Détection automatique de la falsification d'images

M. Gardella[1], **T. Nikoukhah**[1], **Q. Bammey**[1], M. Colom[1], R. Grompone[1], J-M. Morel[1], P. Musé[2], D. Teyssou[3].
[1] *Centre Borelli, ENS Paris-Saclay, CNRS, Université Paris-Saclay, France,*
[2] *IIE, Facultad de Ingeniería, Universidad de la República, Uruguay,*
[3] *Medialab, Agence France-Presse, France.*

Email : {marina.gardella, tina.nikoukhah, quentin.bammey}@ens-paris-saclay.fr

Mots Clés : Falsification d'images, modèle de formation d'images, détection *a contrario*, preuve forensique, géométrie stochastique

Biographie – Marina Gardella, Tina Nikoukhah et Quentin Bammey sont doctorants au Centre Borelli. Dans leurs thèses, ils se concentrent sur la détection de falsifications dans les images, à travers les analyses du bruit (Marina Gardella), du démosaïquage (Quentin Bammey) et de la compression JPEG (Tina Nikoukhah). Ces algorithmes sont utilisés par l'Agence France-Presse et le Service de la Police Nationale Technique et Scientifique.

Resumé :

L'expansion de la photographie numérique a fortement impacté le secteur de l'information et de la communication. La reproductibilité de l'image numérique, ses transformations et sa diffusion massive, ont exacerbé le problème de son éventuelle altération. Dès sa formation et son stockage à partir d'un capteur électronique, l'image brute subit une série d'opérations : débruitage, démosaïquage, balance des blancs, corrections, compression. Dans l'image ainsi produite, ces opérations laissent des traces, imperceptibles à l'œil nu mais néanmoins souvent statistiquement significatives et donc détectables.

Aux opérations « classiques » de la formation de l'image peuvent s'ajouter des retouches locales, appelées falsifications, qui causent des perturbations dans le modèle de formation de l'image. Ces photographies numériques sont de plus en plus utilisées sur les réseaux sociaux pour créer et diffuser de fausses informations. Ainsi, pour détecter efficacement une falsification, nous tentons de reconstruire la chaîne de traitement d'une image, puis nous cherchons à détecter les anomalies dans ce modèle, autrement dit, de trouver les parties de l'image qui n'ont pas été traitées de la même manière que le reste.

L'état de l'art de la détection de falsifications s'appuie sur ces considérations pour proposer des révélateurs numériques, à savoir des opérateurs capables de mettre en évidence les zones semblant avoir été falsifiées [5]. Un examen de ces méthodes montre néanmoins qu'elles souffrent de lacunes dans l'évaluation de la confiance qui peut être attribuées à leurs détections. En l'absence d'une modélisation probabiliste et statistique, il n'est pas possible de quantifier la sûreté des détections.

À travers nos algorithmes [1, 6, 8], nous visons à recréer un historique complet des images analysées. Nous accompagnons les incohérences détectées dans l'image d'une validation statistique basée sur les méthodes de détection *a contrario*, utilisant des arguments de grande déviation [4]. Ce procédé nous permet de définir nos détections à travers la probabilité que de tels évènements se produisent par hasard, afin d'éviter les faux positifs non significatifs. Ainsi, nos outils proposent une analyse automatique des images, et ne nécessitent ni interprétation, ni donc expertise dans le domaine.

La figure 1 montre des images truquées publiées sur les réseaux sociauxainsi que les résultats de détection des algorithmes de l'état de l'art (1b et 1e) et la détection faite par nos algorithmes (1c et 1f).

Les algorithmes développés sont publiés et mis à disposition en ligne afin de pouvoir être utilisés par le plus grand nombre, en particulier par les journalistes de fact-checking. Lors du projet EnVisu4, financé par l'International Fact-Checking Network (IFCN), ils ont été intégrés dans le plugin de