Méthodes de réduction de la sensibilité à la base d'apprentissage en stéganalyse

Quentin Giboulot, Rémi Cogranne, Dirk Borghys, Patrick Bas

Méthodes de réduction de la sensibilité à la base d'apprentissage en stéganalyse

Quentin GIBOULOT[1], Rémi COGRANNE[1], Dirk BORGHYS[2], Patrick BAS[3]

[1]Laboratoire de Modélisation et de Sûreté des Systèmes (LM2S) Equipe Traverse Cyber-sécurité,
Université de Technologie de Troyes, Troyes, France.

[2]Département de Mathématiques
Académie Militaire Royale de Belgique, Bruxelles, Belgique

[3]Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL),
CNRS – École Centrale de Lille, Université de Lille, Lille, France

quentin.giboulot@utt.fr, remi.cogranne@utt.fr
dirk.borghys@rma.ac.be, patrick.bas@centralelille.fr

Résumé – Le Cover-Source Mismatch (CSM, ou inadéquation de la source d'images de couverture) a été depuis bien longtemps identifié comme l'un des problèmes les plus importants pour la détection d'information cachée. Des travaux récents ont montré que les causes du CSM sont principalement dues à la chaîne de traitement et relativement peu de la chaîne d'acquisition. Ce court papier propose de tirer parti de ces résultats en comparant expérimentalement deux approches pour lutter contre le CSM en utilisant la connaissance de la chaîne de traitement. Nous développons en particulier une méthodologie efficace pour identifier les différentes chaînes de traitements composant une base d'image. À partir de cette méthode d'identification, nous montrons expérimentalement qu'il est possible de retrouver un cadre où le CSM est quasi-absent soit en effectuant une stéganalyse ciblée sur chaque chaîne de traitement, soit en diversifiant suffisamment la base d'apprentissage.

Abstract – Cover-Source Mismatch (CSM) has long been identified as one of the most important issues for hidden information detection. Recent works have shown that the CSM finds its roots mainly in the image acquisition and processing pipeline. This short paper builds on those recent research results by comparing two approaches for mitigating the CSM using the knowledge of the image processing pipeline. In particular, we have developed an efficient methodology for identifying the differeent processing pipelines used to generate a dataset. Based on this identification method we experimentally show that it is possible to set the steganalysis problem almost free from CSM using either a targeted approach, tailored to each and every processing pipeline, or diversifying widely the training dataset.

1 Contexte de l'étude et motivations

Le phénomène du Cover-Source Mismatch (CSM) dans le cadre de la stéganalyse reposant sur l'apprentissage supervisé se caractérise de façon informelle par le manque de généralisation d'un détecteur entraîné sur une certaine source d'image lorsque celui est testé sur d'autres sources. Bien que cette sensibilité à la base d'apprentissage a été reconnue par la communauté comme un problème fondamental de la stéganalyse [1], le contexte méthodologique actuel de la discipline, reposant notamment sur une utilisation relativement standardisée de la base d'images BOSSBase, n'est pas propice à son étude approfondie ou à l'élaboration de stratégie à même de minimiser son impact.

Le CSM a été pour la première fois documenté dans [2] où il est montré que l'utilisation d'images provenant de deux appareils photographiques différents durant la phase d'entraînement et de test peut mener à une grande perte de performance de la stéganalyse même si ces images proviennent du même modèle d'appareil photo. De même, pendant la compétition BOSS, les performances obtenues par les meilleurs compétiteurs étaient significativement inférieures sur les images provenant de l'appareil photo non présent dans la base d'apprentissage [3]. Ainsi, la référence [4], supposant implicitement que le modèle d'appareil est le facteur déterminant du CSM, a quantifié l'impact de ce facteur sur les performances en stéganalyse et a montré qu'utiliser une base d'apprentissage contenant de nombreux modèles distincts d'appareils photographiques permet de réduire l'impact du CSM.

Plus tard, des travaux, [5] et [6], ont montré que la chaîne des traitements des images a également un impact sur les performances en stéganalyse. En particulier, la référence [5] montre que lorsqu'une image est sous-échantillonnée, le choix du noyau d'interpolation ainsi que le taux de sous-échantillonnage est déterminant des performances du détecteur. Dans le même courant d'idée, l'étude [6] a montré que si une image est redimensionnée par extraction plutôt que par ré-échantillonnage ou interpolation, les algorithmes d'insertion considérés comme les plus sécurisés sans le cadre standard de la BOSSBase,

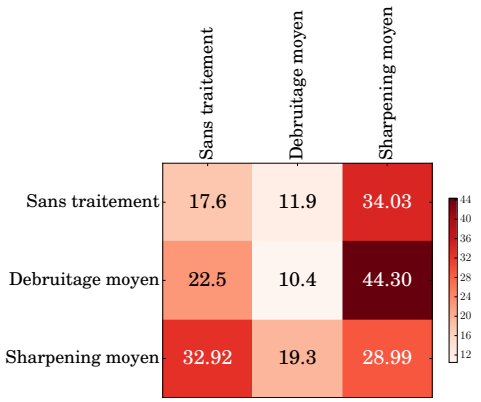

FIG. 1: Probabilité minimale d'erreur empirique (1) en fonction des traitements appliqués sur la base d'apprentissage (en ligne) et sur la base de test (en colonne).

peuvent devenir très détectables selon le choix du taux de redimensionnement.

Nos travaux, [7] et [8] ont été les premiers à étudier de façon systématique l'impact sur le CSM de l'ensemble de la chaîne d'acquisition et de traitements des images. Ces travaux ont en particulier montré l'impact considérable que pouvait avoir des éléments jusqu'ici non considérés tels que l'algorithme de dématriçage ou les opérations de traitement comme le débruitage et l'amélioration du piqué.

Les résultats de cet article sont tirés d'une étude systématique que nous avons effectuée sur le CSM [9]. Dans ce travail, nous avons en particulier montré que le CSM entre une base d'apprentissage et une base de test provient majoritairement des différences dans la chaîne de traitement, l'impact de la chaîne d'acquisition n'étant réellement important que lorsque les deux bases suivent la même chaîne de traitement.

Pour illustrer l'impact de la chaîne de traitement, la Figure 1 montre les différences de probabilité d'erreur entre différentes BOSSBase ayant suivi différents traitements. Chaque ligne correspond à un détecteur entraîné sur une de ces bases tandis que chaque colonne correspond à la base d'image sur laquelle le détecteur a été testé. La diagonale correspond ainsi à une probabilité d'erreur empirique en l'absence de CSM tandis que les autres cellules de chaque colonne correspondent à un cas où le détecteur a été entraîné sur une base d'image ayant subi un traitement différent de la base de test. On observe bien que la probabilité d'erreur sur la diagonale est toujours la plus faible pour chaque colonne. De plus, dans le cas où les traitements ne sont pas identiques entre la base d'apprentissage et la base de test, la différence de probabilité d'erreur avec la diagonale peut atteindre 25% montrant bien que la stéganalyse dans le cadre de l'apprentissage supervisé est extrêmement sensible à la base d'apprentissage.

2 Deux stratégies de réduction du CSM

Pour réduire l'impact du CSM, nous proposons d'étudier deux approches opposées :

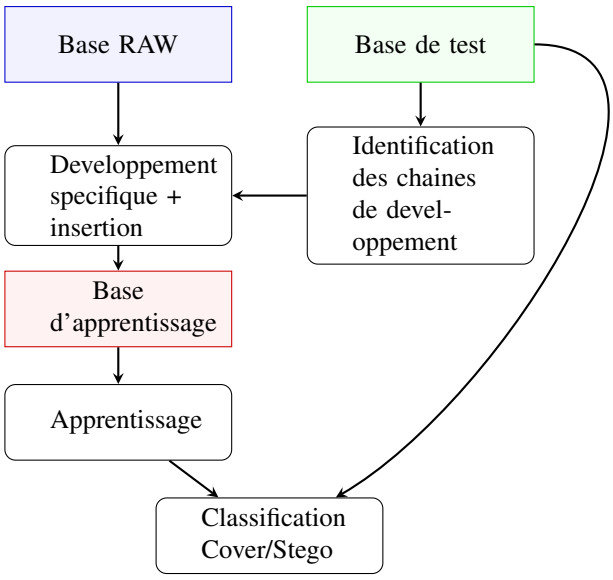

FIG. 2: Schéma illustrant l'ensemble du processus de la stéganalyse ciblée.

1. L'approche atomistique consistant à subdiviser une grande base d'image en bases plus petites ayant des propriétés similaires. L'idée est ainsi d'entraîner un détecteur spécialisé sur chaque nouvelle base d'image ainsi construite. Cette approche est inspirée des méthodes dites de "stéganalyse assistée par le *forensics*" ([10, 11]) où les auteurs ont utilisé des outils provenant de l'analyse *forensics* des images pour former des sous-ensembles d'images aux propriétés statistiques similaires. Cette approche semble naturelle une fois observée que la performance du détecteur peut énormément varier selon la/les source(s) des images de la base d'apprentissage et a été étudiée empiriquement dès les premiers travaux sur le CSM [4].
2. L'approche holistique consistant à diversifier au maximum les sources de la base d'apprentissage, permettant ainsi au détecteur d'apprendre une règle de classification plus générale qui dépendra peu de la source. Cette stratégie a été proposée explicitement dans [12] pour être ensuite utilisée dans [13] et [14].

Dans un cadre idéal, le stéganalyste a connaissance de la source utilisée pour générer les images de la base étudiée. Nous appellerons ce cas limite où le CSM est absent, il servira d'étalon à la performance des deux approches proposées.

3 Protocole expérimental

Les expériences ont été effectuées sur 22 versions différentes de la BOSSBase correspondant chacune à une chaîne de traitements distincte. Chaque version a été développée avec Rawtherapee v5.4 en utilisant différents dématriçages ou traitements d'image (débruitage et amélioration du piqué avec dif-

férents paramètres). Chaque image a ensuite été découpée au centre en dimension 512x512 puis convertie au format JPEG. Seul le canal luminance a été utilisé.La stéganographie est réalisée avec l'algorithme nsF5. La stéganalyse est effectuée à partir des caractéristiques DCTR [15] et du classifieur linéaire à faible complexité (CLFC) proposé dans [16]. La performance des détecteurs est mesurée à partir de la probabilité minimale d'erreur empirique (sous hypothèse de classes équilibrées).

$$P_E = \min_{P_{FA}} \frac{(P_{MD} + P_{FA})}{2} \quad (1)$$

Les performances obtenues dans le cadre idéal d'absence de CSM sont le fruit d'un entraînant sur 5000 paires d'images un détecteur spécifique pour chaque traitement. Chacun de ces détecteurs a ensuite été testé sur une base distincte de 5000 paires d'images ayant été sujettes aux mêmes traitements.

La première stratégie, suivant cette fois-ci l'approche atomistique, fonctionne en deux étapes. Premièrement, un classificateur permettant d'identifier le plus précisément la chaîne des traitements appliquée à chaque image est entraîné. En pratique le classificateur est CLFC entraîné dans un contexte multiclasse sur les 22 chaînes générées précédemment. Une chaîne des traitements est ainsi associée à chaque image. L'étape de stéganalyse à proprement parler consiste alors à entraîner un détecteur spécialisé pour chacune des chaînes de traitements ainsi estimées. Cette stratégie essaie ainsi de se rapprocher au maximum du cadre de la stéganalyse ciblée.

La seconde stratégie proposée, qui suit l'approche holistique, consiste à entraîner un détecteur sur une base d'apprentissage contenant une grande diversité de sources. Pour cela nous avons effectué deux expériences. La première consiste à entraîner un détecteur sur 5000 paires d'images cover/stego comprenant environ 227 paires d'images pour chaque chaîne de traitement. Pour la seconde expérience, le détecteur est entraîné sur 5000 paires d'images cover/stego par chaîne de traitements soit 110 000 paires d'images au total. Dans les deux cas, le détecteur est testé sur 5000 paires d'images non présentes dans la base d'apprentissage.

4 Principaux Résultats, Interprétations et Analyses

Dans un premier temps il est nécessaire de sélectionner les caractéristiques qui serviront à classer les différents historiques de traitements pour l'utilisation de l'approche atomistique. La Figure 3 présente la précision du détecteur selon les caractéristiques utilisées et le facteur de qualité (QF) de la compression JPEG. On observe que les caractéristiques CC-JRM [17] sont bien plus performantes que les DCTR indépendamment du facteur de qualité avec en une précision moyenne de 97% contre 89% pour DCTR pour QF100 et 81% contre 73% pour QF75. La perte de précision lorsqu'un faible facteur de qualité est employé est facilement expliquée par l'impact de la compression JPEG qui, appliquée en dernière opération, tend à homogénéiser les bases d'images. Le classificateur de chaîne des traitements pour l'approche atomistique utilisera donc les caractéristiques CC-JRM comme représentation des images.

La Figure 4 montre les performances de chaque stratégie sur chaque traitement utilisé pour un facteur de qualité JPEG 100 et l'algorithme d'insertion nsF5 avec un taux d'insertion de 0.04 bits par coefficient AC (bpAC). Des résultats similaires ont été obtenus pour J-UNIWARD à 0.3 bpAC. La stéganalyse ciblée est, comme cela est attendu, la stratégie offrant les meilleures performances. L'approche atomistique possède des performances équivalentes comme en témoigne la moyenne des P_E qui est égale pour ces deux stratégies. Cela est également attendu puisque la précision du classificateur chaîne des traitements est proche des 100%, l'approche atomistique est donc ici une excellente approximation de la stéganalyse ciblée. L'approche holistique possède également de bonnes performances, bien qu'inférieures à l'approche atomistique à condition d'utiliser un nombre suffisant d'images pour chaque chaîne des traitements. L'utilisation de 227 images/chaîne des traitements est ici largement insuffisante pour s'approcher des résultats de la stéganalyse ciblée.

Notons également que l'utilisation de ces stratégies n'est pas nécessaire dans le cas où le facteur de qualité est relativement bas (i.e. proche de 75). En effet, dans ce cas, la compression JPEG uniformise les sources amoindrissant très largement l'influence de la chaîne des traitements précédents. Par conséquent l'impact du CSM sera grandement affaibli et les trois stratégies auront des performances équivalentes.

5 Conclusions

Le but de cet article était de proposer et d'étudier les performances de quelques stratégies permettant de rendre la stéganalyse plus robuste au CSM. Nous avons proposé 3 stratégies différentes : (1) la stéganalyse ciblée, cas idéal où la source de la base de test est connue, (2) l'approche atomistique reposant sur l'identification de la chaîne de traitements pour chacune des images de la base de test, (3) l'approche holistique reposant sur la construction d'une base d'apprentissage la plus diversifiée possible. La conclusion de nos expériences est que (1) offre des performances optimales. (2) offre des performances comparables tant que l'identification de chaîne des traitements est très précise. Les performances de (3) sont inférieures à (2) mais tend à s'en approcher à condition d'augmenter la taille de la base d'apprentissage. Ce travail est une première proposition pour l'étude de méthodes permettant d'améliorer la robustesse en stéganalyse et de faire face à l'immense diversité des images numériques qui est le premier obstacle à une stéganalyse pratique, en dehors d'un cadre de recherche académique.

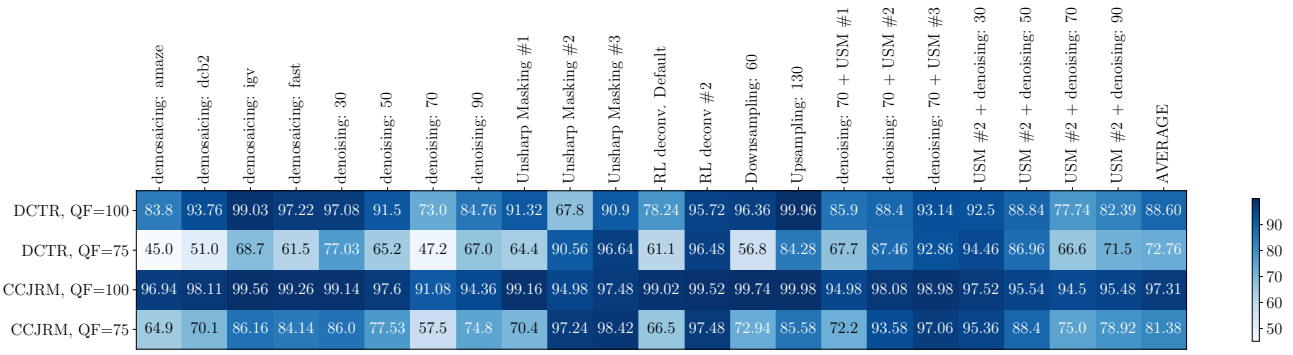

FIG. 3: Précision (%) de la classification des historiques de traitements pour différentes caractéristiques et différents facteurs de qualité JPEG

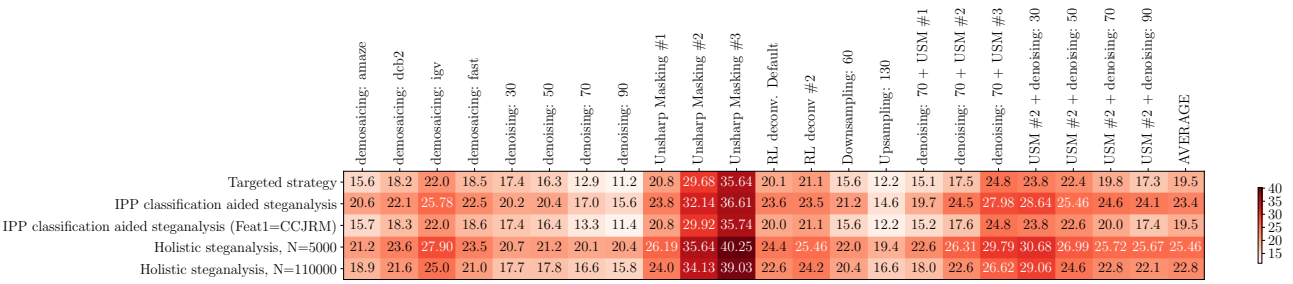

FIG. 4: P_E (%) en fonction de la stratégie employée en utilisant nsF5 comme algorithme d'insertion, QF=100. Les caractéristiques CC-JRM sont utilisées pour la classification de chaîne des traitements et les caractéristiques DCTR pour la stéganalyse.

References

[1] A. D. Ker & *al.*. Moving steganography and steganalysis from the laboratory into the real world. In *Proc. of ACM IH&MMSec*, pages 45–58, New York, NY, USA, 2013. ACM.

[2] M. Goljan, J. Fridrich, and T. Holotyak, "New blind steganalysis and its implications," in *in Proc. Electronic Imaging*, SPIE, 2006, pp. 1–13.

[3] P. Bas, T. Filler, and T. Pevný. Break our steganographic system — the ins and outs of organizing boss. In *Information Hiding*, pp. 59–70, LNCS, 2011.

[4] J. Kodovský, V. Sedighi, and J. Fridrich. Study of cover source mismatch in steganalysis and ways to mitigate its impact. In *Electronic Imaging*, volume 9028 of *Proc. SPIE*, page 90280J, Feb 2014.

[5] J. Kodovsky and J. Fridrich. Effect of image downsampling on steganographic security. *IEEE Trans. on Information Forensics and Security*, 9(5):752–762, May 2014.

[6] V. Sedighi, J. Fridrich, and R. Cogranne. Toss that bossbase, alice! In *Electronic Imaging*, Proc. IS&T, Feb 2016.

[7] Q. Giboulot, R. Cogranne, and P. Bas. Steganalysis into the wild: How to define a source? In *Electronic Imaging*, Proc. IS&T, pages 318–1 – 318–12, Jan 2018.

[8] D. Borghys, P. Bas, and H. Bruyninckx. Facing the cover-source mismatch on jphide using training-set design. In *Proc. of ACM IH&MMSec*,pp. 17–22, 2018.

[9] Q. Giboulot, R. Cogranne, D. Borghys, and P. Bas, "Roots and solutions of cover-source mismatch in image steganalysis: a comprehensive study," submitted to *IEEE Trans. on Information Forensics and Security*, Feb. 2019.

[10] M. Barni, G. Cancelli, and A. Esposito. Forensics aided steganalysis of heterogeneous images. In *Proc. ICASSP*, pages 1690–1693, March 2010.

[11] X. Hou & *al.*. Forensics aided steganalysis of heterogeneous bitmap images with different compression history. *KSII Trans. on Internet and Information Systems (TIIS)*, 6(8):1926–1945, 2012.

[12] I. Lubenko and A. D. Ker. Steganalysis with mismatched covers: Do simple classifiers help? In *Proc. of the on Multimedia and Security*, MM&Sec '12, pages 11–18, New York, NY, USA, 2012. ACM.

[13] A. D. Ker and T. Pevný. A mishmash of methods for mitigating the model mismatch mess. In *Electronic Imaging*, Proc. SPIE, volume 9028, pages 1601–1615, 2014.

[14] J. Pasquet, S. Bringay, and M. Chaumont. Steganalysis with cover-source mismatch and a small learning database. In *2014 Proc. EUSIPCO*, pages 2425–2429. IEEE, 2014.

[15] V. Holub and J. Fridrich. Low-complexity features for jpeg steganalysis using undecimated dct. *Information Forensics and Security, IEEE Trans. on*, 10(2):219–228, Feb 2015.

[16] R. Cogranne & *al.*. Is ensemble classifier needed for steganalysis in high-dimensional feature spaces? In *Proc. WIFS*, pp. 1–6, IEEE, 2015.

[17] J. Fridrich, M. Goljan, and D. Hogea. Steganalysis of jpeg images: Breaking the f5 algorithm. In *Information Hiding*, pp. 310–323, LNCS, 2003.