La vie secrète des images JPEG

Détection de falsification via les traces de compression

Tina Nikoukhah

Cet article est le résumé de la thèse de doctorat[1] de son auteur, qui a reçu un accessit du prix de thèse Gilles Kahn 2023, décerné par la Société informatique de France.

Lors de la capture d'une image numérique, plusieurs étapes de traitement sont nécessaires avant d'obtenir le fichier final au format JPEG. Le capteur de l'appareil photo enregistre d'abord des données brutes, qui sont ensuite soumises à divers traitements, tels que le dématriçage, la balance des blancs et les corrections optiques. Enfin, l'image est compressée au format JPEG, un processus largement utilisé pour réduire la taille des fichiers, mais qui entraîne une perte d'information. Cette perte se manifeste souvent par l'apparition de blocs carrés, appelés artefacts de compression. Ces artefacts, caractéristiques de la compression JPEG, peuvent être analysés pour détecter des retouches ou falsifications, car toute modification locale de l'image perturbe l'homogénéité des traces de compression.

Contexte et motivation

Lors de la prise de vue avec un appareil photographique, les données brutes subissent généralement plusieurs traitements avant d'arriver à l'image finale. Une photographie numérique est débruitée, dématricée, corrigée des aberrations

1. https://theses.fr/2022UPASM027.

DOI:10.48556/SIF.1024.24.151

chromatiques et des distorsions optiques, puis subit des transformations non linéaires pour rehausser son contraste. Enfin, elle est compressée pour faciliter le stockage et la transmission. Le standard JPEG (*Joint Photographic Experts Group*) est le format de compression d'images numériques le plus couramment utilisé aujourd'hui pour les photographies. Cette compression entraîne une perte d'information et donc de qualité, se traduisant par l'apparition d'artefacts sous la forme de carrés, formant ce qu'on appelle la grille JPEG, comme illustré dans la figure 1.

Les étapes de compression JPEG sont décrites brièvement ci-après. Une transformation de l'espace de couleur est suivie d'un sous-échantillonnage des canaux de chrominance. Chaque canal est ensuite partitionné en blocs non chevauchants de 8×8 pixels et la transformée en cosinus discrète (DCT) de type II en 2D est appliquée à chacun de ces blocs. Chaque bloc de 8×8 pixels subit une étape de quantification effectuée dans le domaine spectral. Une table de quantification (liée à la qualité de compression, notée QF, allant de 1 à 100) fournit un facteur de quantification pour chaque composante DCT. Les artefacts JPEG sont clairement visibles lorsque l'image a été fortement compressée et sont presque imperceptibles lorsque la qualité de compression est élevée (i.e. le taux de compression est faible). Cependant, une grille est toujours présente en compression avec perte ($QF \leq 99$). À cette étape, certains coefficients DCT sont annulés lorsqu'ils ont une petite valeur par rapport au facteur de quantification. Ainsi, tous les blocs de taille 8×8 ont un nombre de zéros qui dépend à la fois de la qualité de compression et du contenu de l'image. Enfin, les coefficients DCT quantifiés sont compressés sans perte en exploitant, entre autres, la présence de valeurs nulles.

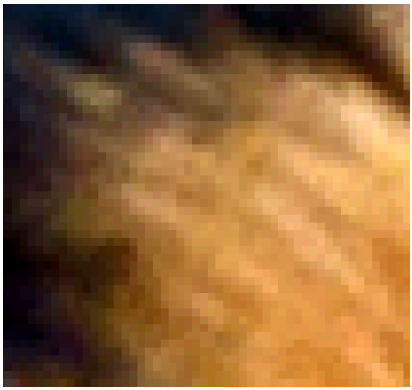

Fig. 1. Image non compressée.

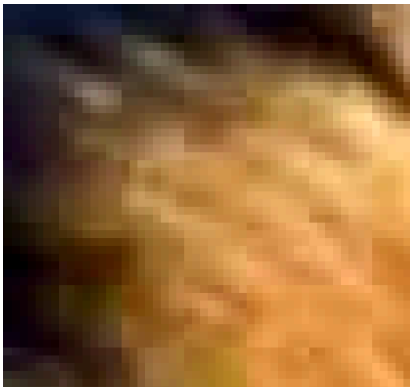

Fig. 2. Image fortement compressée.

À ces opérations classiques dans la chaîne de traitement d'une image peuvent se rajouter des modifications globales comme des changements de couleur, de luminosité ou des retouches locales, appelées falsifications. Ces opérations sont à la portée de tous grâce aux logiciels de retouche manuelle tels que *Gimp* ou *Photoshop* mais également ceux prenant en entrée du texte, tels que *Stable Diffusion* ou *Midjourney*. Le gommage, la suppression, le clonage ou l'insertion d'objets dans une image causent une rupture d'homogénéité des traces de compression. Pour détecter une falsification, les algorithmes basés sur la compression JPEG identifient des grilles locales non alignées avec la grille globale de l'image, ou des zones où la grille est absente. De plus, une zone retouchée peut être représentée comme une zone ayant une table de quantification différente de celle de l'en-tête du fichier ou de celle estimée globalement.

L'état de l'art de la détection de falsification s'appuie sur ces considérations pour proposer des révélateurs numériques [5], [8], [10], [11], [20], [3], [7], à savoir des opérateurs capables de mettre en évidence les zones semblant avoir été falsifiées. Un examen de ces méthodes montre néanmoins qu'elles souffrent de lacunes dans l'évaluation de la confiance qui peut être attribuée à leurs détections. D'où la nécessité de nouvelles méthodes donnant des résultats limitant le nombre de fausses alarmes.

Pour les algorithmes développés pendant ce travail de thèse, j'accompagne les détections dans l'image d'une validation statistique basée sur les méthodes *a contrario* [4]. Ce procédé permet de définir des détections à travers la probabilité que de tels événements se produisent par hasard, afin d'éviter les faux positifs. Ainsi, mes outils proposent une analyse automatique des images, ne nécessitant ni interprétation ni expertise dans le domaine.

Résultats principaux

L'historique de compression d'une image est définie comme l'origine de la grille globale et la table de quantification de toute compression subie par l'image. Si l'image est au format JPEG, ces informations doivent être comparées à celles de l'en-tête du fichier. Dans le cas où l'image ne serait pas au format JPEG, ce type d'analyse devient indispensable pour retrouver les informations d'une éventuelle compression JPEG antérieure.

Je propose une méthode, appelée *Qtable* [18], pour estimer la table de quantification JPEG qui se base uniquement sur les informations de l'image, sans utiliser de données provenant de l'en-tête du fichier. Pour cela, les coefficients DCT du canal de luminance de l'image sont analysés. La méthode se concentre sur les 63 coefficients AC et laisse le coefficient DC (qui a des propriétés différentes) non traité. L'histogramme de chacun des 63 coefficients est analysé et chaque valeur de quantification q entre 1 et 255 est évaluée. Ainsi, la méthode

proposée peut être appliquée à un format d'image non compressée pour détecter une précédente compression JPEG.

L'analyse des coefficients DCT et de leurs histogrammes révèle que de nombreuses valeurs ont été mises à zéro. En effet, c'est la principale raison pour laquelle la compression JPEG est efficace pour stocker une image tout en réduisant l'espace de stockage. La méthode proposée dans [16] repose sur cette observation et peut détecter la présence et les coordonnées des origines de grille globale d'une image. Par elle-même, cette méthode peut détecter si une image ayant subi une compression JPEG a été recadrée. De plus, elle peut être employée en tant qu'étape préliminaire à toute méthode de détection de falsification d'images, car nous avons observé que le fait de connaître l'origine de la grille de l'image et de la réaligner en conséquence améliore l'efficacité de ces méthodes.

Dans le cas d'une image falsifiée, les méthodes décrites précédemment indiquent que l'image a subi plusieurs compressions, car la table Q estimée est différente de celle qui est stockée dans l'en-tête du fichier et qu'il y a plusieurs origines de grille significatives. Cela n'atteste pas une falsification de l'image, mais plutôt fournit des informations intéressantes sur son historique, susceptibles de susciter des doutes. Pour détecter une altération, une approche consiste à repérer des grilles locales qui ne sont pas alignées avec la grille globale de l'image, ou à identifier une zone où la grille est absente. De plus, une zone retouchée peut se manifester par une table de quantification différente de celle mentionnée dans l'en-tête du fichier ou de celle estimée globalement.

La première approche est présentée dans [19] et [17]. La méthode exploite les traces d'artefacts en blocs pour récupérer localement la grille intégrée dans l'image par la compression JPEG. L'algorithme renvoie une liste de grilles associées à chaque partie de l'image. Le seul paramètre qui nécessite une intervention humaine est le choix de la taille des fenêtres utilisées. Enfin, la méthode ZERO [15], [14] qui est plus précise, plus rapide et qui ne nécessite aucun paramètre a été élaborée.

L'algorithme ZERO compte le nombre de zéros pour détecter l'origine de la grille dans chaque fenêtre d'une image. En effet, ZERO exploite le fait que la compression JPEG met des valeurs DCT à zéro et identifie la présence d'une grille JPEG lorsqu'un nombre significatif de zéros est observé pour une origine de grille donnée. Chaque pixel peut appartenir à 64 blocs différents qui se chevauchent et vote pour choisir l'origine de la grille du bloc avec le plus de zéros. La méthode est appliquée de trois manières : d'abord globalement pour détecter l'origine de grille globale de l'image. Ensuite, localement pour détecter les grilles étrangères et enfin, à une version alternative de l'image pour détecter les zones où la grille globale est absente.

Les algorithmes développés se distinguent nettement des approches conventionnelles par leur gestion rigoureuse des faux positifs, notamment en ne

détectant aucun événement dans des images n'ayant jamais subi de compression JPEG, là où d'autres méthodes échouent. Cette efficacité repose sur l'adoption de l'approche *a contrario*, un cadre qui, contrairement aux méthodes basées sur le seuillage simple d'un score de confiance, utilise des tests d'hypothèses pour contrôler l'espérance des fausses alarmes en tenant compte du nombre de tests effectués. Cette méthodologie offre une rigueur mathématique permettant des décisions de détection sans interprétation experte et garantit une fiabilité supérieure des résultats.

En collaboration avec deux autres doctorants, nous avons développé une méthodologie et un jeu de données, nommés *Trace* [2], pour évaluer la fiabilité des méthodes de détection de falsification d'images en se focalisant sur leur sensibilité face à des artefacts spécifiques. Notre approche, en plus d'évaluer, révèle les points forts et faibles des méthodes testées. La méthode ZERO se distingue par sa capacité à éviter les faux positifs, affichant d'excellentes performances comparées à 16 autres algorithmes, malgré certaines limitations comme l'omission de petites falsifications ou la délimitation imprécise des zones altérées.

Impact des travaux

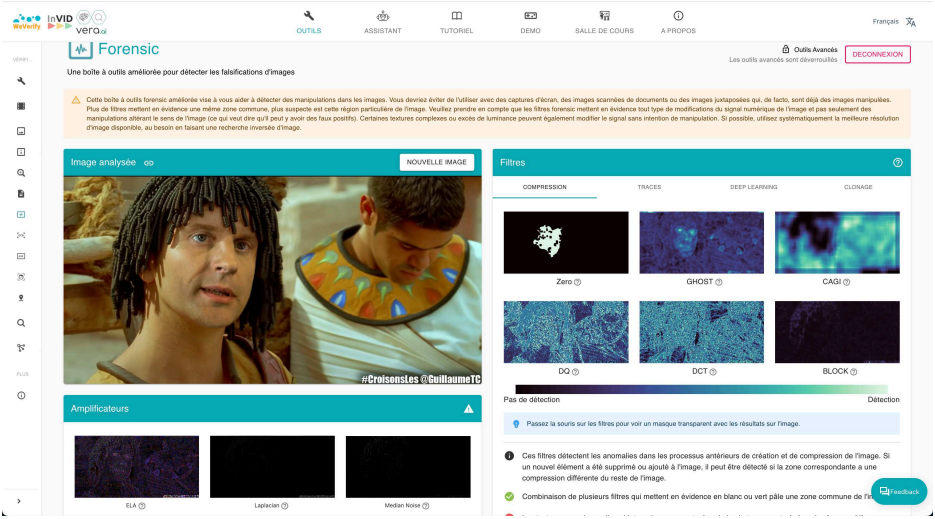

Fig. 3. Capture d'écran de l'outil InVID-WeVerify de l'AFP. La méthode zero [14] se trouve parmi les outils d'analyse d'images (l'image testée provient du compte X de @GuillaumeTC).

Pour assurer la reproductibilité de mes travaux, je publie mes algorithmes sur IPOL (*Image Processing On Line*, ipol.im), une revue de science ouverte. Cette revue accorde une attention particulière aux détails mathématiques et réalise une évaluation stricte par les pairs pour vérifier que le code source correspond aux spécifications de l'article. Chaque article publié dans le journal IPOL détaille la méthode utilisée, présente le pseudo-code, et inclut le code source correspondant. De plus, un démonstrateur en ligne est mis à disposition, offrant la possibilité de tester la méthode avec ses propres images [17], [14], [18].

Cela a permis à ces méthodes d'être largement testées, et en raison de sa rapidité, de sa fiabilité et de sa précision, l'algorithme ZERO a été inclus dans le plugin InVID-WeVerify [9] (comme illustré figure 3). Cet outil, créé par l'Agence France-Presse (AFP), est dédié à la vérification d'informations. Il est accessible à tous et est utilisé principalement par les journalistes et *fact-checkers* dans 224 pays [1]. À travers cet outil, la méthode ZERO a notamment permis de détecter des cas réels de falsification, comme illustré dans cet article de France24 [6].

Par ailleurs, j'ai également adapté les sujets de cette thèse pour différents publics. Ils sont enseignés dans le cadre du cours de traitement des images, faisant partie de la formation en « Éducation aux médias et à l'information » dans le cadre du diplôme universitaire entre l'ENS Paris-Saclay et l'ESJ Lille. Ces idées sont également présentées lors de conférences et diffusées à travers des publications destinées au grand public [12], [13], y compris aux jeunes.

Références

[1] Agence France-Presse. 2023. Le medialab de l'AFP. Retrieved from `https://www.afp.com/fr/lagence/medialab`.

[2] Quentin Bammey, Tina Nikoukhah, Marina Gardella, Rafael Grompone von Gioi, Miguel Colom, and Jean-Michel Morel. 2022. Non-semantic evaluation of image forensics tools: Methodology and database. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3751–3760.

[3] Davide Cozzolino and Luisa Verdoliva. 2020. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security* 15:144–159. `https://doi.org/10.1109/TIFS.2019.2916364`.

[4] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. 2007. *From gestalt theory to image analysis. Interdisciplinary applied mathematics*. Springer, New York. `https://doi.org/10.1023/A:1026593302236`.

[5] Hany Farid. 2009. Exposing digital forgeries from JPEG ghosts. *IEEE transactions on information forensics and security* 4, 1:154–160.

[6] France24. 2023. Une photo de bébé israélien calciné générée par IA : Quand les outils de détection brouillent le débat public. Retrieved from `https://observers.france24.com/fr/moyen-orient/20231017-une-photo-de-b%C3%A9b%C3%A9-isra%C3%A9lien-calcin%C3%A9-g%C3%A9n%C3%A9r%C3%A9e-par-ia-quand-les-outils-de-d%C3%A9tection-brouillent-le-d%C3%A9bat-public`.

[7] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20606–20615.

[8] Chryssanthi Iakovidou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Content-aware detection of JPEG grid inconsistencies for intuitive image forensics. *Journal of Visual Communication and Image Representation* 54:155–170.

[9] InVID-WeVerify. 2023. InVID-WeVerify plugin. Retrieved from `https://weverify.eu/verification-plugin/`.

[10] Neal Krawetz and Hacker Factor Solutions. 2007. A picture's worth. *Hacker Factor Solutions* 6, 2.

[11] Hannes Mareen, Dante Vanden Bussche, Fabrizio Guillaro, Davide Cozzolino, Glenn Van Wallendael, Peter Lambert, and Luisa Verdoliva. 2022. Comprint: Image forgery detection and localization using compression fingerprints. In *International conference on pattern recognition*, 281–299.

[12] Tina Nikoukhah. 2020. Tout ce que les algorithmes de traitement d'images font pour nous. *Interstices*. Retrieved from `https://hal.inria.fr/hal-02929192`.

[13] Tina Nikoukhah. 2021. Les traces de compression pour détecter les photomontages. *Interstices*. Retrieved from `https://hal.inria.fr/hal-03220223`.

[14] Tina Nikoukhah, Jérémy Anger, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. 2021. ZERO: a Local JPEG Grid Origin Detector Based on the Number of DCT Zeros and its Applications in Image Forensics. *Image Processing On Line* 11:396–433.

[15] Tina Nikoukhah, Jérémy Anger, Thibaud Ehret, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. 2019. JPEG grid detection based on the number of DCT zeros and its application to automatic and localized forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 110–118.

[16] Tina Nikoukhah, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. 2019. Détection de grille JPEG par compression simulée. In *XXVIIe colloque sur le traitement du signal et des images*, 641–644.

[17] Tina Nikoukhah, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. 2020. Local JPEG Grid Detector via Blocking Artifacts, a Forgery Detection Tool. *Image Processing On Line* 10:24–42.

[18] Tina Nikoukhah, Miguel Colom, Jean-Michel Morel, and Rafael Grompone von Gioi. 2022. A Reliable JPEG Quantization Table Estimator. *Image Processing On Line* 12:173–197.

[19] Tina Nikoukhah, Rafael Grompone von Gioi, Miguel Colom, and Jean-Michel Morel. 2018. Automatic JPEG grid detection with controlled false alarms, and its image forensic applications. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, 378–383.

[20] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2019. ManTra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.