarrow, \rightarrow, \downarrow, \uparrow, \nwarrow, \nearrow, \swarrow, \searrow$). Afin d'augmenter la représentativité des différents motifs possibles à travers les images, nous utilisons la probabilité conditionnelle $C(d_1|d_2, d_3)$, définie par exemple pour le sens de parcours $\rightarrow$:

$$C_{\rightarrow}(d_1|d_2, d_3) = \Pr[(p_{i,j} - p_{i,j+1}) = d_1 \quad (1)$$
$$|p_{i,j+1} - p_{i,j+2} = d_2, p_{i,j+2} - p_{i,j+3} = d_3)]$$

où Pr est la probabilité empirique, et $p_{i,j}$ représente l'intensité du pixel de coordonnées (i, j). Notons que les $p_{i,j}$ étant des entiers, par exemple compris entre 0 et 255, les d_k sont aussi des entiers, d'où le caractère discret des données par la suite.

Afin de limiter le nombre de caractéristiques, la probabilité conditionnelle n'est habituellement calculée que pour les valeurs de différences $d_k \in \{-3, \ldots, 3\}$ ce qui permet de générer $7^3 = 343$ caractéristiques pour chaque direction. Toujours dans le but limiter le nombre de caractéristiques, les matrices représentant les directions $\{\leftarrow, \rightarrow, \downarrow, \uparrow\}$ sont moyennées, il en va de même pour les directions $\{\nwarrow, \nearrow, \swarrow, \searrow\}$. Les caractéristiques SPAM totalisent ainsi $2 \times 343 = 686$ caractéristiques. Cette méthode de stéganalyse a servi de clef de voûte à la construction de modèles riches contenant plusieurs dizaines de milliers de caractéristiques [2] et qui ont remporté le challenge international BOSS [1]. Les modèles riches nécessitent cependant l'utilisation d'un ensemble de classifieurs et sont lourds à utiliser lors de l'extraction des caractéristiques et de l'apprentissage. Ces caractéristiques en grand nombre sont également difficilement interprétables.

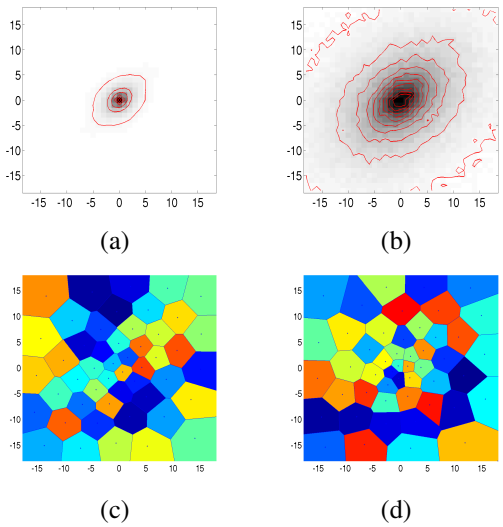

FIGURE 1 – Probabilités conjointes des caractéristiques SPAM (d_2, d_3), pour les directions $\{\leftarrow, \rightarrow, \downarrow, \uparrow\}$ (a) pour BossBase (b) pour TEXTBASE ; cellules de Voronoï de quantification obtenues par (c) k-means, (d) balanced k-means pour TEXTBASE .

L'objectif de ce travail est d'obtenir des caractéristiques plus discriminantes au sens de la stéganalyse tout en gardant leur nombre constant et égal à 686. Nous nous attachons en particulier à étudier la stéganalyse d'images texturées, une catégorie d'images difficiles à analyser. A titre d'exemple, pour un taux d'insertion de 1 b.p.p., les caractéristiques SPAM offrent un taux de bonne classification de 100% sur la base d'images BossBase [1] essentiellement composée d'images de paysages, alors que celui-ci n'atteint que 64.8% sur une base d'image texturées TEXTBASE constituée de 3950 photos de textures naturelles, voir partie 3. Nous expliquons cette baisse importante de performance par le fait que les caractéristiques SPAM, si elles modélisent l'essentiel des différences d'une image régulière sur l'intervalle $[-3, 3]^3$, ne rend compte que d'une faible proportion des variations possibles dans une image texturée. Les figures 1(a) et 1(b), qui représentent respectivement les probabilités conjointes de couples de points tirés aléatoirement dans des images des bases BossBase et TEXTBASE , permettent d'apprécier les différences entre les dispersions statistiques de ces deux bases d'images. Afin de conserver un nombre fixe de caractéristiques, l'une des solutions pour extraire des caractéristiques pertinentes à partir de la probabilité conditionnelle est de quantifier le triplet (d_1, d_2, d_3) de manière appropriée.

3 Construction de TEXTBASE

La notion de *texture* n'est pas bien définie. Nous assimilons ici une texture à une image présentant un caractère suffisamment désordonné et contenant de l'énergie sur une gamme d'échelles étendue. Nous avons construit TEXTBASE , une base de 3950 images texturées 512×512 en niveaux de gris. Cette

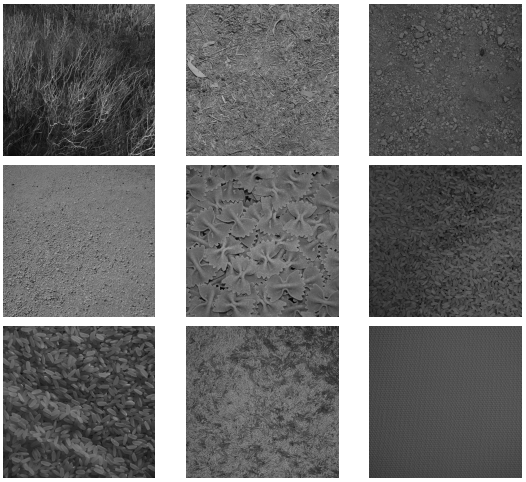

FIGURE 2 – Exemples variés d'images issues de TEXTBASE .

base est spécifiquement destinée à tester des outils de stéganalyse et comprend des versions stéganographiées avec les méthodes LSB ± 1 et HUGO pour une gamme de taux d'insertion. TEXTBASE sera prochainement disponible.

Pour éviter le biais dû aux habitudes du photographe, les clichés ont été pris avec des distances focales variées, la luminosité n'est généralement pas la même d'un groupe de photos à un autre. Les photos ont été prises dans des pays différents, durant des périodes différentes. La distance aux objets et les positions dans lesquelles l'appareil a été tenu varient. Le seul élément parfaitement commun entre ces images est l'appareil qui a été utilisé pour les réaliser et la personne ayant pris les clichés. TEXTBASE a été élaborée en visant les critères suivants : obtenir plusieurs catégories de textures, chaque image devant être une texture quasiment pure (pas de mélange de textures) sans éléments étranger. TEXTBASE se compose de textures provenant de plusieurs scènes (mur, crépi, sols divers, tissus, végétation, aliments...)

Pour mesurer qualitativement le niveau de désordre de ces textures par un critère qualitatif lié au problème de la stéganalyse, nous avons utilisé l'homogénéité des cartes de probabilité d'insertion d'un message caché utilisée par l'algorithme de stéganalyse HUGO [5]. Pour disposer d'un critère quantitatif et plus objectif, nous avons choisi de découper les images de taille $M \times N$ en $K \times L$ blocs $\mathcal{B}_{k,\ell}$, $1 \leq k \leq K$, $1 \leq \ell \leq L$ de taille $M/K \times N/L$. Nous calculons les entropies locales $S_{k,\ell}$ de chaque bloc dont la quantité $\mathcal{H}_{\|,\ell} = |S_{k,\ell} - \frac{\log(MN)}{KL}|$ mesure l'écart à l'entropie maximale. La quantité $\mathcal{D} = \sum_{k,\ell} \mathcal{H}_{\|,\ell}$ mesure alors l'écart moyen des entropies locales à l'entropie maximale qui doit être inférieur à un seuil déterminer empiriquement (e.g. $0,6529$ ici) pour qu'une image soit considérée comme une texture homogène. Après sélection, nous avons conservé 3950 images de taille 512×512 représentant bien des textures désordonnées et homogènes sur l'ensemble de l'image, voir figure 2.

4 Quantification optimisée

Trois stratégies de quantification ont étés étudiées. Les SPAM standard consistent à ne considérer que les valeurs des $d_i \in [-3, 3]$, soit les $2T + 1 = 7$ valeurs voisines de zéro (T=3). L'extension la plus naturelle pour tenir compte de la plus grande dispersion des gradients dans les textures consiste à choisir une *quantification scalaire uniforme 1D* (QS uni.) sur chacune des variables d_i, $i \in \{1, 2, 3\}$, le pas de quantification q choisi étant un entier : cette stratégie permet d'explorer la plage $[-3 * q, 3 * q]$ mais ne tient pas compte de la forme de la distribution des d_i.

Une *quantification vectorielle* (QV) de Lloyd-Max tient compte de la distribution sous-jacente en minimisant l'erreur quadratique moyenne de quantification. Afin de prendre en compte les statistiques jointes, nous avons effectué séparément une quantification vectorielle 2D sur le couple de variables (d_2, d_3) et une quantification adaptative sur d_1 décrite dans le paragraphe suivant. L'usage de la probabilité conditionnelle $C(d_1|d_2, d_3)$ impose de quantifier indépendamment d_1 et (d_2, d_3). La quantification vectorielle est obtenue en adaptant l'algorithme *K-means* pour pouvoir travailler sur des points à coordonnées entières ($d_i \in \mathbb{N}$) et non pas continues. Afin d'empêcher que les centres des clusters restent "piégés" dans des minima locaux, les données sont préalablement légèrement bruitées par un bruit uniforme dans $[-0.5; 0.5]^2$. Le K-means est entrainé à partir de 2.10^6 couples de différences (d_2, d_3) extraits aléatoirement de l'ensemble de la base pour estimer $7^2 = 49$ centres. La figure 1(c) représente leur diagramme de Voronoï. Enfin la différence d_1 est quantifiée en utilisant la stratégie décrite dans le point suivant.

Nous avons aussi exploré une *quantification scalaire adaptative indépendante 1D* sur chaque coordonnée. Dans ce cas, la largeur de chaque cellule de quantification est choisie de manière à ce que le nombre moyen de coefficients par cellule soit constant. Pour une quantification sur $2T + 1 = 7$ centres sur TEXTBASE , on obtient $\{-7.5; -4; -1.5; 0; 1.5; 1.5; 7.5\}$ dans l'intervalle $[-9, 9]$.

Enfin, nous avons utilisé une méthode de *quantification vectorielle équilibrée* dénommée *balanced K-means* [6]. Cette méthode s'appuie sur une heuristique qui modifie l'algorithme de K-means pour favoriser des clusters de taille égale (nombre d'échantillons par cluster approximativement identique). En résumé, tandis que les échantillons appartiennent à un espace de dimension D, les centres sont définis dans un espace de dimension $D + 1$. La $D + 1$-ème coordonnée joue le rôle d'une pénalité : un cluster rassemblant trop d'échantillons voit sa pénalité (son altitude) augmenté, tandis qu'un cluster trop petit voit son centre moins pénalisé. Les cellules de Voronoï sont définie à partir des intersections entre des hyperplans en dimension $D + 1$ et l'espace de dimension D dans lequel vivent les points. La figure 1(d) représente un diagramme de Voronoï obtenu sur TEXTBASE pour (d_2, d_3). Notons que les cellules centrales sont alors plus concentrées que pour le K-means classique.

Stratégie	Performance en classification
QS uni., $q = 1$	65,5%
QS uni., q = 2 / *3* / 4 / 5	67,1% / *68, 5%* / 68.0% / 67,6%
QV 2D sur $[-18, 18]$	69,5%
QV 2D bal. sur $[-18, 18]$	69,9%
QS adapt. $[-18, 18]$	70,2%
QV 2D bal. sur $[-9, 9]$	74,8%

TABLE 1 – Taux de bonne de classification obtenus sur TEXTBASE pour une insertion +/-1 LSB de 1 b.p.p.

5 Résultats et perspectives

Les résultats ont été obtenus à partir de TEXTBASE (3950 images 512×512 en niveaux de gris sur 8 bits) par validation croisée sur 1000 paires (base test, base d'apprentissage) sélectionnées aléatoirement. Les performances de classification sont estimées par cross-validation. Le classifieur de référence le plus simple est l'analyse discriminante linéaire qui a déjà montré son efficacité en stéganalyse. Malgré sa simplicité, une comparaison avec d'autres classifieurs montre que la LDA est relativement efficace sur ce problème. Le tableau 1 permet de comparer les conséquences des différentes stratégies de quantification en termes de performances de stéganalyse. Nous travaillons toujours avec une insertion +/-1 LSB de 1 bit/pixel, le problème devenant rapidement très difficile pour des taux d'insertion trop faibles. Remarquons que nous avons observé de meilleurs résultats en utilisant les probabilités conditionnées $P(d_1|d_2, d_3)$ qu'avec les histogrammes joints $P(d_1, d_2, d_3)$. Nous supposons que cette observation s'explique par le fait que $P(d_1|d_2, d_3) = P(d_1, d_2, d_3)/P(d_2, d_3)$ de sorte que les probabilités conditionnelles sont des quantités renormalisées prenant des valeurs du même ordre de grandeur pour tout (d_2, d_3).

Les résultats issus de ces comparaisons amènent aux conclusions suivantes par ordre croissant de performances, voir tableau 1. L'utilisation d'une quantification uniforme 1D de pas $q \geq 1$ améliore les performances, le taux de bonne classification maximal de 68,5% étant atteint pour un pas $q = 3$ sur l'intervalle $[-9, 9]$. La quantification vectorielle 2D via K-means sur (d_2, d_3) permet d'obtenir un gain de performance par rapport à la quantification uniforme en passant à 69,5% lorsque d_1 est quantifiée de façon adaptative (histogramme approximativement uniforme pour les valeurs quantifiées). Il apparaît qu'une quantification adaptative 1D appliquée indépendemment aux d_i donne un taux de bonne classification de 70.2% comparable mais légèrement supérieur à celui obtenu via une quantification vectorielle 2D. Enfin, notons qu'une quantification vectorielle équilibrée (K-means balanced) fournit des résultats comparables (69,9 %) sur $[-18, 18]$ et des résultats nettement supérieurs (74.8 %) sur $[-9, 9]$.

Il apparaît également que le classifieur ensemblé apporte un gain de performance non négligeable (voir Table 2) en travaillant avec des SPAM standards (alors d'autres résultats montrent que ce gain n'est pas observable sur la base BOSSBase).

	Classifieur	
Quantification	LDA	Ensemble
QS uni., $q = 1$	65,5%	74,1%
QV 2D bal. sur $[-9, 9]$	74.8 %	74,2 %

TABLE 2 – Gain obtenu grâce à la quantification comparé au gain dû au classifieur.

Cette différence de comportement nous amène à penser que les caractéristiques extraites des images texturées ont des distributions plus "compliquées" que celles extraites sur des images naturelles classiques. Notons également que le gain apporté par le classifieur ensemblé disparaît lorsque l'on utilise des caractéristiques issues de la quantification QV équilibrée sur $[-9, 9]$ (nous avons utilisé le code fourni par J. Fridrich avec ses paramètres par défaut pour l'essentiel).

6 Conclusion & Perspectives

Ce travail exploratoire sur l'optimisation de la quantification des caractéristiques d'images texturées pour la stéganalyse a permis de montrer l'importance de cette étape initiale. Des gains de performance importants sont observés lorsque l'espace des caractéristiques est décrit de façon homogène via une quantification assurant que le poids statistique des cellules de quantification est approximativement identique. Puisque ce sont finalement des matrices de co-occurence qui sont utilisées comme entrées du classifieur final, nous explorons maintenant la possibilité d'optimiser la quantification de façon supervisée : les matrices de co-occurence doivent au mieux préserver l'information discriminante.

Références

[1] P. Bas, T. Filler, T. Pevny. "Break Our Steganographic System" : The Ins and Outs of Organizing BOSS. In *Information Hiding*, vol. 6958/2011 of *LNCS*, pp. 59–70, 2011.

[2] J. Fridrich and J. Kodovský. Rich models for steganalysis of digital images. *IEEE Trans. Information Forensics and Security*, 7(3) :868–882, 2011.

[3] T. Pevný. Co-occurrence steganalysis in high dimensions. In *Proc. SPIE 8303, Media Watermarking, Security, and Forensics*, vol. 8303, pp. 83030B, 2012.

[4] T. Pevný, P. Bas, and J. Fridrich. Steganalysis by subtractive pixel adjacency matrix. *IEEE Trans. Info. Forensics and Security*, 5(2) :215–224, 2010.

[5] T. Pevný, T. Filler, and P. Bas. Using high-dimensional image models to perform highly undetectable steganography. In *Proc. of 12th ICIH*, vol. 6387 of *LNCS*, pp. 161–177. Springer-Verlag, 2010.

[6] R. Tavenard, H. Jegou, and L. Amsaleg. Balancing clusters to reduce response time variability in large scale image search. In *Proc. of Int. Work. on CBMI*, pp. 19–24, 2011.