Stéganographie et stéganalyse: tendances actuelles et perspectives.

Jean-François Couchot[1]

[1]Institut FEMTO-ST - Département DISC - Equipe AND.
Univ. Bourgogne Franche-Comté (UBFC), France

24 mars 2017 / Besançon
Journée recherche et développement

Outline

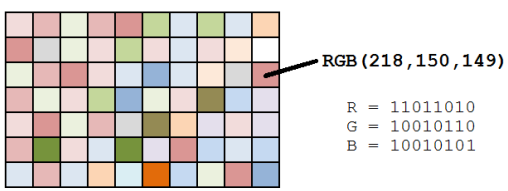
Plan

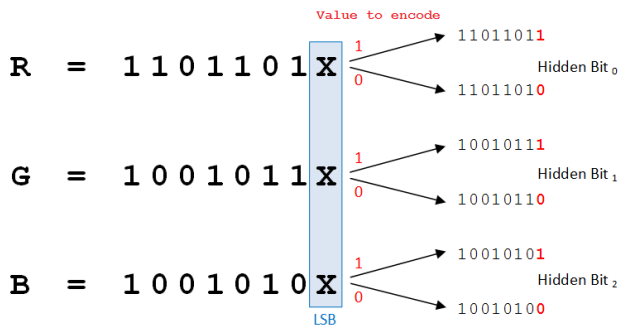
Stéganographie dans des images numériques spatiales

Intuition

L'art de dissimuler un message (souvent crypté) dans une image anodine et de manière imperceptible.

Un exemple dans le domaine spatial (Wikipédia)





Cas d'utilisation de la stéganographie

Message imperceptible (avant tout)

- 2010–'Illegals' : réseau d'espions russes envoyant des données dans des images stéganographiées
- 2011–Duqu (virus Windows) : clefs envoyées dans des images JPEG 54x54 pixels stéganographiées
- ...

Tatouage : difficile d'enlever le message (avant tout)

- DRM dans les images de photographes

Stéganalyse : dual de la stéganographie

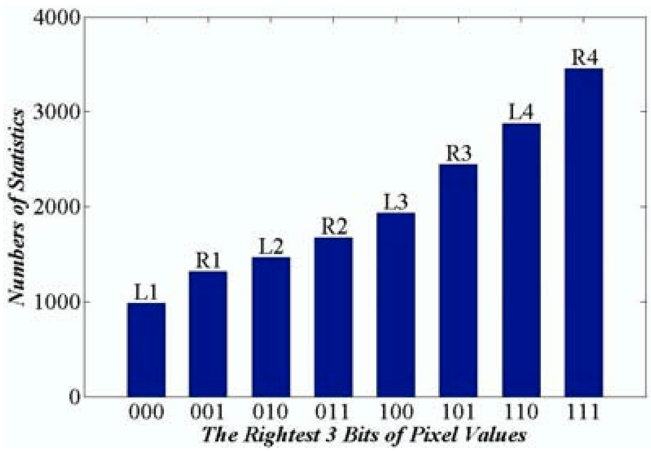
Intuition

Méthodes permettant de décider si une image contient un message.

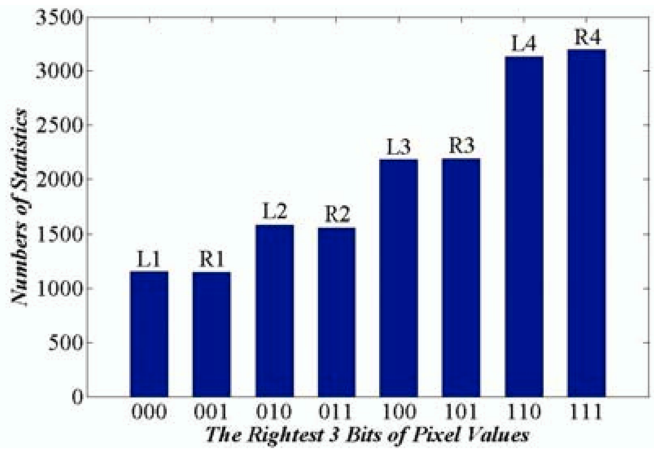
Une démarche de classification

1. Construction des caractéristiques statistiques des images
2. Comparaison avec celles de supports vierges connus
3. Verdict

Exemple : comparaison des histogrammes LSB



(a) Original image.



(b) After LSB substitution.

Le jeu : stéganographie contre stéganalyse

Objectifs contradictoires

- Un stéganographieur :
  - un choix intelligent des pixels à modifier ;
  - un algorithme de modification des pixels efficace.
- Un stéganalyseur :
  - un ensemble de caractéristiques statistiques discriminantes (d'images) ;
  - un algorithme de classification efficace.

Évaluation de la sécurité d'un stéganographieur

- Principe de Kerckhoffs : le stéganalyseur connaît
  - le stéganographieur
  - la taille des messages embarqués dans les images
- Stéganalyse difficile $\Leftrightarrow$ stéganographieur sécurisé

Plan

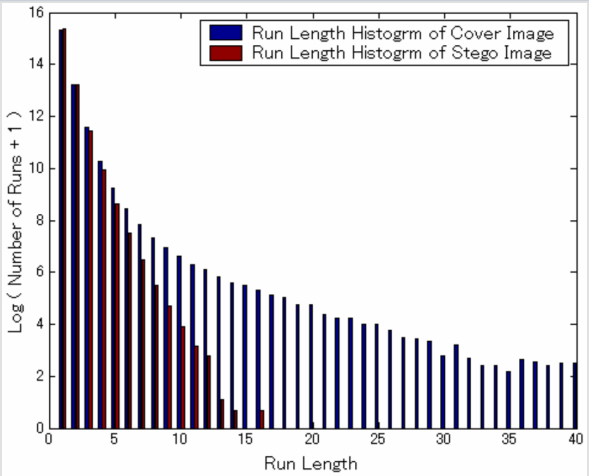
LSB : remplacement et correspondance

Remplacement des LSBs

- Remplacement des LSBs de l'image par les bits du message
- Très peu sécurisée : histogramme caractéristique

Correspondance de LSBs [1]

- Pixel : inchangé si LSB égal au bit courant du message
- Pixel : ± 1 aléatoirement
- Attaquable par histogramme de longueurs de valeurs égales [2]



1. Sharp, T. (2001, April). An implementation of key-based digital signal steganography. In International Workshop on Information Hiding (pp. 13-26). Springer Berlin Heidelberg.

2. Yu, X.Y. and N. Babaguchi, 2008. Run length based steganalysis for LSB matching steganography. Proceedings of the IEEE International Conference on Multimedia and Expo, June 23-April 26, Hannover, Germany, pp : 353-356.

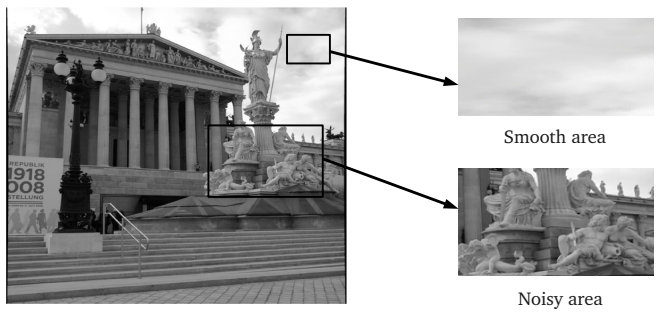
Stéganographie adaptive [4]

Problématique

A partir d'un support X à n éléments : trouver le moyen de transmettre le message m dans l'image stéganographiée Y en minimisant la distorsion.

Étape 1 : carte de distorsion

- Modéliser le *coût de modification* par une fonction de distorsion D additive : $D(X, Y) = \sum_{i=1}^{n} \rho(i)|X_i - Y_i|$
- Trouver l'image Y qui minimise D : $Y = \arg\min D(X, Y)$.
- Intuitivement : valeur de la distorsion
  - zone facilement modélisable, uniforme $\rightarrow$ élevée ;
  - zone de texture ou chaotique $\rightarrow$ faible.



Étape 2 : modification efficaces des bits par STC [3]

- Généralisation des codes de Hamming à n'importe quelle taille.

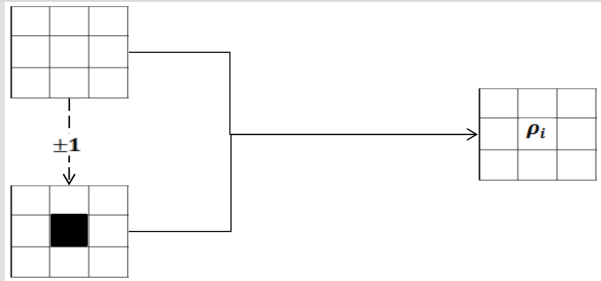
3. Filler, T., Judas, J., & Fridrich, J. (2011). Minimizing additive distortion in steganography using syndrome-trellis codes. IEEE Transactions on Information Forensics and Security, 6(3), 920-935.

4. Fridrich, J., & Filler, T. (2007, February). Practical methods for minimizing embedding impact in steganography. In Electronic Imaging 2007 (pp. 650502-650502). International Society for Optics and Photonics.

Exemples de stéganographie adaptive

Basée sur les modifications de caractéristiques : HUGO [5]

- Distorsion $\rho_i = |f(y_i) - f(x_i)|$ où
- f : valeur des caractéristiques SPAM autour de i



Basée sur les filtres de convolution

- S-UNIWARD [6] : $\rho_U(X) = \sum_{i=1}^{3} \frac{1}{|X \star K^i| + \sigma} \star |K^i|^{\curvearrowleft}$, où K est un filtre d'ondelettes de Daubechies-8
- HILL [7] : $\rho_H(X) = \frac{1}{|X \star H_1| \star L_1} \star L_2$, où $H_1, L_1, L_2 \ldots$

5. Pevný, T., Filler, T., & Bas, P. (2010, June). Using high-dimensional image models to perform highly undetectable steganography. In International Workshop on Information Hiding (pp. 161-177). Springer Berlin Heidelberg.

6. V. Holub, J. Fridrich, T. Denemark, *Universal Distortion Function for Steganography in an Arbitrary Domain*, EURASIP J. on Inf. Se., 2014(1)

7. B. Li, M. Wang, J. Huang, X. Li, *A New Cost Function for Spatial Image Steganography*, 2014 IEEE International Conference on Image Processing (ICIP). pp. 4206-4210, 2014.

Exemples de stéganographie adaptive (suite)

Basée sur un a priori : STABYLO [8]

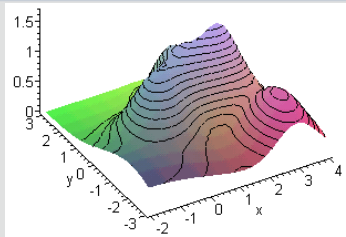
- Bords dans une image : concentrent les fortes variations
- Modifications des pixels de bords : a priori peu détectables

Basée sur des distributions probabilistes : MiPOD [9]

Estimation des paramètres de la distribution statistique d'une fenêtre de pixels avant/après.

Basée sur l'analyse vectorielle : Ky [10]

- Trouver les pixels sur des lignes de niveau les plus irrégulières grâce à des dérivées secondes.



8. Couchot, J. F., Couturier, R., & Guyeux, C. (2015). STABYLO : steganography with adaptive, bbs, and binary embedding at low cost. annals of telecommunications-annales des télécommunications, 70(9-10), 441-449.

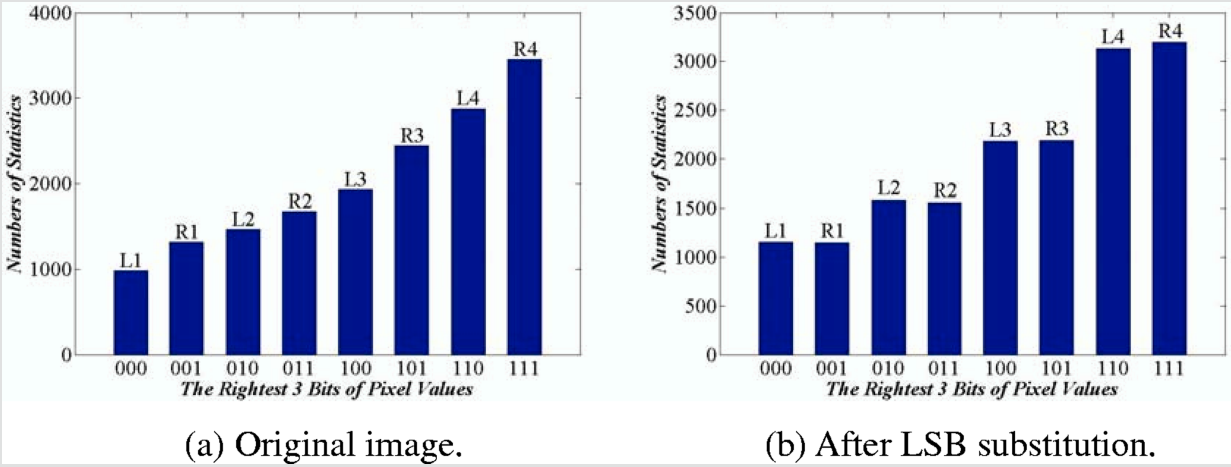
9. V. Sedighi, R. Cogranne and J. Fridrich, *Content-Adaptive Steganography by Minimizing Statistical Detectability*. IEEE Transactions on Information Forensics and Security. 11(2) : 221-234 (2016)

10. Couchot, J. F., Couturier, R., Fadil, Y. A., & Guyeux, C. (2016). A Second Order Derivatives based Approach for Steganography. SECRYPT 2016 : 424-431.

Plan

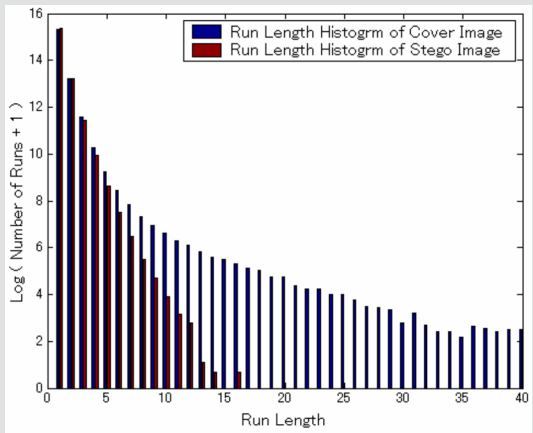
Des histogrammes discriminants

Remplacement des LSBs



(a) Original image. (b) After LSB substitution.

Correspondance de LSBs



Une approche en deux étapes

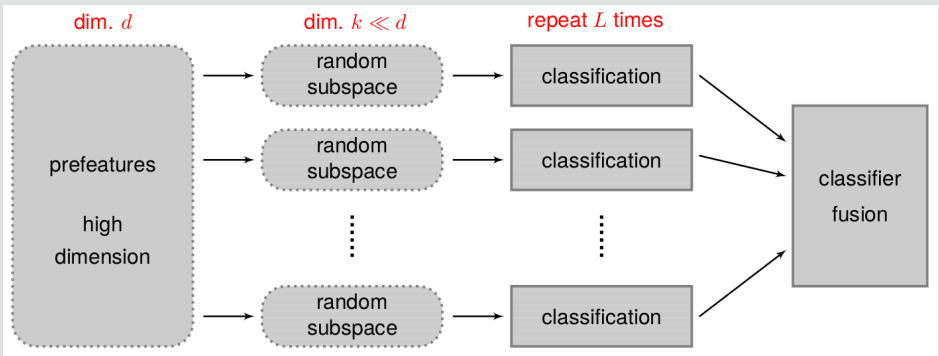
Étape 1 : construction des caractéristiques (modèle d'images)

Caractéristiques statistique des images et de leur composante de bruit :

- révélant les dépendances entre les pixels des images ;
- sensibles aux modifications, mais indépendantes du contenu de l'image.

Étape 2 : outil de classification

- N'importe quel outil par apprentissage : FLD, SVM, NN
- Compatible avec les caractéristiques statistiques
- Pratiquement : FLD Ensemble classifieur [11] (EC)



11. J. Kodovský, J. Fridrich,& V. Holub, Ensemble Classifiers for Steganalysis of Digital Media. IEEE Transactions on Information Forensics and Security, Vol. 7, No. 2, pp. 432-444, April 2012.

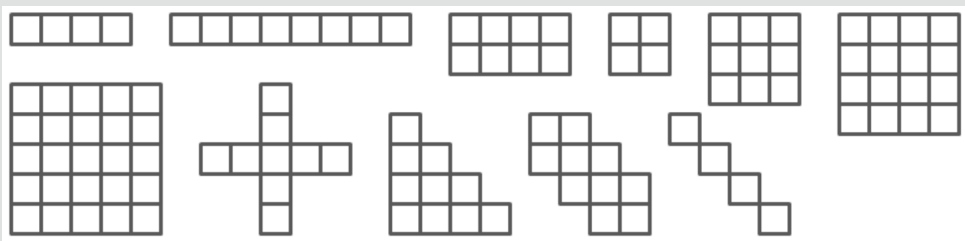
Des modèles d'images de plus en plus riches

Modèle SPAM [12] : 686 dimensions

- Calcul des variations D dans les huit directions ($\rightarrow, \swarrow, \downarrow, \ldots$)
- Pour chacune d'elle, calcul de la matrice de probabilités de modification $p(D = u | D = v)$ pour certains u, v

Modèles riches [13] : ≈34000 dimensions

- Calcul des variations du bruit : $z_{ij} = x_{ij} - Pred(Vois(x_{ij}))$
- Quantifications et troncatures variées
- Co-occurences entre pixels voisins :



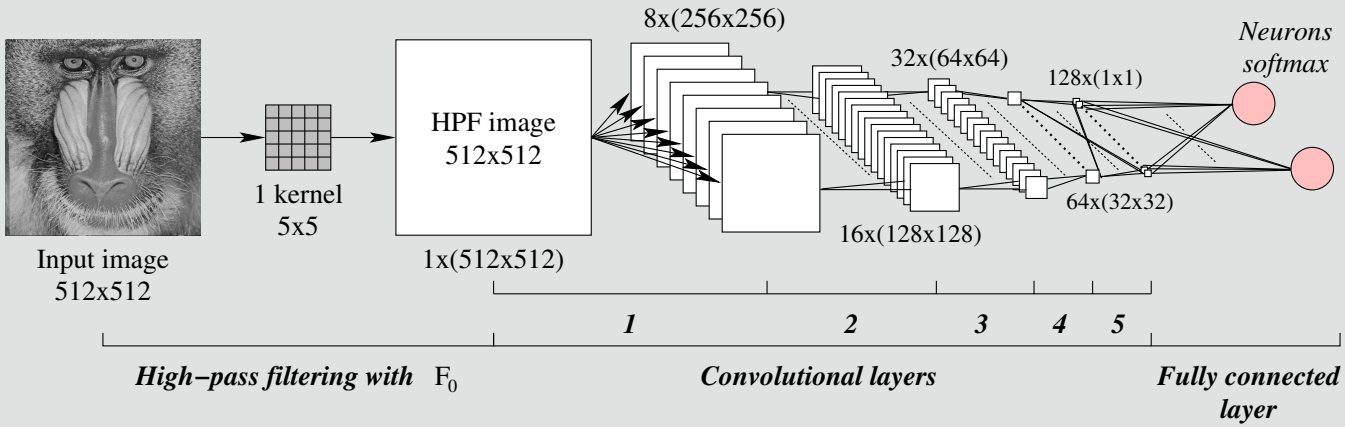
12. Pevny, T., Bas, P., & Fridrich, J. (2010). Steganalysis by subtractive pixel adjacency matrix. IEEE Transactions on information Forensics and Security, 5(2), 215-224.

13. V. Holub and J. Fridrich, Phase-Aware Projection Model for Steganalysis of JPEG Images, Proc. SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics XVII, vol. 9409, 2015.

Deep learning : une seule étape

Réseaux de neurones par convolution [14]

- Caractéristiques discriminantes inférées par le réseau
- Architecture :



14. G. Xu, H.-Z. Wu, Y.-Q. Shi, Structural Design of Convolutional Neural Networks for Steganalysis, IEEE Signal Processing Letters, vol. 23, num. 5, pp. 708–712.

Stéganographie contre stéganalyse : bilan

Règles du jeu

- Base BOSS [15] de 10,000 images en niveaux de gris 512×512
- Connu : stéganographieur et taille du message embarqué

Résultats : erreur moyenne de classification

	S-UNIWARD		MiPOD		HILL	
	0.1	0.4	0.1	0.4	0.1	0.4
Caffe[14] (CNN)	42.67	19.76	X	X	41.56	20.76
SRM + EC[9]	39.84	18.06	41.18	21.42	42.96	23.31

- Approches fournissant des résultats globalement équivalents

15. Bas, P., Filler, T., & Pevný, T. (2011, May). " Break Our Steganographic System" : The Ins and Outs of Organizing BOSS. In International Workshop on Information Hiding (pp. 59-70). Springer Berlin Heidelberg.

Plan

Réduire l'erreur de détection

Réseaux de neurones par convolution ?

- D'autres filtres, d'autres tailles, . . .
- Sans résultats

Exploiter le meilleur des deux approches ?

- Pour une image donnée : quel stéganalyseur est le plus efficace ?
- Critères pour orienter vers un des outils

Classification de SRM+EC / du CNN

Notre implantation en plus et en aveugle

	S-UNIWARD		MiPOD		HILL	
	0.1	0.4	0.1	0.4	0.1	0.4
Caffe[14]	42.67	19.76	X	X	41.56	20.76
TensorFlow (aveugle)	47.38	20.52	43.72	19.36	46.79	20.25
SRM + EC[9]	39.84	18.06	41.18	21.42	42.96	23.31
SRM + EC (aveugle)	40.57	20.85	41.18	21.42	43.35	23.99

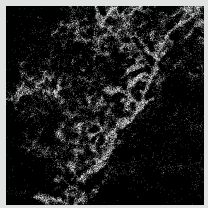
Exemples de classification par CNN

Stéganographie effectuée avec MiPOD avec un ratio de 0.4 bpp

- Situations de succès pour le CNN TensorFlow :

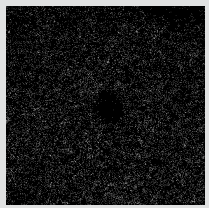
1388.pgm


8873.pgm

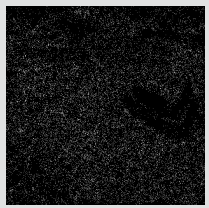



- Situations d'échec pour le CNN TensorFlow :

1911.pgm

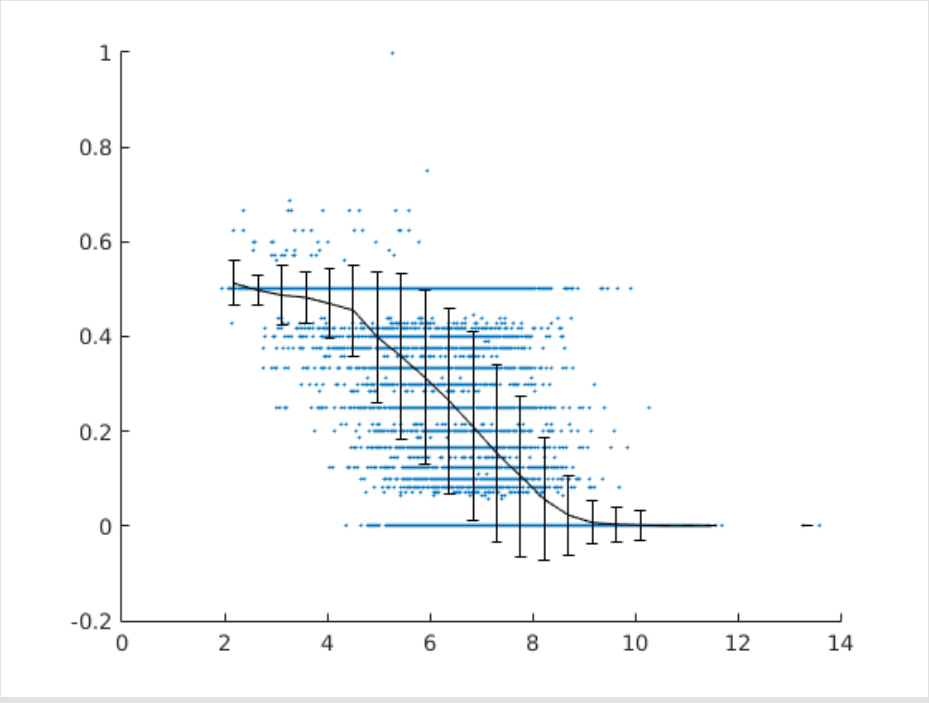



3394.pgm

Erreur de classification avec CNN-Tensor Flow p.r. $\overline{\rho_U}$

$\overline{\rho_U}$: distorsion moyenne selon S-UNIWARD



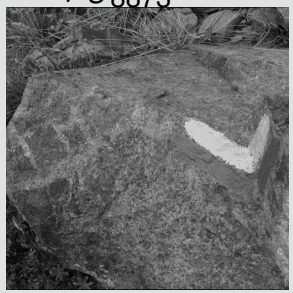
Quiz

Devinez le $\overline{\rho U}$ pour chaque image.

$\overline{\rho U}_{1388} = ?$

$\overline{\rho U}_{8873} = ?$

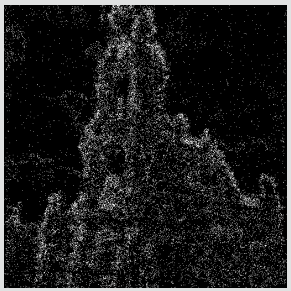
$\overline{\rho U}_{1911} = ?$



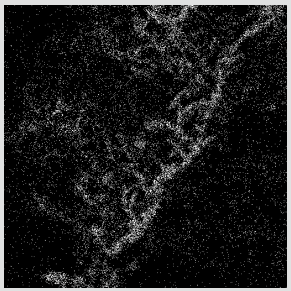
$\overline{\rho U}_{3394} = ?$

Quiz

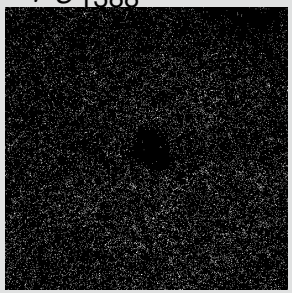
Devinez le $\overline{\rho U}$ pour chaque image.



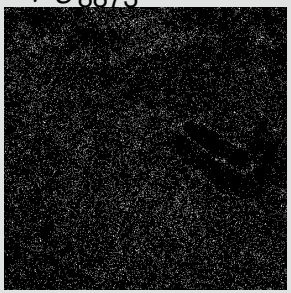
$\overline{\rho U}_{1388} = 7.05$



$\overline{\rho U}_{8873} = 7.39$



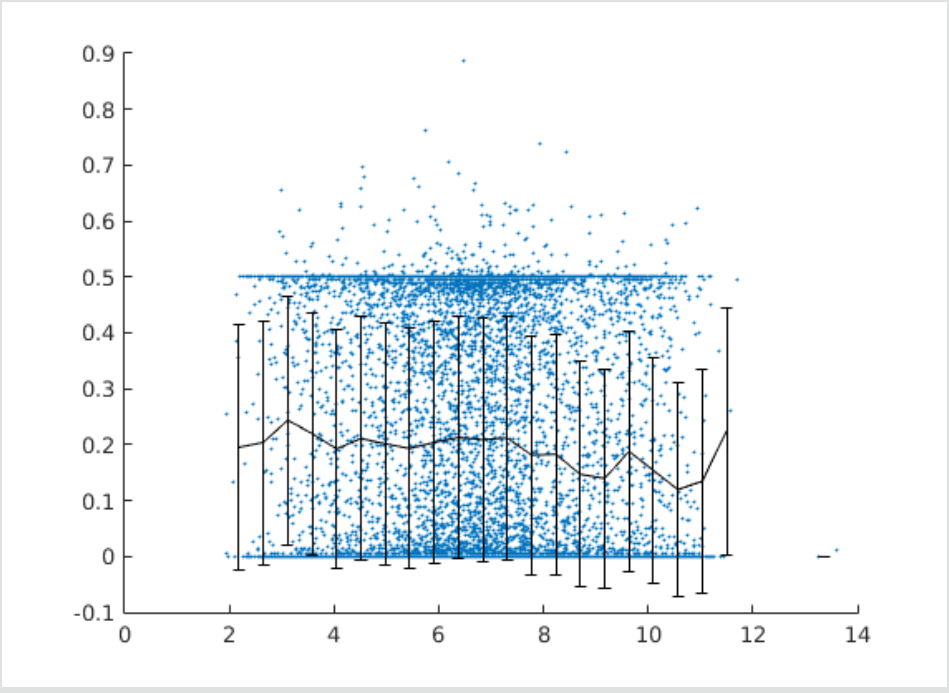
$\overline{\rho U}_{1911} = 2.1$



$\overline{\rho U}_{3394} = 3.06$

Erreur de détection selon SRM+EC p.r. $\overline{\rho U}$

200 évaluations



Comment choisir le meilleur stéganalyseur ?

Proposition de méthode

- Calcul préalable : abscisse $\overline{\rho_U^{\cap}}$ de l'intersection entre les courbes.

- Pour chaque image I :
  - calcul de la distorsion moyenne $\overline{\rho_U}(I)$ selon S-UNIWARD
  - si $\overline{\rho_U}(I) < \overline{\rho_U^{\cap}}$, utiliser le verdict de SRM+EC
  - sinon, utiliser le verdict du CNN

Résultats d'erreur moyenne de classification

Taux d'embarquement : 0.4

	SRM+EC	$\overline{\rho_U^{\cap}}$	CNN	Notre approche	SRM+EC	CNN
S-UNIWARD	20.01	7.1	8.25	14.82	18.06	19.76
S-UNIWARD aveugle	22.05	6.9	9.50	15.87	20.85	X
MiPOD	23.89	6.6	9.26	15.65	21.42	19.36
HILL	24.51	6.6	9.78	16.22	23.31	20.76
HILL aveugle	25.41	6.6	9.78	16.61	23.99	20.25

Taux d'embarquement : 0.1

	SRM+EC	$\overline{\rho_U^{\cap}}$	CNN	Notre approche	SRM+EC	CNN
S-UNIWARD	40.08	9.2	23.36	38.06	39.84	42.67
S-UNIWARD aveugle	41.00	9.2	23.36	38.88	40.57	X
MiPOD	42.13	8.0	25.84	37.82	41.18	43.72
HILL	43.48	8.9	21.88	40.24	42.96	41.56
HILL aveugle	44.30	8.3	27.72	40.64	43.35	46.79

Bilan

Résumé de cette proposition

- Choix guidé du stéganalyseur le plus efficace
- Meilleurs résultats connus en termes de stéganalyse

Développement sur GPU

- Matériel
  - Développement du réseau $\rightarrow$ 1 NVIDIA GPU Titan X
  - Exécution du réseau au Mesocentre $\rightarrow$ 4 NVIDIA GPU Tesla K40
- Durée d'entraînement sur la NVIDIA GPU Titan X
  - ratio d'embarquement de 0.4 bpp $\rightarrow \approx$ 3 jours pour des résultats "très bons "
  - ratio d'embarquement de 0.1 bpp $\rightarrow \approx$ 7 jours pour pour des résultats "bons "

Environ 15,000 heures de calcul en utilisant le Mesocentre

Publié en mai 2017

Couchot, J.-F., Couturier, R., & Salomon, M. *Improving Blind Steganalysis in Spatial Domain using a Criterion to Choose the Appropriate Steganalyzer between CNN and SRM+EC.* 32nd International Conference on ICT Systems Security and Privacy Protection - IFIP SEC 2017, May 29 - 31, 2017, Rome, Italy.

Plan

Conclusion

Résumé de la présentation

- Jeu « stéganographie contre stéganalyse » : ne fait que commencer
- Distorsion en stéganographie : ondelettes, analyse vectorielle, probabilités
- Stéganalyse : caractéristiques, deep learning

Et après ?

- Pratique : stéganographie par CNN, + de puissance de calcul
- Théorique : pourquoi les CNN font aussi bien sans connaissance ?

Compétences pour cette thématique

- Extraction d'info dans des données de grande taille (SVD, PCA)
- Des mathématiques discrètes, des probabilités, développement Tensor-Flow
- Outils de classification
- Au cœur des thématiques de AND-DISC-FEMTO-ST

Equipe AND-DISC-FEMTO-ST