Contribution des signaux résiduels pour la détection de la permutation de visages dans les vidéos hypertruquées

P. Tessé C. Charrier E. Giguet

Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

{paul.tesse, christophe.charrier, emmanuel.giguet}@unicaen.fr

Résumé

L'évolution fulgurante de l'apprentissage profond et plus particulièrement la découverte des réseaux antagonistes génératifs (RAG) a révolutionné le monde du Deepfake. Les falsifications sont de plus en plus réalistes et par conséquent de plus en plus difficiles à détecter. Attester si un contenu vidéo est authentique est de plus en plus sensible et le libre accès aux technologies de falsification rend la menace d'autant plus inquiétante. De nombreuses méthodes ont été proposées pour détecter ces faux et il est difficile de savoir quelles méthodes de détection sont encore d'actualité face aux progrès. Dans cet article, nous présentons notre approche pour la détection de permutation de visages dans les vidéos hypertruquées basée sur l'analyse des signaux résiduels.

Mots clefs

Vidéos hypertruquées, permutation de visages signaux résiduels, investigation numérique, apprentissage profond.

1 Introduction

Notre société hyperconnectée voit transiter des quantités de contenus multimédia de plus en plus importantes, que ce soit via la télévision, la vidéo surveillance, les réseaux sociaux et plus généralement internet. Ceci est dû aux progrès réalisés ces dernières années en matière de création et de partage de contenus vidéos. En couplant ces progrès avec les avancées réalisées dans le domaine de l'apprentissage machine, et plus particulièrement de l'apprentissage profond, nous assistons à une hausse très significative du nombre de faux contenus multimédia, en particulier les vidéos hypertruquées, aussi appelées *deepfakes*. De nouveaux outils de falsification très performants sont librement accessibles et de plus en plus simples d'utilisation. Certains de ces modèles sont d'ores et déjà intégrés à des réseaux sociaux tels que Snapchat et accessibles à tout utilisateur sous le nom de "filtres". Cette démocratisation des outils de falsification vidéo est à l'origine de la hausse significative du nombre de fake news, vidéos de propagande, tentative d'usurpation vidéo, etc. La détection de ces vidéos falsifiées représente par conséquent un enjeux sociétal majeur. En effet, il est de plus en plus difficile d'attester l'authenticité d'une vidéo, ce qui est très préoccupant dans notre société où chaque jour, les heures de visionnage uniquement sur Youtube se comptent en milliards.

La détection des vidéos hypertruquées est un sujet particulièrement ardent ces derniers temps bien que de nombreux chercheurs travaillent sur le sujet depuis des années. De nombreux articles traitent de ce sujet sous des angles variés. Parmi les approches les plus performantes, celles basées sur l'apprentissage profond sont majoritairement plébiscités, ce qui n'est pas sans poser de problèmes en terme d'explicabilité et du biais récurrent induit durant la phase d'apprentissage, voire du transfert d'apprentissage. Afin de pallier ces deux inconvénients, l'approche que nous avons retenue est fondée sur l'utilisation conjointe d'informations extraites des signaux résiduels et de réseaux de neurones.

La structure de l'article est la suivante. Une formalisation du problème est proposée dans la section 2. La section 3 dresse un panorama des méthodes de détection de permetutation de visages, basées notamment sur l'utilisation des modèles génératifs adverses et sur les techniques issues de la criminalistique des images. La section 4 décrit la méthode d'analyse que nous proposons. Les résultats sont présentés en section 5. La conclusion met en avant les perspectives de ce travail.

2 Formalisation du problème

La problématique étudiée étant la détection des deepfakes vidéos basés sur la permutation de visages, les éléments essentiels pris en considération dans la formalisation du problème sont les suivants :

- les vidéos sont de durée variable avec un trucage pouvant survenir à n'importe quel endroit ou moment ;
- un mécanisme de détection des visages est nécessaire puisqu'il permet de cibler la zone à étudier ;
- le modèle doit être le plus robuste et généralisable possible ;
- le modèle devant pouvoir être utilisé pour éclairer la Justice, une attention toute particulière doit être accordée à l'explicabilité des résultats ;
- le modèle doit fonctionner sans référence pour prononcer son diagnostic ;
- l'analyse d'images synthétiques n'est pas prise en compte dans ces travaux.

Ces aspects pris en compte, notre objectif est de développer un module prenant en entrée une vidéo et retournant en sortie un verdict concernant l'authenticité de cette dernière. Le problème est donc envisagé comme un problème de classification binaire où les classes sont "authentique" et "falsifiée".

3 Etat de l'art

De très nombreuses méthodes de détection de vidéos hypertruquées ont été proposées au cours des dernières années. Parmi les méthodes existantes, nous nous sommes tout d'abord intéressés aux méthodes d'apprentissage profond qui ont montré un niveau de performance élevé dans les tâches de classification au détriment de l'explicabilité du verdict [1]. C'est pourquoi nous avons laissé de côté ces modèles et avons concentré nos efforts sur les méthodes d'analyse des signaux résiduels, celles-ci étant totalement explicables. Voici les deux signaux résiduels que nous avons sélectionnés jusqu'à présent.

3.1 Evaluation de la qualité des images

Parmi les méthodes les plus répandues, on retrouve la mesure de la qualité des images (IQA–*Image Quality Assessment*). En effet, de nombreuses études ont montré que la qualité des images est altérée suite à la falsification [2]. Cette information est *de facto* pertinente et sera exploitée comme telle dans la tâche de classifation. Etant donné que nous ne disposons pas de l'image de référence, on s'attachera à utiliser une mesure de qualité des images *sans référence*. Parmi toutes les méthodes existantes, nous avons sélectionné l'indice de qualité BRISQUE (*Blind/Referenceless Image Spatial Quality Evaluator*) [3]. Ce dernier ne calcule pas les caractéristiques spécifiques aux distorsions, telles que l'effet le flou, de ringing ou de bloc, mais utilise les statistiques de scènes naturelles des coefficients de luminance normalisés localement pour quantifier les éventuelles pertes de « naturel » dans l'image dues à la présence de distorsions, ce qui aboutit à une mesure holistique de la qualité.

3.2 Analyse du spectre fréquentiel

Une autre approche consiste à étudier le spectre fréquentiel des images. Ce changement de représentation est motivé par un constat très intéressant présenté dans [4]. En effet, les auteurs ont mis en exergue un phénomène lié à l'utilisation des GANs dans les modèles générateurs de deepfake tel que le StyleGAN [5]. L'utilisation des opérations d'upsampling est nécessaire dans le processus de génération afin d'augmenter la dimensionnalité tout au long du processus. Cette opération utilise une opération d'interpolation qui est à l'origine d'une augmentation de l'utilisation des hautes fréquences dans la représentation de l'image. Cette introduction de hautes fréquences est alors un indice qui peut être exploité afin de déterminer si une vidéo est authentique ou non.

3.3 Autres signaux résiduels

D'autres signaux résiduels sont également exploitables. Nous avons pour l'instant concentré nos efforts sur les deux premiers mais l'on peut en citer de nombreux autres tels que l'analyse de la Lateral Chromatic Aberration [6] ou encore des Color Filter Array [7] Artefacts, qui se concentrent sur l'analyse des d'artéfacts induits par les différences dans les systèmes d'acquisition des sources des images mélangées pour générer le deepfake.

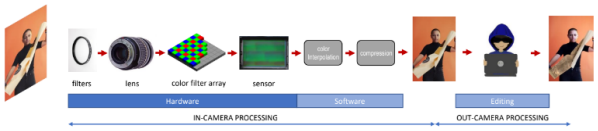

FIGURE 1 – *Système d'acquisition image numérique [8]*

4 Architecture proposée

Afin de combiner la puissance des modèles d'apprentissage profond avec l'explicabilité des méthodes basées sur l'analyse des signaux résiduels que nous avons présentés précédemment, nous proposons l'architecture suivante. L'architecture proposée, telle qu'illustrée dans la figure 3, se décompose en quatre étapes :

1. La vidéo est prétraitée pour obtenir les frames (F) et ne conserver que le visage qui est la zone de l'attaque pour plus de précision et une optimisation en terme de coûts.
2. Les images sont ensuite passées à différents extracteurs de caractéristiques (FE) qui vont extraire des caractéristiques pertinentes telles que le score de qualité via la mesure BRISQUE, la représentation fréquentielle de l'image ou le ratio des hautes fréquences.
3. Ces différentes caractéristiques sont ensuite concaténées en une seule représentation pour le Classifieur afin d'augmenter la robustesse de ce dernier.
4. Le Classifieur qui sera à terme un modèle d'apprentissage profond à définir quant à lui procède à la classification binaire entre les classes *Authentique* et *Falsifiée*.

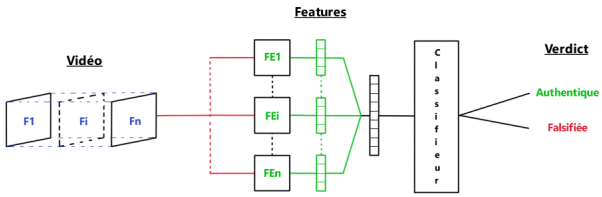

FIGURE 2 – *Architecture proposée*

L'intérêt de cette architecture est selon nous de proposer une alternative en boîte grise. En effet, les extracteurs de caractéristiques sont des boîtes blanches puisqu'ils n'utilisent pas d'apprentissage profond et seul le classifieur sera une boîte noire. De cette manière nous pensons pouvoir conserver un bon équilibre entre performance et explicabilité. Enfin, cette architecture est évolutive puisque la modularité permet d'ajouter simplement de nouveaux extracteurs de caractéristiques et seul le classifieur sera à entraîner, ce qui assure une meilleure durabilité du modèle dans le temps.

5 Expérimentations et résultats

Afin d'étudier ces signaux résiduels et leur pertinence plus en détail, nous avons réalisé plusieurs expérimentations. A l'heure actuelle nous n'avons pu nous intéresser qu'à BRISQUE ainsi qu'à la représentation fréquentielle. Ces expérimentations ont été réalisées sur les vidéos issues des base de données VidTIMIT [9] et DeepfakeTIMIT [10] qui contiennent respectivement les échantillons authentiques et falsifiés. Ces vidéos de haute qualité ont été traitées de sorte à ne conserver que les visages dans les images d'origines. Il est important de préciser que nous utilisons un SVM en guise de classifieur dans cette étude préliminaire au vu du peu de données que nous avons.

5.1 Indice de qualité

En reprenant l'implémentation fournie par les auteurs sur Github [3], nous avons été en mesure de calculer le score de qualité pour une image. Notre modèle recevant une vidéo en input, nous nous sommes intéressés au calcul de ce score à l'échelle de la vidéo. C'est pourquoi nous avons calculé la moyenne et l'écart-type de ce score à partir du score de chaque frame. Voici un échantillon des résultats obtenus en appliquant notre extracteur de caractéristiques sur les deux bases de données pour les vidéos authentiques (Tableau 1) et les vidéos truquées (Tableau 2).

VideoId	faks0-sa1	fcft0-sa2	mdab0-sx49	BDD
mean±std	21.4 ± 1.81	31.9 ± 1.68	26.6 ± 1.80	24.28 ± 2.88

TABLEAU 1 – *BRISQUE Scores vidéos Authentiques*

VideoId	faks0-sa1	fcft0-sa2	mdab0-sx49	BDD
mean±std	31.5 ± 2.15	42.5 ± 1.40	38 ± 1.45	32.9 ± 1.93

TABLEAU 2 – *BRISQUE Scores vidéos Hypertruquées*

On peut constater, et ce à l'échelle de l'ensemble des paires de vidéos authentiques/falsifiées, que la qualité moyenne semble se dégrader systématiquement et ce de manière significative. Nous rappelons que le score varie entre 0 et 100 avec 0 qui correspond à la qualité optimale. Pour ce qui est de l'écart-type, la variation est moins significative mais celle-ci a tendance à diminuer contrairement à la moyenne. Cette tendance dans les résultats semble conservée à l'échelle des bases de données au vu de la colonne BDD. Cela tend à confirmer que ces scores pourraient bien servir de caractéristiques pour notre classifieur.

5.2 Hautes fréquences

De la même manière que pour les tests sur BRISQUE, nous avons repris l'implémentation des auteurs [4] et l'avons reprise afin de permettre de générer la représentation fréquentielle d'une vidéo. Encore une fois, nous avons généré les résultats sous la forme de paires dont un échantillon est présenté sur la figure 3.

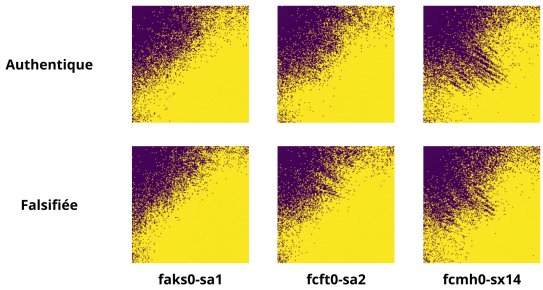

FIGURE 3 – *Visualisation spectre fréquentiel vidéos*

On peut observer une hausse des hautes fréquences que nous avons essayé de quantifier plus finement avec la différence entre les ratios des vidéos authentiques et falsifiées, correspondant au rapport entre le nombre de valeurs de pixels supérieures à 150 et le nombre de pixels total.

VideoId	faks0-sa1	fcft0-sa2	mdab0-sx49	BDD
ΔHF	+2.1%	+2.6%	+1.8%	+0.5%

TABLEAU 3 – *Variation des hautes fréquences*

Les résultats du tableau 3 confirment bien notre analyse qualitative des spectres. Il y a une augmentation légère mais qui peut rester perceptible pour notre classifieur qui est persistante d'après les résultats à l'échelle de la base de données (BDD) bien que l'augmentation soit plus faible. Ceci peut s'expliquer par le fait que nous calculons une première moyenne entre les frames, puis entre toutes les vidéos ce qui produit un effet de lissage. Néanmoins, on constate qu'il reste une variation qui pourrait être exploitable par notre classifieur. Dans notre cas, nous avons fait le choix d'utiliser ce ratio en guise de caractéristique étant donné qu'il s'agit d'un score normalisé représenté par un simple scalaire.

5.3 Classification par SVM

Afin de statuer sur la pertinence des caractéristiques présentées, nous avons testé la détection des deepfakes en utilisant le classifieur SVC de Scikit Learn 3 avec les réglages par défaut. Pour cela nous avons extrait les différentes caractéristiques des vidéos issues des bases de données Vid-

TIMIT et DeepfakeTIMIT. Les caractéristiques ainsi obtenues ont été divisées en un jeu d'entraînement (Train) et un jeu de validation. Ce même processus a été appliqué à un échantillon de la base de données Celeb-DeepFake [11] afin de générer des données de test (Test) pour avoir un aperçu des performances en généralisation. Les résultats obtenus avec les différents ensembles sont présentés dans le tableau 4. Les résultats présentés ont été obtenus en appliquant un Bootstrap à 999 réplications. La composition des ensembles utilisés est présentée dans le tableau 5.

SVM	BRISQUE	Somme HFs	Concaténés
Train	88% $\pm$ 0.008	60% $\pm$ 0.006	86% $\pm$ 0.006
Val	87% $\pm$ 0.02	59% $\pm$ 0.02	85% $\pm$ 0.03
Test	48% $\pm$ 0.01	45% $\pm$ 0.03	46% $\pm$ 0.01

TABLEAU 4 – *Précision Classification SVM*

Ensembles	Train	Validation	Test
Source(s)	VidTIMIT DeepfakeTIMIT	VidTIMIT DeepfakeTIMIT	CelebDeepFake
Taille	580	286	103
Repartition	A=436/F=256	A=110/F=64	A=51/F=52

TABLEAU 5 – *Composition des ensembles où A correspond au nombre de vidéos non falsifées et F au nombre de vidéos hypertruquées*

Nos résultats sont au dessus des 50% ce qui signifie que nos prédictions sont plus fiables que le hasard bien que l'on observe une baisse systématique et significative des performances en généralisation. Cette baisse de performance peut être due à plusieurs facteurs tels que la quantité de données qui reste assez faible, le fait que les données d'entraînement ne soient issues que d'un seul jeu de données, ou encore tout simplement le modèle en lui-même qui reste trop simple. Les résultats relatifs à l'utilisation du ratio des hautes fréquences nous laisse penser qu'il est nécessaire d'utiliser un CNN afin d'exploiter au maximum les informations contenues dans le spectre et non pas un simple ratio. De plus, la combinaison des deux semble bien indiquer que l'utilisation du ratio des hautes fréquences n'améliore pas les performances obtenues avec BRISQUE.

6 Conclusion

Nous avons présenté dans cet article nos travaux préliminaires relatifs à la détection de vidéos hypertruquées, aussi appelées *Deepfakes*. Nous avons pu démontrer que les signaux résiduels constituent bel et bien une piste sérieuse de caractéristiques pertinentes et explicables. Il est en effet possible pour un classifieur, comme le montre les résultats obtenus, d'exploiter ces signaux afin de résoudre notre problème de détection. Nous n'avons pour le moment pu tester que deux extracteurs de caractéristiques avec un simple SVM en guise de classifieur. C'est pourquoi il nous faut procéder à davantage de tests sur ces derniers afin de confirmer ces premiers résultats expérimentaux. Dans un second temps nous incorporons d'autres signaux résiduels tout en améliorant le classifieur afin d'améliorer les performances de notre architecture. Enfin, un travail de passage à l'échelle reste à effectuer afin d'obtenir le plus de précision et de recul possible quant à l'évaluation de ces performances.

Références

[1] David Güera et Edward J Delp. Deepfake video detection using recurrent neural networks. Dans *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[2] Javier Galbally et Sébastien Marcel. Face anti-spoofing based on general image quality assessment. *Proceedings - International Conference on Pattern Recognition*, pages 1173–1178, 08 2014.

[3] Anish Mittal, Anush Krishna Moorthy, et Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12) :4695–4708, 2012.

[4] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, et Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. Dans *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

[5] Tero Karras, Samuli Laine, et Timo Aila. A style-based generator architecture for generative adversarial networks. Dans *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 4396–4405, 2019.

[6] Owen Mayer et Matthew C. Stamm. Accurate and efficient image forgery detection using lateral chromatic aberration. *IEEE Transactions on Information Forensics and Security*, 13(7) :1762–1777, 2018.

[7] Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, et Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5) :1566–1577, 2012.

[8] Luisa Verdoliva. Media forensics and deepfakes : An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14 :910–932, 2020.

[9] C. Sanderson et B.C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. *Lecture Notes in Computer Science (LNCS)*, 5558 :199–208, 2009.

[10] Pavel Korshunov et Sébastien Marcel. Deepfakes : a new threat to face recognition ? assessment and detection. *ArXiv*, abs/1812.08685, 2018.

[11] Pu Sun Honggang Qi Yuezun Li, Xin Yang et Siwei Lyu. Celeb-df : A large-scale challenging dataset for deepfake forensics. Dans *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.