22e édition du colloque CORESA COmpression et REprésentation des Signaux Audiovisuels

7-9 juin 2023
Lille, France

Comité local d'organisation :

Marius BILASCO
Mohamed DAOUDI
Olivier LOSSON
Ludovic MACAIRE
Benjamin MATHON
Deise SANTANA MAIA

Sponsors

CRIStAL Centre de Recherche en Informatique, Signal et Automatique de Lille
Université de Lille
Région Hauts-de-France

Partenaires

cnrs
IMT Nord Europe École Mines-Télécom IMT-Université de Lille
GdR ISIS

Table des matières

1 Image, vidéo et géométrie **7**

Annotation semi-automatique de base de données d'images complexes non standardisées 8
Guillaume Picaud, Marc Chaumont, Gérard Subsol, Luc Teot

Détection de contenu explicite dans les vidéos 12
Hugo Jean, Emmanuel Giguet, Christophe Rosenberger

Réseau de neurones convolutif pour l'extraction d'attributs de texture à partir d'images multispectrales 16
Anis Amziane, Olivier Losson, Benjamin Mathon, Ludovic Macaire

2D versus 3D Convolutional Spiking Neural Networks Trained with Unsupervised STDP for Human Action Recognition 20
Mireille El-Assal, Pierre Tirilly, Ioan Marius Bilasco

Principal Geodesic Analysis of Merge Trees (and Persistence Diagrams) . 21
Mathieu Pont, Jules Vidal, Julien Tierny

2 Codage et compression **23**

Analysis of the influence of errors in DNA-based image coding 24
Jorge Encinas Ramos, Davi Lazzarotto, Michela Testolina, Touradj Ebrahimi

Etude de la faisabilité d'une compensation efficace de la latence par extrapolation des images vidéo 28
Hind Kanj, Anthony Trioux, Marco Cagnazzo, François-Xavier Coudoux, Patrick Corlay, Michel Kieffer

Multiple description video coding for real-time applications using HEVC 32

Trung Hieu Le, Marc Antonini, Marc Lambert, Karima Alioua

Optimisation du décodage par liste de vidéos corrompues basée sur une architecture CNN 35
Yujing Zhang, Stephane Coulombe, François-Xavier Coudoux, Anthony Trioux, Patrick Corlay

Transformer-Based Image Compression Without Positional Encoding . . 39
Bouzid Arezki, Fangchen Feng, Anissa Mokraoui

Étude comparative des méthodes de prédiction de l'échelle de débit basées sur l'apprentissage pour le streaming vidéo adaptatif 43
Ahmed Telili, Wassim Hamidouche, Sid Ahmed Fezza, Luce Morin

Exploring Temporal Consistency in Image-Based Rendering for Immersive Video Transmission 44
Smitha Lingadahalli Ravi, Félix Henry, Luce Morin, Matthieu Gendrin

3 Sécurité **45**

Analyse d'images secrètes bruitées 46
Erwan Reinders, Bianca Jansen van Rensburg, Pauline Puteaux, William Puech

Méthode Jointe de Tatouage et Compression Draco pour les Objets 3D . 50
Bianca Jansen van Rensburg, Adrian Bors, William Puech

Stéganographie robuste et sans erreur dans des images JPEG en utilisant les sorties des codeurs JPEG 54
Jan Butora, Pauline Puteaux, Patrick Bas

4 Applications **55**

Détection d'anomalies dans des vidéos acquises par drone pour la maintenance préventive de lignes électriques 56
Guillaume Fourret, Gérard Subsol, Christophe Fiorio, Marc Chaumont, Samuel Brau

Vers un outil d'inspection temps réel de l'état d'avancement d'un chantier de construction par RA 60
Mathis Baubriaud, Stephane Derrode, Rene Chalon, K. Kernn

5 Session posters **65**

Actual Fabric Digitalization 66
Thu Ha Do, Minh Chau Huynh, Xuyuan Tao, Pascal Bruniaux, Ludovic Koehl, Kim Phuc Tran, Xianyi Zeng

Contribution des signaux résiduels pour la détection de la permutation de visages dans les vidéos hypertruquées 70
Paul Tesse, Christophe Charrier, Emmanuel Giguet

Evaluation de la qualité sans référence des nuages de points basée sur les statistiques de co-occurrence 3D 74
Souheib Riache, Mohamed-Chaker Larabi

Transformer multimodal pour la détection du stress 78
Kevin Feghoul, Deise Santana Maia, Mohamed Daoudi, Ali Amad

X-RCRNet: an explainable deep learning network for COVID-19 detection using ECG beat signals 82
Marc Junior Nkengue, Xianyi Zeng, Ludovic Koehl, Xuyuan Tao

6 Liste des auteurs **87**

Image, vidéo et géométrie

Annotation semi-automatique de base de données d'images complexes non standardisées

Guillaume PICAUD[1,3] Marc CHAUMONT[1,2] Gérard SUBSOL[1] Luc TEOT[3]

[1]Equipe ICAR, LIRMM, Univ. Montpellier, CNRS, Montpellier, France

[2]Univ. Nîmes France

[3]Cicat-Occitanie, Montpellier, France

{ guillaume.picaud, marc.chaumont, gerard.subsol}@lirmm.fr

l-teot@chu-montpellier.fr

Résumé

Les plaies chroniques représentent un enjeu sanitaire mondial affectant lourdement le quotidien des patients et représentant un coût important pour les systèmes de santé. La prise en charge des plaies chroniques repose en partie sur une analyse visuelle, ce qui motive l'utilisation d'algorithme d'analyse automatique. Des bases de données d'images spécialement dédiées à l'entraînement d'algorithme Deep Learning ont été assemblées en respectant des conditions d'acquisitions rigoureuses. Cependant, une telle standardisation des images ne reflète pas le quotidien des soignants car la plupart des soins sont réalisés au domicile du patient et les infirmiers n'ont que leur smartphone pour prendre des photos. Le réseau Cicat-Occitanie est un réseau de conseil destiné à aider les équipes soignantes dans la prise en charge des patients porteurs de plaies complexes. Avec le temps, le réseau a constitué une base de données de plus 130 000 images de plaies chroniques. Bien que ces images soient décrites d'un point de vue médical, aucune annotation par boîte englobante n'est disponible. Or, l'annotation représente une charge de travail importante. Cet article propose une méthode d'annotation semi-automatique d'une base de données non labellisée via l'entraînement itératif d'un algorithme de détection. Elle vise à réduire la charge de travail de l'annotateur.

Mots clefs

Plaies chroniques, Deep Learning, détection d'objet, annotation semi-automatique

1 Introduction

Selon la Haute Autorité de Santé, une plaie est considérée comme chronique après 4 à 6 semaines d'évolution. De multiples facteurs peuvent favoriser leurs apparitions au sein de populations à risque comme les personnes âgées, les diabétiques ainsi que les personnes à mobilité réduite. Les plaies chroniques représentent un problème majeur de santé publique. Elles présentent des complications comme l'amputation voire le décès. Elles ont aussi un coût pour la société. L'assurance maladie a estimé à plus d'un milliard d'euros la seule gestion des escarres et ulcères à domicile pour l'année 2011. Leur prévalence est en hausse [1] du fait de la croissance démographique mondiale ainsi que du vieillissement des populations occidentales.

La richesse de la littérature dédiée aux plaies chroniques témoigne de l'intérêt croissant des chercheurs, notamment dans le domaine du Deep Learning [2]. Quelques bases de données sont aujourd'hui accessibles comme le DFUC2020 [3] sur la détection, DFUC2021 [4] sur la classification, HealTech [5] également utilisable sur le thème de la classification ou encore FUseg [6] concernant la segmentation. Ici, les conditions d'acquisition sont standardisées, facilitant la convergence des algorithmes de Deep Learning. Pour répondre au besoin des équipes soignantes, il existe donc un réel besoin de concevoir des bases de données médicales annotées, moins standardisées et de proposer des IA plus robustes face aux variations des conditions d'acquisition.

Le « Réseau Cicat-Occitanie » [1] déploie sur l'ensemble de la région éponyme une aide destinée aux équipes de premier recours (médecins et infirmiers) dans le but d'améliorer la prise en charge des patients. Cette aide porte notamment sur l'analyse des plaies chroniques par des experts à distance au cours d'une téléconsultation [7]. Ainsi, près de 19 000 patients ont été suivis par le réseau, constituant une base de données structurée et significative contenant plus de 133 000 images rattachées aux dossiers médicaux archivés. La diversité des équipes et du matériel, le manque de temps ainsi que les conditions au domicile du patient rendent difficile l'acquisition d'images standardisées. La grande variabilité des images du point de l'éclairage, la distance, l'angle, le zoom, le flou, etc... complexifie leur analyse automatisée. Des exemples d'images issues de la base de données Cicat-Occitanie sont consultables à la figure 1.

1. https ://www.cicat-occitanie.org/a-propos-1

En l'état, la base d'image du Cicat-Occitanie ne possède aucune annotation par boîte englobante. L'annotation de base de données par un opérateur coûte cher en temps et en ressources. Une automatisation de ce travail par des algorithmes de Deep Learning serait utile mais ne garantit pas l'absence d'erreurs d'annotation. Nous trouvons dans la littérature des travaux dédiés à la réduction du temps de travail de l'annotateur. Une solution étudiée dans [8] et [9] est fondée sur l'Active Learning afin que l'opérateur humain n'ait qu'une fraction de la base de données à annoter. Ces deux travaux ainsi que [10] ont aussi démontré l'efficacité d'une interface ergonomique dédiée à l'annotation afin de réduire le temps de travail de l'opérateur. Dans cet article, nous proposons une méthode d'annotation semi-automatique, incrémentale, dont le but est de réduire le temps de travail de l'annotateur tout en garantissant la qualité des annotations. Cette méthode permet de vérifier manuellement l'ensemble de la base de données en s'appuyant sur des outils open source faciles à utiliser.

FIGURE 1 – *Exemples d'images issues de la base de données Cicat-Occitanie. Certaines images sont de bonne qualité tandis que certaines sont plus difficiles à analyser par le flou ou l'occultation de la plaie.*

2 Annotation supervisée incrémentale

Dans cette section, nous présentons notre méthode incrémentale pour annoter la base de données Cicat-Occitanie de manière semi-automatique à l'aide de boîtes englobantes. Nous proposons tout d'abord une vue d'ensemble du protocole d'annotation avant de présenter en détail les différentes étapes.

2.1 Vue d'ensemble

La figure 2 présente le processus itératif global d'annotation semi-automatique d'une base de données B par boîtes englobantes. Il consiste à entraîner incrémentalement un algorithme de détection R_i à partir d'une sous-partie de B correctement annotée B_i^*. On obtient une version R_{i+1}. On peut alors réaliser une inférence sur le lot L_i d'images non annotées issu de B. Les prédictions $L_i^{'}$ sont alors vérifiées, validées ou corrigées par l'annotateur. Le lot corrigé L_i^* est intégré dans une nouvelle version de la base de données d'entraînement B_{i+1}^*.
L'initialisation du processus nécessite de choisir une base de données de référence B_0^* différente de B et qui ne sera pas utilisée par la suite pour la constitution des B_i pour tout i>0. Il faut également choisir un algorithme de détection pré-entraîné R_0 puis l'entraîner sur B_0^* donnant ainsi la version R_1 de l'algorithme. Nous l'utilisons alors pour prédire les boîtes englobantes d'un lot d'images L_0, issu de la base de données à annoter. Ce lot d'images désormais doté de boîtes englobantes à vérifier se nomme $L_0^{'}$. Les prédictions sont acceptées, corrigées ou supprimées par l'annotateur et le lot d'images corrigé se nomme alors L_0^*.

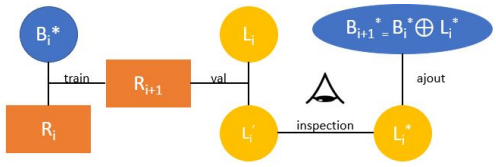


FIGURE 2 – *Schéma du processus itératif d'annotation semi-automatique. Nous entraînons le réseau R_i sur la base de données d'entraînement annotée B_i^*. Nous utilisons R_{i+1}, le résultat de l'entraînement, sur le lot d'images L_i afin d'obtenir les prédictions $L_i^{'}$. L'opérateur vérifie, corrige ou supprime produisant ainsi L_i^* qui sera utilisé pour construire B_{i+1}^*.*

2.2 Présentation et préparation des bases de données

Le Diabetic Foot Ulcer Challenge est une compétition annuelle[2] pour la promotion et le développement d'algorithmes capables d'analyser des photos de plaies du pied diabétique. En 2020, le thème de la compétition était la détection par boîtes englobantes [3]. La base de données mise à disposition jouit de conditions d'acquisition rigoureuses : un nombre restreint d'experts a utilisé un nombre restreint d'appareils photos dans un nombre restreint de salles de consultations garantissant ainsi l'homogénéité de la base de données. 2000 images annotées composent la base d'entraînement et seront utilisées dans l'expérience en tant que B_0^*.
La base de données du Cicat-Occitanie rassemble elle plus de 133 000 images de plaies chroniques de tout type (escarre, ulcère, plaie du pied diabétique, etc…). Un premier filtrage de la base de données est réalisé afin de standardiser les images aux dimensions 450x600 en excluant celles dont le ratio d'origine est trop éloigné de la valeur 1,33. B est composé de 9000 images extraites de la base de données Cicat-Occitanie par tirage aléatoire sans remise. Nous

2. https ://dfu-challenge.github.io/

l'avons divisé en 3 lots de tailles identiques, L_0, L_1 et L_2. En plus, un dernier lot de 538 images, L_{val}, a été conçu. Il est annoté manuellement à l'aide du logiciel opensource LabelImg [3].

3 Expériences

3.1 Implémentation

Comme algorithme de détection, nous avons choisi Yolov5 développé par Ultralytics [11]. Ce choix est motivé par sa simplicité de prise en main et par sa capacité à détecter des objets de petites comme de grandes tailles. Nous disposons pour nos entraînements d'une carte graphique GeForce GTX TITAN X 12Go ce qui nous permet d'utiliser la version Yolov5x à 86 millions de paramètres. R_0 correspond à Yolov5x pré-entraîné sur la base de données ImageNet avec le jeu d'hyperparamètres proposé par Ultralytics pour l'entraînement sur la base de données VOC. Tous les entraînements sont réalisés avec des mini-batchs de 8 images durant 60 epochs. Nous initialisons la base de données B_0^* avec celle du DFUC2020.

3.2 Métriques d'évaluation

L'algorithme de détection propose quatre coordonnées servant à localiser la boîte englobante ainsi qu'une valeur appelée score de confiance qui quantifie la confiance de l'algorithme dans la prédiction de chaque boîte.
Les vrais positifs T_P, les faux positifs F_P ainsi que les faux négatifs F_N sont déterminés par rapport à une valeur seuil du calcul de l'Intersection over Union, dont l'équation est établie ci-dessous, entre les boîtes englobantes issues de la prédiction de l'algorithme et celle correspondant à la vérité terrain. Si l'IoU entre la prédiction et la vérité terrain dépasse le seuil de 0.5 alors la prédiction sera comptée comme T_P, sinon elle sera comptée comme F_P. Si la boîte englobante vérité terrain d'une plaie n'est associée à aucune prédiction T_P alors nous compterons un faux négatif F_N. Nous sommes ainsi en mesure de calculer les métriques Précision et Rappel, dont les abréviations sont respectivement P et R, à partir de toutes les boîtes englobantes prédites.

$$IoU = \frac{Vt \cap Pred}{Vt \cup Pred} \quad P = \frac{T_P}{T_P + F_P} \quad R = \frac{T_P}{T_P + F_N}$$

La Précision moyenne à 50 ou *l'*AP_{50} est mesurée en fixant le seuil d'IoU à 50% puis en faisant varier le seuil du score de confiance. Ainsi, nous obtenons pour chaque seuil un couple (Précision, Rappel) nous permettant de construire la courbe précision sur rappel. L'AP_{50} correspond à la mesure de l'aire sous cette courbe.
La Précision moyenne entre 0.5 et 0.95 ou $AP_{[0.5:0.95]}$ est mesurée en fixant cette fois-ci le seuil du score de confiance à 0.01 puis en faisant varier le seuil d'IoU entre 0.5 et 0.95 par pas de 0.05. Nous pouvons alors construire la courbe précision sur rappel et mesurer son aire.

3. https ://github.com/tzutalin/labelImg

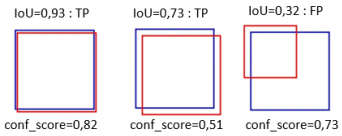


FIGURE 3 – *Illustration du calcul de l'IoU et la prédiction ainsi que du score de confiance prédit par l'algorithme.*

3.3 Performances du processus d'annotation semi-automatique

La version R_1 a été évaluée sur les 2000 images composant la base de données test du DFUC2020 donc distincte de B_0^*. On obtient des performances de 0.64 en mAP et de 0.69 en $F1_{score}$. Le tableau 1 présente les performances des versions de l'algorithme, au fil des itérations, évaluées sur L_{val}.

B_i^*	P	R	AP_{50}	$AP_{0.5:0.95}$
B_0^*=DFUC2020	0.93	0.84	0.88	0.62
B_1^*=L_1^* (2377)	0.92	0.88	0.94	0.66
B_2^*=$B_1^* \bigoplus L_2^*$ (4608)	0.92	0.89	0.94	0.66
B_3^*=$B_2^* \bigoplus L_3^*$ (6894)	0.91	0.88	0.94	0.68

TABLEAU 1 – *Evolution des performances des versions de l'algorithme de détection sur le lot de validation L_{val} en fonction de la base de données d'entraînement B_i^*.*

Le tableau 2 présente l'évaluation des performances de l'inférence $L_i^{'}$ du détecteur R_{i+1} par rapport à l'ensemble d'images acceptées ou corrigées par l'annotateur L_i^*.

L_i^*	P	R	AP_{50}	$AP_{0.50:0.95}$
DFUC2020	0.94	0.86	0.93	0.86
L_1^*(2377)	0.94	0.93	0.97	0.92
L_2^*(2231)	0.94	0.94	0.97	0.93
L_3^*(2286)	0.94	0.93	0.98	0.93

TABLEAU 2 – *Performances des prédictions proposées par l'algorithme de détection R_{i+1} sur les lots L_i au cours des itérations.*

Le tableau 3 illustre l'impact de l'algorithme de détection semi-automatique sur le temps de travail de l'annotateur afin d'inspecter un lot.

R_i	Lot	t_{total}	t_{moyen}/image
aucun	L_{val} (538)	49m	5.5s
R_1	$L_1^{'}$ (3000)	2h05m	2.5s
R_2	$L_2^{'}$ (3000)	1h44m	2.1s
R_2	$L_3^{'}$ (3000)	1h33m	1.9s

TABLEAU 3 – *Evaluation du temps de travail nécessaire à l'annotateur pour vérifier un lot en fonction de la version de l'algorithme réalisant la pré-annotation.*

3.4 Discussion

A l'aide du tableau 1, nous remarquons que chaque itération semble conserver voire améliorer légèrement les prédictions de l'algorithme sur L_{val}.
Le tableau 2 retranscrit le comportement de l'annotateur face aux propositions de l'algorithme de détection. La précision semble être maintenue au cours des itérations. Le nombre de F_P ne semble pas se réduire malgré l'augmentation itérative de la base de données d'entraînement B_i^*. Cependant, le rappel augmente ce qui signifie que l'apprentissage itératif permet de réduire le nombre de F_N. On remarque dans la figure 4 que l'algorithme réalise des erreurs comme la détection de certains motifs dans l'arrière-plan ou la confusion de zones anatomiques.

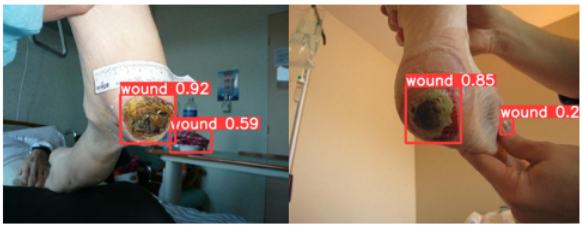


FIGURE 4 – *Confusion de détection de plaies avec un motif en arrière-plan ou une autre zone anatomique.*

Le tableau 3 permet de mesurer l'impact sur le temps de travail de l'annotateur. Dès la première itération, ce temps est divisé par 2 par rapport à une annotation purement manuelle. Cela montre que la plupart des prédictions de l'algorithme sont acceptées par l'annotateur. De plus, la valeur t_{moyen}/image continue de diminuer légèrement au cours des itérations. Pour autant, le temps de lecture de l'image par l'annotateur reste incompressible.

4 Conclusion et perspectives

Dans cet article, nous avons proposé une méthode incrémentale d'annotation semi-automatique par boîte englobante permettant de diviser par 2 le temps de travail d'un annotateur. Cette méthode a été mise en application sur une base de données d'images complexes n'ayant pas d'acquisition standardisée. L'initialisation de la base de données et de l'algorithme peuvent être réalisées à partir de ressources publiques. Les itérations permettent ensuite d'adapter plus finement l'algorithme à la base de données cible.
La compétition DFUC2020 [4] a montré que certaines modifications architecturelles permettent de dépasser les performances de Yolov5 comme notamment la convolution déformable [12, 13] dont l'application dans l'architecture Faster R-CNN a atteint la première place de la compétition. Ainsi, il serait intéressant d'améliorer le processus d'annotation du Cicat-Occitanie en appliquant ce processus itératif sur de nouvelles architectures de détection.

4. https ://github.com/ryohachiuma/DFU-challenge

Remerciements

Nous souhaitons remercier l'Association Nationale de la Recherche et de la Technologie ainsi que le réseau Cicat-Occitanie pour financer et soutenir la thèse au travers d'une Convention industrielle de formation par la recherche.

Références

[1] L. Martinengo et al. Prevalence of Chronic Wounds in The General Population : Systematic Review and Meta-Analysis of Observational Studies. *Annals of epidemiology*, (29) :8–15, Janvier 2019.

[2] R. Zhang et al. A Survey of Wound Image Analysis Using Deep Learning : Classification, Detection, and Segmentation. *IEEE Access*, 10 : 79502–79515, 2022.

[3] B. Cassidy et al. The DFUC 2020 Dataset : Analysis Towards Diabetic Foot Ulcer Detection. *touchREV Endocrinol*, (17) :5–11, 2021 .

[4] M. Hoon Yap et al. Analysis Towards Classification of Infection and Ischaemia of Diabetic Foot Ulcers. Dans *IEEE EMBS International Conference on Biomedical and Health Informatics BHI*, Juillet 2021.

[5] O. Subba Reddy et al. HealTech-A System for Predicting Patient Hospitalization Risk and Wound Progression in Old Patients. Dans *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2463–2472, 2021.

[6] C. Wang et al. FUSeg : The Foot Ulcer Segmentation Challenge. Dans *arXiv* , 2022.

[7] L. Téot et al. Complex Wound Healing Outcomes for Outpatients Receiving Care via Telemedicine, Home Health, or Wound Clinic : A Randomized Controlled Trial. *The International Journal of Lower Extremity Wounds*, (19) :197–204, Juin 2020.

[8] K. Gokalp Ince et al. Semi-Automatic Annotation For Visual Object Tracking. Dans *CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE, oct 2021.

[9] D. P. Papadopoulos et al. We don't need no bounding-boxes : Training object class detectors using only human verification, 2017.

[10] A. M. Obeso. Image annotation for Mexican buildings database. Dans *Optics and Photonics for Information Processing X*, volume 9970, pages 201–208. SPIE, 2016.

[11] G. Jocher et al. YOLOv5 by Ultralytics, 5 2020, https ://github.com/ultralytics/yolov5.

[12] D. Jifeng et al. Deformable convolutional networks. Dans *Proceedings of the IEEE International Conference On Computer Vision*, pages 764–773, 2017.

[13] X. Zhu et al. Deformable ConvNets V2 : More Deformable, Better Results. Dans *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9300–9308, 2019.

Détection de contenu explicite dans les vidéos

Hugo Jean Emmanuel Giguet Christophe Rosenberger

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

{hugo.jean, emmanuel.giguet, christophe.rosenberger}@unicaen.fr

Résumé

L'analyse de contenu vidéo joue une rôle majeur dans la protection des enfants sur Internet et dans le travail des forces de l'ordre lors d'investigation numérique. Avec la rapide croissance de la taille des supports de stockage et les plateformes de partage, le besoin de solutions rapides et robustes d'analyse de vidéos devient de plus en plus nécessaire. Dans cet article, nous proposons un modèle de détection de contenu explicite dans les vidéos. Notre approche se base l'utilisation de réseau convolutionnel pour extraire des paramètres visuels de haut niveau et un réseau à mémoire pour exploiter la dimension temporelle de la vidéo. Nous validons notre approche sur un jeu de données composé de vidéos (avec différentes résolutions et durées) démontrant son intérêt opérationnel.

Mots clefs

Analyse de vidéos, apprentissage profond, contenu explicite.

1 Introduction

De nos jours, la diffusion de vidéos représente 80% du trafic sur Internet. Les contenus vidéos sont transmis dans des volumes en croissance exponentielle, la taille des outils de stockage et la réduction de leur prix permettent un stockage quasi illimité de vidéos. Ainsi, un individu peut facilement stocker chez lui plusieurs milliers de vidéos.

Dans un contexte d'enquête criminelle, les supports de stockage de suspects sont en général perquisitionnés pour la recherche de preuves. Les crimes tels que l'exploitation d'enfants et le partage non autorisé de contenu à caractère sexuel ont de plus en plus comme support les vidéos. Cependant, la taille de ces données et leur nature ne permet pas un traitement rapide par les forces de l'ordre, qui doivent souvent exploiter ces données sous la pression d'un délai, typiquement la durée d'une garde a vue. Ces contraintes sont propices à des erreurs humaines (données compromettantes non détectées). Aujourd'hui, les avancées technologiques en matière de traitement automatique des images ne sont plus à présenter, leur utilisation est universelle. Ceci ouvre des opportunités pour le développement de nouvelles méthodes automatiques et robustes.

L'objectif de ce travail est de proposer une méthode de détection du caractère explicite d'une vidéo avec de meilleures performances par rapport à l'état de l'art, à la fois en considérant le taux de reconnaissance et le temps de calcul. L'analyse est globale mais peut être aussi appliquée pour identifier des frames avec du contenu explicite. Cette méthode vise à être utilisée de façon opérationnelle dans des investigations numériques de disques durs. D'autres usages peuvent être envisagés comme la diffusion de vidéos avec contrôle parental ou la vérification automatique du contenu de vidéos lors d'un téléchargement sur un serveur.

Dans cet article, nous proposons une approche basée sur l'utilisation de modèle convolutionnel pour l'extraction de paramètres visuels de haut niveau. Nous utilisons ensuite un réseau avec mémoire pour utiliser la dimension temporelle d'une vidéo. Notre architecture est présentée dans la Figure 2 avec une vue de haut niveau.

Le plan de l'article est le suivant. La section 2 décrit les principales méthodes de la littérature sur la détection du caractère explicite dans les vidéos. La méthode proposée est décrite dans la section 3. La section 4 présente le protocole expérimental et les résultats obtenus. Nous concluons cet article dans la section 5 et définissons plusieurs perspectives à ce travail.

2 Etat de l'art

Le papier [1] proposant le dataset utilisé dans cet article recense plusieurs travaux et donne leurs performances relatives dans le tableau 1. La méthode θ_x consiste à simplement compter le nombre d'images considérées comme explicites et considérer la vidéo comme explicite si cette valeur dépasse un seuil défini. L'approche θ_y réalise la même chose mais avec un pourcentage de la vidéo, par exemple si 10% de toutes les frames sont explicites alors la vidéo est considérée comme explicite elle aussi. Enfin, la méthode θ_z compte le nombre de frames **successives** explicites et ainsi considère la vidéo explicite si par exemple 5 frames successives sont explicites.

Les résultats présentés ici se basent tous sur une analyse séquentielle des frames, soit en analysant toutes les frames ou alors jusqu'à ce qu'un seuil soit atteint (seuil de frames

Modèle	θ_x Compteur d'image	θ_y % d'image	θ_z images successives explicite
Mask R-CNN	85.63%	86.38%	85.63%
YOLOv4	87.25%	87.75%	87.00%
SSD	81.16%	83.48%	82.88%
Cascade Mask R-CNN	84.88%	86.63%	86.13%

TABLEAU 1 – *Performance des méthodes de l'état de l'art sur la base LSPD.*

classifiées comme explicites par exemple). L'utilisation de toutes les frames d'une vidéo permet en effet de prendre une décision globale, cependant des réseaux convolutionnel 3D utilisant un groupe de frames et non plus une seule frame ont prouvé leur utilité et leur robustesse. On peut citer par exemple, X3D [2] ou encore Resnet3D [3]. Ces réseaux utilisent une sélection de frames tirées d'une méthode d'échantillonage définie, par exemple ci dessous l'architecture des réseaux X3D.

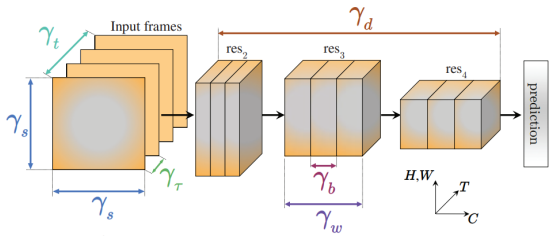


FIGURE 1 – *Base de X3D.*

Ici, nous nous intéresserons à γ_t tout simplement le nombre de frames en entrée du réseau et γ_τ l'écart entre les frames sélectionnées. Ces deux paramètres permettent entre autre d'influencer la taille du réseau et donc de choisir entre précision et rapidité. Le réseau X3D est pensé pour une utilisation sur mobile et est donc par définition rapide de base. Il existe cependant des versions plus complexes pour une utilisation classique.

3 Méthode proposée

L'approche proposée se base sur l'utilisation d'un modèle basé sur un réseau de neurones convolutionnel. Nous présentons tout d'abord le jeu de données utilisé pour l'entraînement du modèle et l'évaluation de performance. Nous détaillons ensuite le modèle ainsi que la stratégie d'analyse.

3.1 Dataset

Nous utilisons ici le dataset LSPD [1], celui ci contient 4000 vidéos (2000 de chaque classe) avec la répartition en durée dans la figure 2. Il est actuellement le dataset disponible au public le plus gros et le plus divers, ils est composé d'une partie image et vidéo, ici nous nous intéressons uniquement à la partie vidéo.

Label	< 1min	<5min	<10min	<20min	>20 min
Explicite	746	745	233	179	106
Normal	986	661	175	125	53

TABLEAU 2 – *Durée des vidéos dans LSPD.*

3.2 Modèle proposé

Le modèle proposé utilise un extracteur de paramètres de haut niveau ici X3D-M pour la vidéo, avec γ_t = 16 et en utilisant non pas une distribution uniforme comme proposé dans le papier original, mais une distribution normale décalée vers la fin de la vidéo. En effet, dans les vidéos, les contenus explicites sont souvent situés au milieu et à la fin. Nous utilisons ensuite un RNN et enfin un classifieur classique. Le modèle est décrit dans la figure 2. Cette version M de X3D fournit en sortie d'extracteur des vecteurs de taille 2048 que nous réduisons à une taille fixe de 512 par soucis de modularité.

Cette architecture permet une utilisation modulaire, on peut facilement ajouter une modalité (typiquement l'audio de la vidéo) et réaliser une fusion de paramètres avant le passage dans le RNN. Cela nous permettra dans le futur de pouvoir réaliser des tests sur l'intérêt et la robustesse de ces méthodes multimodales. Cela nous permet aussi d'inter-changer facilement et rapidement l'extracteur de paramètres, le classifieur et le réseau à mémoire.

4 Évaluation de la performance

Nous définissons dans cette section, le protocole et les résultats obtenus.

4.1 Protocole expérimental

Sur les 4000 vidéos composant le dataset, nous en sélectionnons 80% pour le set d'entraînement, 10% pour le set de validation et enfin 10% pour la base de test. Ainsi, nous avons 3200 vidéos (1600 par classe) d'entraînement, 400 de test et de validation.

L'implémentation de ce modèle est réalisé sous PyTorch, le modèle a été entraîné sur un cluster de 7 cartes Nvidia 1080 Ti avec 11 Go chacune. Cependant, il est important de noter que nous utilisons un cluster pour accélérer l'entraînement et non pas parce que nous sommes bridés par la taille du modèle (comme expliqué plus haut nous utilisons un modèle principalement utilisé pour les téléphones lors de l'inférence).

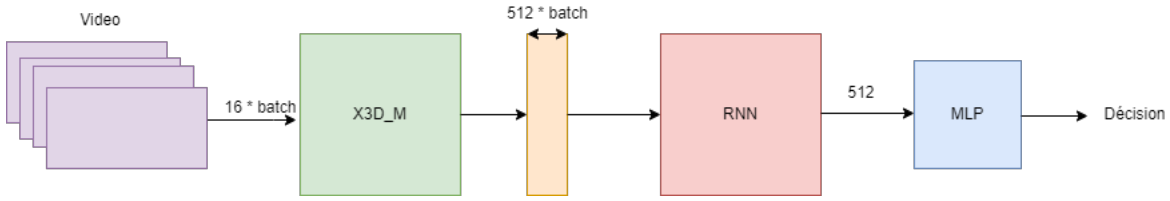


FIGURE 2 – *Architecture du modèle.*

4.2 Résultats

Le tableau 3 présente les résultats obtenus lors de l'évaluation de notre modèle sur notre base de test. Nous avons aussi proposé un modèle sans RNN et directement un classifieur après l'extraction des paramètres.

Head type	Performance
MLP	94%
RNN	96%

TABLEAU 3 – *Résultats sur LSPD.*

Les résultats obtenus sont excellents pour une classification binaire et sont meilleurs que les résultats par utilisation de modèle par frame unique présentés dans le tableau 1.

Lors de l'exécution du modèle pour la phase de test, nous avons calculé le temps d'inférence du modèle pour un paquet de 16 frames (voir le tableau 4). On peut noter un speedup d'environ 50 lors du passage sur GPU, ainsi l'utilisation d'un GPU est grandement favorable et reste assez facile d'accès de part la taille du modèle d'environ 16 MB en mémoire lors de l'inférence.

Processeur	Temps moyen d'inférence
CPU	907
GPU	20

TABLEAU 4 – *Temps d'inférence moyen (en ms).*

Nous présentons dans la figure 3 une illustration de l'analyse d'une vidéo en appliquant le modèle localement pour identifier les séquences avec un contenu explicite. Dans une investigation numérique opérationnelle, l'expert pourra définir un seuil et visualiser des résumés vidéos dans les séquences identifiées par la méthode proposée. Il pourra plus facilement décider si le contenu est répréhensible ou non.

5 Conclusion et perspectives

La méthode proposée permet une reconnaissance efficace et rapide du caractère explicite d'une vidéo facilitant grandement le travail d'investigation numérique sur un disque dur par exemple. L'approche proposée permet aussi d'identifier des séquences problématiques au regard de son contenu explicite. Des applications de l'approche peuvent également concerner le contrôle parental en masquant des frames inadéquates pour des enfants.

Dans cet article nous avons abordé la sélection de frames. Dans le papier où l'architecture X3D est proposé [2], la méthode tire uniformément γ_t frames avec un écart minimum de γ_τ. Pour notre entraînement, nous avons utilisé une distribution normale centrée sur la deuxième partie de la vidéo. Une des principales pistes de recherche est de pouvoir extraire une distribution réelle du contenu explicite dans les vidéos. On pourra par exemple utiliser les vidéos dont la durée est supérieure à 10 minutes et en analyser celles-ci frame par frame pour en tirer une distribution normalisée sur la durée de la vidéo. On pourra ensuite utiliser cette distribution en tant que méthode d'échantillonage de nos vidéos pour l'apprentissage. Cela devrait nous permettre d'accélérer l'apprentissage et d'améliorer la robustesse de nos modèles.

Un autre point déjà abordé dans ce papier est l'utilisation d'autre modalité pour notre classification par exemple l'audio qui semble être la modalité la plus facile d'accès. La classification d'audio aujourd'hui s'effectue à partir de spectrogramme de celui-ci, spectogramme ensuite fourni à un CNN pour bénéficier de l'état de l'art de ceux-ci comme par exemple proposé dans [4]. L'ajout de celle-ci pourrait se faire assez simplement due à la construction modulaire du réseau. Il suffirait d'utiliser une concaténation avant la couche RNN 4. On peut aussi mentionner les différentes techniques de fusion de paramètres disponibles pour les modèles multimodales proposées comme par exemple dans [5], ou même encore les méthodes par ensemble qui peuvent être utilisées pour obtenir de meilleurs résultats.

Références

[1] Phan Duy, Thanh Nguyen, Quang Nguyen, Hoang Tran, Ngoc-Khoi Khac, et Lung Vu. Lspd : A large-scale pornographic dataset for detection and classification. *International Journal of Intelligent Engineering and Systems*, 15 :198, 02 2022.

[2] Christoph Feichtenhofer. X3D : expanding architectures for efficient video recognition. *CoRR*, abs/2004.04730, 2020.

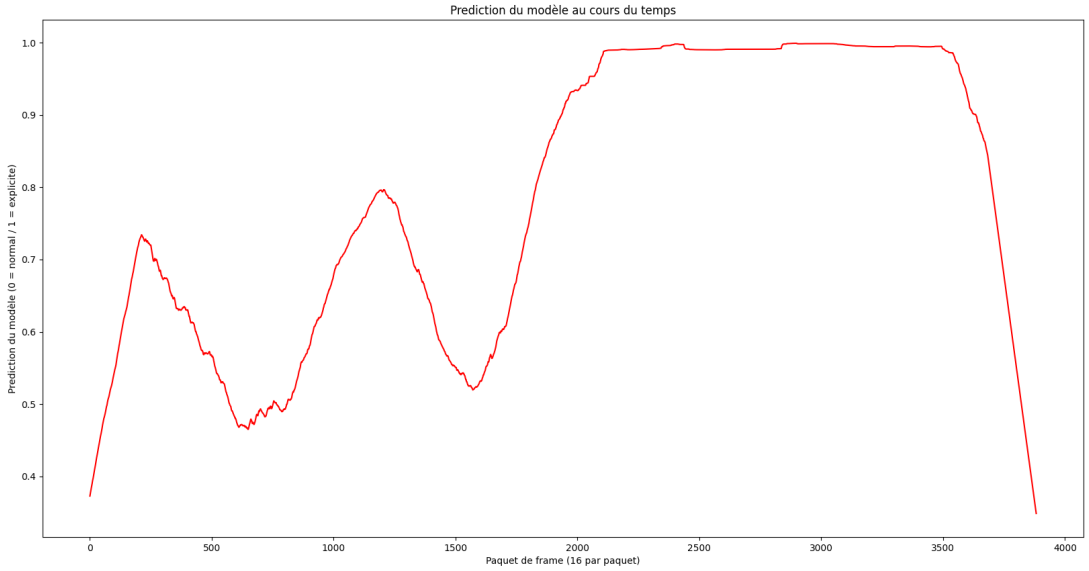


FIGURE 3 – *Illustration d'une analyse des frames d'une vidéo.*

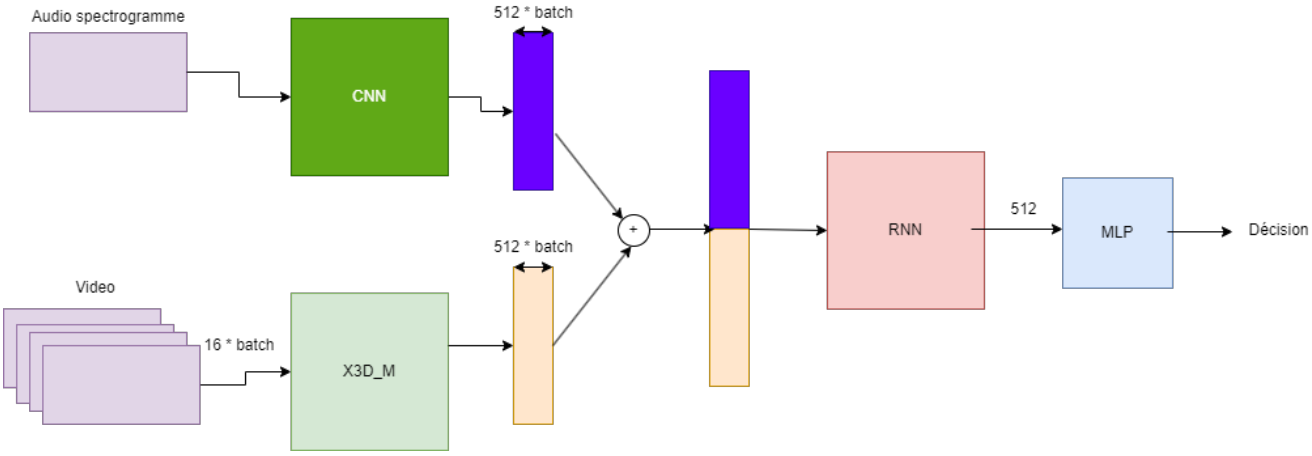


FIGURE 4 – *Architecture du modèle combinant l'audio et la vidéo.*

[3] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, et Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *CoRR*, abs/1711.11248, 2017.

[4] Truc Nguyen et Franz Pernkopf. Lung sound classification using co-tuning and stochastic normalization, 2021.

[5] Konrad Gadzicki, Razieh Khamsehashari, et Christoph Zetzsche. Early vs late fusion in multimodal convolutional neural networks. Dans *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–6, 2020.

Réseau de neurones convolutif pour l'extraction d'attributs de texture à partir d'images multispectrales

Anis Amziane, Olivier Losson, Benjamin Mathon, et Ludovic Macaire
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Résumé

Les caméras multispectrales de type "snapshot" équipées d'une matrice de filtres optiques multispectraux (MSFA) acquièrent instantanément plusieurs bandes spectrales et fournissent une image brute dans laquelle un seul canal est disponible pour chaque pixel. Les caractéristiques de texture sont classiquement extraites d'images entièrement définies qui sont estimées par dématriçage. Cette procédure peut toutefois générer des artefacts spatio-spectraux. En outre, les coûts de calculs de l'extraction d'attributs de texture ainsi que la dimension de ces derniers augmentent avec le nombre de bandes spectrales échantillonnées par les filtres de la caméra. Dans cet article, nous proposons une approche originale basée sur un réseau neuronal convolutif appelé MSFA-Net pour capturer des interactions spatio-spectrales dans les images brutes à coûts de calcul réduits. Les expériences de classification d'images multispectrales et de segmentation d'images acquises en conditions extérieures montrent que l'approche proposée surpasse plusieurs descripteurs de l'état de l'art.

Mots clefs

Imagerie multispectrale, texture, matrice de filtres multispectraux (MSFA), classification, segmentation.

1 Introduction

Les caméras multispectrales intègrent plusieurs filtres optiques, ce qui permet d'observer les surfaces des matériaux dans plusieurs bandes spectrales. Selon le type de filtres qui échantillonnent la lumière incidente (radiance), les images multispectrales peuvent contenir de l'information spectrale associée au domaine du visible (VIS), du proche infrarouge (NIR) et/ou de l'infrarouge. Les dispositifs « multi-shot » [1] produisent une image multispectrale en empilant plusieurs *frames* acquises successivement. À l'inverse, les dispositifs « snapshot » fournissent une image multispectrale à partir d'une seule acquisition [2]. Les caméras snapshot multicapteurs utilisent des prismes dichroïques pour scinder le faisceau de lumière entrant sur plusieurs capteurs selon des plages de longueurs d'onde. Ils sont donc coûteux et ne peuvent échantillonner que quelques bandes spectrales. Les dispositifs snapshot monocapteur intègrent une matrice de filtres optiques multispectraux (MSFA) recouvrant le capteur, comme le filtre de Bayer (CFA) largement utilisé en imagerie couleur, afin d'échantillonner spatialement et spectralement la radiance incidente en fonction de l'emplacement des photo-capteurs. Chaque filtre du MSFA est sensible à une bande spectrale étroite spécifique, de sorte que chaque pixel de l'image *brute* ainsi acquise représente une seule bande. Les autres bandes manquantes sont estimées par dématriçage pour reconstruire l'image multispectrale pleinement définie [3]. Certaines applications (comme l'identification d'adventices en plein champ) nécessitent des signatures spectrales indépendantes de l'éclairage. Pour ce faire, les images de réflectance sont classiquement estimées à partir des images de radiance dématricées, et des attributs de texture en sont extraits. Comme le dématriçage génère des artefacts et augmente les coûts de calcul, certains auteurs proposent de traiter directement les images brutes pour l'estimation de la réflectance [4] ou l'extraction d'attributs [5]. Dans [6], des attributs de texture basés sur les motifs locaux binaires (LBPs) sont directement extraits des images brutes. Dans le même esprit, nous exploitons ici les avantages de l'apprentissage profond et proposons un réseau neuronal convolutif (CNN) qui agit comme un extracteur d'attributs de texture à partir d'images brutes [7].

2 Extraction d'attributs de textures bruts par CNN

2.1 Image brute acquise via un MSFA

Pour classer les images de texture fournies par une caméra mono-capteur échantillonnant B^2 bandes via un MSFA, le dématriçage estime généralement des images pleinement définies (sur B^2 canaux) à partir d'images brutes. De ces images pleinement définies sont extraits des attributs de texture [8]. Cette approche peut se révéler gourmande en temps de calcul et en mémoire, surtout avec des images à haute définition spectrale. De plus, les interactions spatio-spectrales peuvent ne pas être prises en considération de manière efficace, ce qui affaiblit le pouvoir discriminant des attributs extraits.

Pour prendre en compte la corrélation spatio-spectrale, certaines études traitent directement les images brutes [5, 6]. Lorsqu'un descripteur analyse efficacement une image brute, il peut atteindre des performances de classification similaires, voire supérieures, à celles obtenues à partir d'une image pleinement définie, car le dématriçage génère des artefacts susceptibles d'altérer la représentation de la texture. Dans [6], les attributs de texture sont directement

calculés à partir d'images brutes, ce qui évite l'étape de dématriçage et fournit des attributs discriminants. Plus précisément, la méthode analyse une image brute en fonction du motif de base du MSFA et de sa disposition pour construire un descripteur de texture basé sur l'opérateur LBP. En s'inspirant de ces travaux, nous proposons ici une nouvelle architecture CNN adaptée aux images brutes. Les MSFAs utilisés dans ce travail sont définis par la répétition d'un motif de base $B \times B$ qui échantillonne B^2 bandes différentes. Il n'existe pas de consensus concernant la taille du motif de base, et la recherche d'un compromis entre les échantillonnages spatiaux et spectraux reste un problème ouvert difficile [9] qui dépasse le cadre du présent article. Par conséquent, nous suivons les dispositions MSFA de deux caméras snapshot fabriquées par IMEC [10] et opérant dans les domaines VIS ($B = 4$) et NIR ($B = 5$) (voir Fig. 1).

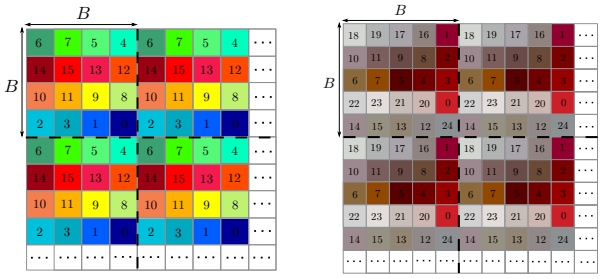


FIGURE 1 – *MSFAs utilisés : IMEC VIS* 4×4 *(*$\lambda^b \in \{469\,nm,\ \ldots,\ 633\,nm\}$*,* $b \in [\![0, 15]\!]$*) (gauche) et NIR* 5×5 *(*$\lambda^b \in \{678\,nm,\ \ldots,\ 960\,nm\}$*,* $b \in [\![0, 24]\!]$*) (droite).*

2.2 Architecture CNN proposée

L'architecture CNN proposée, nommée MSFA-Net, extrait directement des attributs de texture à partir de patchs carrés de l'image brute de taille $X \times X$ pixels, où $X = m \cdot B$ est un multiple de la largeur du motif de base du MSFA. MSFA-Net est composée de trois blocs convolutifs, suivis d'une couche de sous-échantillonnage (« pooling ») qui moyenne les cartes d'attributs et de deux couches entièrement connectées. La première couche de convolution est la plus importante, car elle guide l'extraction d'attributs selon le motif de base du MSFA. Elle utilise 128 noyaux convolutifs $\{H_n\}_{n=0}^{127}$ de taille $B \times B$ et de profondeur 1, avec un pas de B pixels dans les deux dimensions spatiales et sans remplissage par zero (« zero padding »). Un pas de B pixels garantit que chaque coefficient du noyau est toujours associé à la même bande du MSFA pour toutes les convolutions. Cette première couche apprend des interactions spatiales et spectrales entre les valeurs des canaux dans chaque patch brut qui correspond au motif élémentaire du MSFA. La convolution entre un patch brut P^{raw} et un noyau H_n, $n \in [\![0, 127]\!]$, est définie en chaque pixel $(x, y) \in [\![0, m-1]\!]^2$ par :

$$O_n(x,y) = \sum_{i=0}^{B-1} \sum_{j=0}^{B-1} H_n(i,j) \cdot P^{\text{raw}}(B \cdot x + i, B \cdot y + j). \tag{1}$$

Les 128 cartes d'attributs résultantes $\{O_n\}_{n=0}^{127}$, de taille $m \times m$, sont introduites dans le deuxième bloc convolutif qui utilise 256 noyaux de taille 3×3 avec un pas et un remplissage par zero d'un pixel, de sorte que les tailles des cartes d'attributs d'entrée et de sortie soient identiques. Le dernier bloc convolutif utilise 384 noyaux de taille 3×3 avec un pas d'un pixel et sans remplissage par zero. Pour être invariant aux translations spatiales et robuste au bruit, les cartes d'attributs de la dernière couche convolutive sont injectées dans une couche de sous-échantillonnage. cette dernière moyenne les cartes d'attributs canal par canal. Afin d'introduire une non-linéarité et de réduire la dimension des attributs, le vecteur d'attributs de dimension 384 est injecté dans une couche entièrement connectée qui fournit le vecteur de texture final de taille 128.

3 Évaluation expérimentale

3.1 Description

Nous évaluons notre approche de classification et de segmentation d'images multispectrales à l'aide de deux bases : HyTexiLa [11], formée de 112 images pour la classification de textures, et une base de 96 images dont nous disposons, dédiée à la reconnaissance de cultures et adventices (« mauvaises herbes »).

Nous comparons notre descripteur avec ceux de l'état de l'art, à savoir trois descripteurs basés sur un apprentissage profond et quatre autres basés sur l'opérateur LBP. Nous adaptons d'abord à nos images le modèle SegNet-Basic (version simplifiée de SegNet [12]) en retenant uniquement l'encodeur, complété d'une couche de vectorisation d'attributs et deux couches entièrement connectées pour obtenir un vecteur de 512 attributs. Nous testons également le modèle S-CNN [13], composé de trois couches convolutives et de deux couches entièrement connectées, dont la première fournit un vecteur de 1024 attributs. Enfin, nous considérons l'extraction d'attributs par apprentissage résiduel profond [14] grâce à l'architecture à 18 couches (ResNet18), qui fournit un vecteur de 512 attributs. Pour les descripteurs basés sur l'opérateur LBP, nous considérons le LBP marginal (comme descripteur de base) [6], les motifs angulaires locaux (LAP) [15], ainsi que les descripteurs LBP-LCC [16] et M-LBP [6].

3.2 Résultats et discussions

Le tableau 1 montre les résultats de classification obtenus par chaque attribut avec, comme classifieur, le plus proche voisin (1-ppv) couplé avec la distance euclidienne. Parmi les descripteurs non basés sur un apprentissage, M-LBP est plus performant que les autres descripteurs basés sur l'opérateur LBP, car il prend en compte la corrélation

spatio-spectrale au sein de l'image brute et évite l'étape de dématriçage, qui peut affecter la représentation des textures. Globalement, les performances de tous les descripteurs augmentent avec la taille des patchs, en particulier celles du LBP marginal et du LAP, qui sont sensibles au nombre de pixels du patch. Les encodeurs SegNet-Basic, S-CNN et ResNet18 sont peu affectés par la taille des patchs et ne fournissent pas de meilleurs résultats que M-LBP avec le MSFA IMEC 5×5 dans la plupart des cas. L'approche que nous proposons est classée première cinq fois parmi les cas testés, suivie de ResNet18. Cela confirme que les attributs fournis par MSFA-Net sont discriminants malgré leur petite taille. Dans l'ensemble, MSFA-Net fournit des performances meilleures que les autres approches ou comparables à des coûts de calcul nettement inférieurs (voir Fig. 2). Le tableau 1 montre également que les performances de MSFA-Net sont moins affectées par la taille des patchs que celles des descripteurs calculés sans apprentissage, ce qui le rend intéressant pour effectuer des tâches de segmentation.

La Fig. 3 montre les résultats de la segmentation obtenue par MSFA-Net et SegNet-Basic sur deux images tests pour le problème de détection et d'identification des cultures de betteraves et de leurs adventices. Elle montre des performances comparables en matière de détection des adventices entre SegNet-Basic et MSFA-Net. Elle montre également que pour l'identification des betteraves et des adventices, les attributs extraits par MSFA-Net permettent au classifieur de mieux distinguer les betteraves et les feuilles de chénopode.

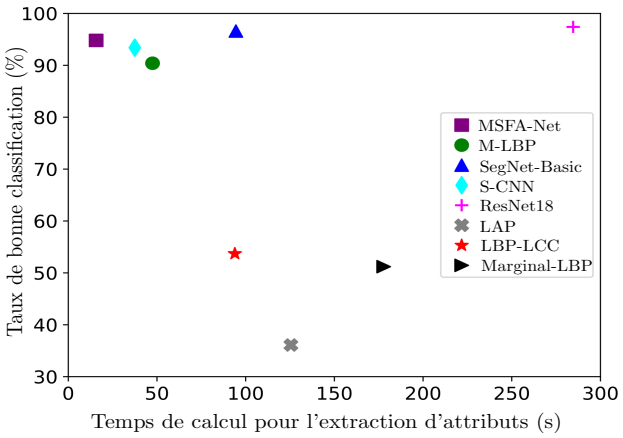


FIGURE 2 – *Taux de bonne classification par 1-ppv vs. temps d'extraction d'attributs pour* $\approx 35,9 \cdot 10^3$ *patchs apprentissage de* 65×65 *pixels (simulés avec le MSFA IMEC* 5×5*) de la base HyTexiLa. Les temps de dématriçage et d'apprentissage des réseaux ne sont pas pris en compte.*

4 Conclusion et perspectives

Dans cet article, nous avons proposé une approche originale pour l'extraction d'attributs de texture à partir d'images brutes grâce à une architecture CNN appelée

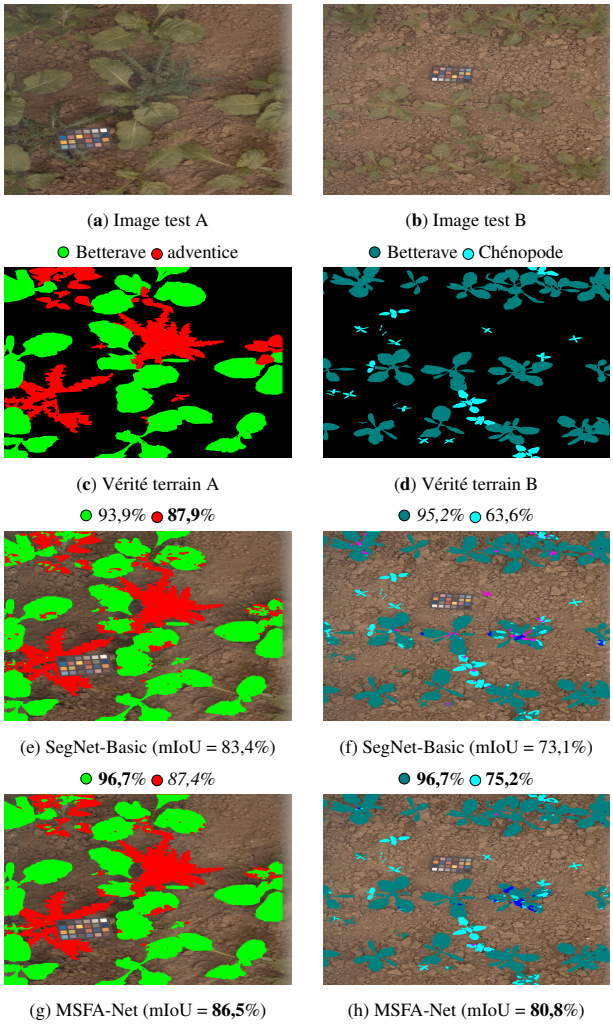


FIGURE 3 – *Résultats de segmentation (Intersection-sur-union moyenne (mIoU) et taux de bonne classification par classe) obtenus par les attributs extraits par SegNet-Basic et MSFA-Net. (a, b) : rendus RGB de deux images multispectrales tests ; (c, d) : vérités terrain ; (e, k) : résultats de détection betterave/adventice ; (f, h) : résultats d'identification betterave/adventice. Les valeurs en gras indiquent les meilleurs résultats. Les couleurs magenta et bleu dans (f, h) correspondent aux pixels chénopode classés comme datura ou chardon, respectivement.*

MSFA-Net. Cette approche évite l'étape de dématriçage qui peut être gourmande en temps de calculs et peut altérer la représentation des textures. Elle nécessite l'apprentissage de beaucoup moins d'hyper-paramètres que les autres architectures CNN testées. Des expériences sur la classification d'images et la segmentation des cultures/adventices montrent que MSFA-Net est globalement plus performante que les autres approches testées, avec des coûts de calcul bien moindres. Les travaux futurs se focaliseront sur la conception d'architectures CNN plus robustes aux perturbations extérieures, liées par exemple à la variation de l'éclairage et aux ombres portées.

TABLEAU 1 – *Taux de bonne classification (%) obtenu par 1-ppv et les attributs extraits à partir d'images brutes ou dématricées, sur la base HyTexiLa. Le meilleur résultat de chaque colonne est affiché en gras, le second meilleur en italique. Le symbole* * *fait référence au MSFA IMEC* 4×4*,* † *à IMEC* 5×5*. L'architecture de ResNet18 utilisée est disponible sur* `https://paperswithcode.com/model/resnet`.

Patch d'entrée	Attribut	taille	IMEC $4 \times 4^*$			IMEC $5 \times 5^\dagger$		
			200×200	124×124	64×64	200×200	125×125	65×65
MSFA	MSFA-Net	$128^{*,\dagger}$	**99,5**	**98,3**	**98,7**	**99,0**	**98,4**	95,1
	M-LBP [6]	$4096^*/6400^\dagger$	*97,2*	*96,9*	*94,6*	96,9	95,4	90,4
Dématricé	SegNet-Basic [12]	$512^{*,\dagger}$	86,2	83,5	86,4	97,1	96,9	*96,6*
	S-CNN [13]	$1024^{*,\dagger}$	82,5	83,6	81,1	94,4	97,4	93,4
	ResNet18	$512^{*,\dagger}$	95,5	88,1	81,7	*98,5*	*97,8*	**97,4**
	LAP [15]	$256^{*,\dagger}$	80,0	69,1	41,3	68,4	65,6	36,1
	LBP-LCC [16]	$512^{*,\dagger}$	87,0	83,8	69,6	70,9	71,4	53,7
	Marginal LBP [6]	$4096^*/6400^\dagger$	81,5	76,4	45,9	77,2	71,6	51,2

Références

[1] Julien Pichette, Wouter Charle, et Andy Lambrechts. Fast and compact internal scanning CMOS-based hyperspectral camera: the Snapscan. Dans *Procs. SPIE*: *Photonic Instrumentation Engineering IV*, volume 10110, pages 1–10, San Francisco, USA, 2017.

[2] Nils Genser, Jürgen Seiler, et André Kaup. Camera array for multi-spectral imaging. *IEEE Transactions on Image Processing*, 29:9234–9249, 2020.

[3] Vishwas Rathi et Puneet Goyal. Generic multispectral demosaicking based on directional interpolation. *IEEE Access*, 10:64715–64728, 2022.

[4] Vlado Kitanovski, Jean-Baptiste Thomas, et Jon Yngve Hardeberg. Reflectance estimation from snapshot multispectral images captured under unknown illumination. Dans *Procs. 29th Color and Imaging Conference*, pages 264–269, Online, 2021.

[5] Wei Zhou, Shengyu Gao, Ling Zhang, et Xin Lou. Histogram of oriented gradients feature extraction from raw Bayer pattern images. *IEEE Transactions on Circuits and Systems II*: *Express Briefs*, 67(5):946–950, 2020.

[6] Sofiane Mihoubi, Olivier Losson, Benjamin Mathon, et Ludovic Macaire. Spatio-spectral binary patterns based on multispectral filter arrays for texture classification. *Journal of the Optical Society of America A*, 35(9):1532–1542, 2018.

[7] Anis Amziane, Olivier Losson, Benjamin Mathon, et Ludovic Macaire. MSFA-Net: a convolutional neural network based on multispectral filter arrays for texture feature extraction. *Pattern Recognition Letters*, 168:93–99, 2023.

[8] Alice Porebski, Mohamed Alimoussa, et Nicolas Vandenbroucke. Comparison of color imaging vs. hyperspectral imaging for texture classification. *Pattern Recognition Letters*, 161:115–121, 2022.

[9] Travis W. Sawyer, Michaela Taylor-Williams, Ran Tao, Ruqiao Xia, Calum Williams, et Sarah E. Bohndiek. Opti-MSFA: a toolbox for generalized design and optimization of multispectral filter arrays. *Optics Express*, 30(5):7591–7611, 2022.

[10] Bert Geelen, Nicolaas Tack, et Andy Lambrechts. A compact snapshot multispectral imager with a monolithically integrated per-pixel filter mosaic. Dans *Procs. SPIE*: *Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VII*, volume 8974, pages 1–8, San Francisco, USA, 2014.

[11] Haris Ahmad Khan, Sofiane Mihoubi, Benjamin Mathon, Jean-Baptiste Thomas, et Jon Yngve Hardeberg. HyTexiLa: high resolution visible and near infrared hyperspectral texture images. *Sensors*, 18(7):2045, 2018.

[12] Vijay Badrinarayanan, Alex Kendall, et Roberto Cipolla. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[13] Vivek Sharma, Ali Diba, Tinne Tuytelaars, et Luc Van Gool. Hyperspectral CNN for image classification & band selection, with application to face recognition. *Technical report KUL/ESAT/PSI/1604, KU Leuven, ESAT, Leuven, Belgique*, 2016.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et Jian Sun. Deep residual learning for image recognition. Dans *Procs. IEEE CVPR*, pages 770–778, Las Vegas, USA, 2016.

[15] Claudio Cusano, Paolo Napoletano, et Raimondo Schettini. Local angular patterns for color texture classification. Dans *Procs. 18th ICIAP Workshops*, volume 9281 de *LNCS*, pages 111–118, Gênes, Italie, 2015.

[16] Claudio Cusano, Paolo Napoletano, et Raimondo Schettini. Combining local binary patterns and local color contrast for texture classification under varying illumination. *Journal of the Optical Society of America A*, 31(7):1453–1461, 2014.

2D versus 3D Convolutional Spiking Neural Networks Trained with Unsupervised STDP for Human Action Recognition

M. El-Assal, P. Tirilly, IM. Bilasco

UMR 9189 – CRIStAL – Centre de Recherche en Informatique, Signal et Automatique de Lille
Univ. Lille, CNRS, Centrale Lille
{mireille.elassal2, pierre.tirilly, marius.bilasco}@univ-lille.fr

originally published at IJCNN 2022
https://doi.org/10.1109/IJCNN55064.2022.9892063

https://arxiv.org/abs/2205.13474

Principal Geodesic Analysis of Merge Trees (and Persistence Diagrams)

M. Pont, J. Vidal, J. Tierny

CNRS, Sorbonne Universite (LIP6)
{mathieu.pont, jules.vidal, julien.tierny}@sorbonne-universite.fr

Published in IEEE TVCG (Transactions on Visualization and Computer Graphics),
Volume : 29, Issue : 2, 01 February 2023.
DOI : 10.1109/TVCG.2022.3215001.

https://ieeexplore.ieee.org/abstract/document/9920234

https://arxiv.org/pdf/2207.10960.pdf

Codage et compression

Analysis of the influence of errors in DNA-based image coding

Jorge Encinas Ramos, Davi Lazzarotto, Michela Testolina, Touradj Ebrahimi
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{jorge.encinasramos, davi.nachtigalllazzarotto, michela.testolina, touradj.ebrahimi}@epfl.ch

Abstract

In the last decade, DNA has been increasingly investigated as an alternative medium for cold data storage, presenting several advantages over standard hard drives such as a higher density, longer lifespan and lower energy consumption. However, such coding methods are limited by biochemical constraints that elevate the probability of errors being added to the coded nucleotides during synthesis, storage, and sequencing. Although such errors can be limited by carefully designing the produced strands, it is unfeasible to avoid them completely. In this paper, we explore the impact of naturally induced errors on the performance of a DNA-based image coding by means of realistic simulations, demonstrating that the quality of the decoded images is severely impacted. We also propose an error correction scheme based on Reed-Solomon codes and Blawat encoding, which successfully removes the produced artifacts.

Keywords

Image compression, DNA-based compression, error correction

1 Introduction

The amount of generated and stored data has been growing at an increasing rate, requiring the construction of more and more data storage centers. Digital data is usually represented as bits having binary values, and the majority of persistent data is stored in magnetic tapes and hard drives. Although current technology has been advancing to increase efficiency, the high rates of storage requirements pose a logistic and environmental challenge due to energy consumption. In this context, DNA has been proposed as a serious candidate for storing rarely accessed data. Contrary to typical binary systems, DNA is suitable to represent information in a quaternary basis through four different nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). DNA coding has, in fact, the advantage of being capable of storing data with much less energy resources while offering much higher density, and if good conservation conditions are met, preservation periods in the order of hundreds of years.

However, this technology still suffers from unique drawbacks and is not yet ready to replace other data storage methods in all scenarios. In practical applications, DNA strands have to be synthesized and stored in low luminosity rooms with controlled temperature. In order to retrieve the original information, the molecules are sequenced in a costly process, making it expensive, time consuming, and impractical for data to be accessed multiple times. Moreover, the entire pipeline cannot be executed without undesirable errors such as deletion, insertion, or substitution of nucleotides. The synthesized strands are subject to biochemical constraints which, if they are not met, can either increase the error rate or produce unstable molecules. Research in the field reveals that, in order to reduce errors, the produced molecules should avoid repeated patterns and homopolymers, i.e. repeated sequences of the same nucleotides, as well as high proportions of C and G nucleotides. Although the real effect of naturally induced errors can only be tested with the real-world implementation of a full pipeline with synthesis and sequencing, a faster setup for testing coding mechanisms can be implemented using error simulators [1] that attempt to reproduce the outcomes of such natural processes.

Figure 1: *Left to right: original image, simulation with in vivo storage model, recovered corrected image*

In spite of all the reported challenges, several implementations of DNA-based coding have been reported in the last few years. In a first attempt to store digital information in DNA, [2] translated 0 and 1 to (A, C) and (T, G), respectively, with the goal of saving a 659-Kbyte digital book. Later, [3] produced the first coding mechanism avoiding the creation of homopolymers by translating data into a ternary basis and using a rotating dictionary to generate nucleotides. In particular, the previously encoded nucleotide was always excluded from the available options to represent the next symbol. Recent solutions incorporate mechanisms for error correction by including redundancy in the binary symbols prior to the translation to DNA, allowing to successfully retrieve the original encoded data even af-

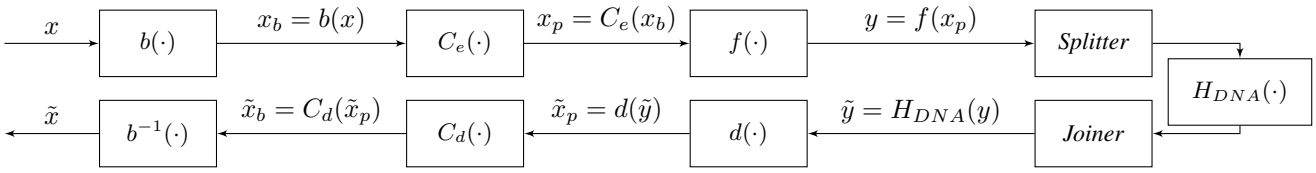


Figure 2: *Proposed workflow diagram*

ter errors being added to the nucleotide sequence, for example by using algorithms such as Reed-Solomon codes [4]. A number of works have also been devoted to developing methods to store image data into DNA. [5] leveraged the DCT from JPEG 1 and used Goldman with Huffman encoding to represent the obtained coefficients. Inspired by this solution, [6] proposed a transcoder to translate the coefficients of already compressed JPEG files into DNA. Both solutions assume a lossless DNA channel and do not take into consideration the errors induced by the storage pipeline. Recently, the usage of neural networks for designing effective and robust compression solutions is being explored. As an example, [7] proposed an image compression solution based on a learning-based convolution autoencoder that can be trained to be robust to substitution noise, which was nevertheless not evaluated against insertion or deletion errors. In this paper, the algorithm from [5] is used as a baseline by modeling the DNA channel with an error simulator [1], demonstrating that even a small percentage of errors causes severe degradation on the decoded image. A pipeline for error correction is then proposed and implemented, allowing for complete recovery of the original information from the distorted DNA strands.

2 Proposed error correction pipeline

The workflow proposed in this paper focuses on adapting existing Forward Error Correction (FEC) codes developed for binary usage to DNA applications. In order to achieve such a target, an Error Correction Block is wrapped around a binary pass-through stage, where the error correction code is applied before the information is returned to the DNA domain.

The proposed recoding procedure is shown in Figure 2. Let the unprotected source FASTA x first be converted into binary in a FASTA–binary conversion, labeled as $b(\cdot)$. This mapping may be as simple as a trivial fixed-length 2-bit mapping, where each base is assigned a fixed binary-represented number ranging from 0 to 3. This binary x_b is protected using the FEC code of choice via its encoding function, $C_e(\cdot)$. The protected binary x_p is then converted into a new FASTA $y = f(x_p) = f(C_e(b(x)))$, this time using a constraint-compliant encoding function $f(\cdot)$ that results in a DNA strand compliant with given biochemical constraints. The DNA sequence y is split into oligos of finite length (in our implementation, 200 nucleotides) and sent to the DNA channel H_{DNA}, which denotes all stages, i.e. synthesis, PCR, storage, and sequencing, of the DNA workflow and the subsequent errors introduced by them.

The DNA channel is here modeled using the MESA DNA error simulator [1], which reproduces the effects of all previously mentioned stages in individual oligos. For this reason, no identification methods such as barcodes were implemented to determine the position of oligos in the bitstream, which are concatenated after simulation to produce the FASTA $\tilde{y} = H_{DNA}(y)$. The protected binary equivalent inherently damaged by the DNA channel $\tilde{x}_p$ is then recovered using a conversion opposite to the one used in encoding, labeled as $d(\tilde{y})$. This binary-encoded information is then sent to the correction block $C_d(\cdot)$ resulting in the binary stream $\tilde{x}_b$. Here, if the redundancy available is enough to correct the errors in their binary form, the value of $\tilde{x}_b$ will be equal to x_b. Lastly, an operation inverse of the first mapping $b^{-1}(\cdot)$ is applied to reveal the estimated original FASTA $\tilde{x} = b^{-1}(C_d(d(\tilde{y})))$, which can be decoded with the corresponding coding algorithm.

The suitable error correction method must be selected considering the mappings and the relation between one nucleotide error and its corresponding binary error. In our simulations, the selected FEC code is a Reed-Solomon code in finite field GF(2^8), denoted as RS($255, 225, 31$). The choice for $f(\cdot)$ and $d(\cdot)$ was a simplified single-cluster Blawat encoding scheme [8], where every byte is converted into a 5-nucleotide tuple. Without clusters, it is impossible to locate the positions of added errors in $d(\cdot)$, but the encoding and decoding processes are less complex. The use of Blawat codes ensures that one substitution error alters at most two bits, that the overall GC content is balanced, as well as a maximum homopolymer length of 3 is reached. Moreover, this approach effectively turns insertions and deletions into additional substitution errors, resulting in a need for increased redundancy at the cost of equal processing of errors.

Since the length of the compressed FASTA in nucleotides is not necessarily an integer multiple of the block code size k, a padding sequence is added to the RS symbols. This padding is later scanned in the decoding process correlating it with the decoded sequence to locate and remove it. This way, a FASTA file with any length can be used with any code size without need for truncation.

The proposed method has the advantage that redundancy can be introduced without constraints, as the FEC code can be applied directly to the binary translation of the input FASTA and then split into oligos after re-encoding to DNA. This splitting procedure is a necessary step due to the DNA

synthesis constraints, which limit the synthesizable oligos to a few hundred nucleotides. This provides higher flexibility, as a different and further-optimized error correction algorithm may be used in this scheme, providing an extensive test bed of codes.

The designed scheme also provides effective protection to FASTA-encoded information that did not originally incorporate a correction scheme. This implies that the recoding mechanism ensures that the protected FASTA complies with the DNA-specific channel constraints independently of the source.

Finally, the scheme may offer the possibility of skipping the first encoding step, i.e. conversion from FASTA to binary, directly feeding binary information as input, with the goal of converting and protecting it for storage in DNA as a general purpose protect-encode block, similar to most existing DNA codec implementations.

3 Results and analysis

In the proposed workflow, the two steps allowing for an increase in the number of nucleotides in the output FASTA y when compared to x are the binary-FASTA conversions and the Reed-Solomon codes. Since $b(\cdot)$ and $f(\cdot)$ convert bits to nucleotides at different rates, increasing the amount of nucleotides by a factor of $5/4$, and since $C_e(\cdot)$ produces n symbols for every block of length k, the total ratio between stored nucleotides and source nucleotides equals to:

$$R_{rec} = \frac{5}{4} \cdot \frac{n}{k} = 1.417 \tag{1}$$

In other words, the proposed error correction method induces a nucleotide rate increase of approximately 41.7%. Without any detection of error position, the number of substitution errors that can be corrected on the GF(2^8) symbols from $\tilde{x}_p$ is:

$$\lfloor \frac{n-k}{2} \rfloor = 15 \tag{2}$$

Since all erasure and insertion errors in $\tilde{y}$ are converted into additional substitution errors when applying the Blawat decoding function $d(\cdot)$, it is possible to use the result from 2 to determine the maximum amount of errors added to the DNA channel that can be corrected by the proposed scheme. Considering that each block in x_p is composed of $n = 255$ GF(2^8) symbols, and each symbol is translated into 5 nucleotides by $f(\cdot)$, under the assumption that errors are sparse and uniformly distributed, then the maximum tolerated error rate is equal to $15/(255 * 5) = 1.18\%$.

The reported values depend on the level of redundancy added by the Reed-Solomon codes. Using a higher amount of redundancy symbols would allow the correction of more errors at the expense of increasing the nucleotide rate. As this work does not aim at reporting the optimal level of redundancy, and as DNA technology continues to advance, the appropriate level required may be chosen based on the assumptions about the DNA channel. For example, under assumptions of errorless synthesis and greatly improved sequencing or storage technologies, the number of redundancy symbols can be decreased as necessary to improve the code rate.

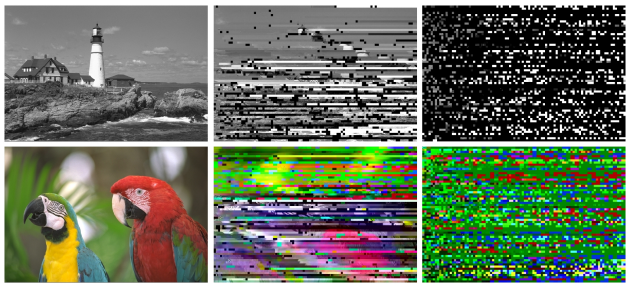


Figure 3: *Left to right: Original image, simulation with in vivo configuration, simulation with in vitro configuration*

The protection scheme described in the previous section was used in conjunction with the JPEG DNA Benchmark Codec [5] as the baseline producing DNA strands from still uncompressed images. Using two images from the Kodak dataset [9] as a test set, the corresponding FASTA files were obtained from the baseline. The error simulator was first applied directly into these sequences without any added protection to evaluate its effect on the decoded image. The results for two images of the test set can be observed in Figure 3.

The original undistorted images are reported in the left column, while the decoded images after error simulation uisng two distinct sets of configuration parameters without protection are reported in the middle and right columns. The different configurations can be obtained by selecting different equipment or technologies used in the DNA-domain workflow such as polymerases, storage hosts (*in vitro* or *in vivo*), or sequencers. The middle column shows the simulation results under a configuration based on an *in vivo* storage model using Escherichia Coli bacteria as the host. This results in a low number of erasures and insertions, and errors are mostly due to substitution. The right-most column contains the decoded results for simulation with *in vitro*-based storage, with error probability set to 0.5%, producing a high number of erasures, and a higher number of errors overall. Results show heavy degradation on the obtained visual result are inflicted because of the simulation. Although using an *in vivo* storage model results in lower loss on information, the final result is anyway wildly different from the original image. These results reveal that successful retrieval of media information without a strong correction scheme is not feasible.

In order to test the efficacy of the proposed error correction scheme, a crop of size 432x432 from test image *kodim-23* was selected and compressed into DNA with the baseline. The generated unprotected FASTA was then encoded using the error correction scheme. Both FASTA files were then served as input to the simulator, using an *in vivo* storage model configuration. Finally, they were both decoded back

Description	MS-SSIM
Unprotected (*in vivo* simulation)	0.0869
Protected (*in vivo* simulation)	0.9937
No simulation	0.9937

Table 1: *MS-SSIM scores for different configurations*

Description	FASTA length [nt]	Nucleotide rate [nt/px]
Unprotected (*in vivo*)	127,248	0.6818
Protected (*in vivo*)	181,050	0.9701

Table 2: *Nucleotide rate for different configurations*

to the image domain. Figure 1 depicts both retrieved of images. Visually, the image recovered from the protected FASTA doesn't present any difference from the original, while the unprotected FASTA produced an image with high visual distortion.

To quantify the obtained results, the MS-SSIM [10] quality metric was computed on a grayscale colormap of the images shown in Figure 1. The results are presented in Table 1.

The objective quality scores show that the distortions introduced by the error simulation have very high impact on the entire workflow, while the objective quality scores with error correction are identical to the image without any error simulation. Therefore, the error correction mechanism allowed to retrieve the same FASTA file as prior to the error simulation, effectively neutralizing the effect of the DNA channel.

Let us consider the bitrate with and without the protection scheme. The results can be seen in Table 2. The redundancy added considerably increases the number of nucleotides per pixel. The FASTA lengths for the protected image are computed before entering the DNA channel, after both encoding steps. This analysis reveals that a bitrate increase of around 42% was obtained, which is inline with the results from Equation 1.

4 Conclusions

In this paper, we presented an efficient pipeline that allows for correction of errors introduced by the synthesis and storage of DNA. Notably, the redundancy level analyzed in this paper allows for robustness against errors with an average bitrate increase, measured in nucleotides per pixels, of approximately 42%. Moreover, the proposed pipeline allows for a simple and flexible adaptation to error correction methods with increased performance. In further work, the optimal level of redundancy in different scenarios can be investigated.

Acknowledgments

The authors would like to acknowledge support from the Swiss National Scientific Research project entitled "Compression of Visual information for Humans and Machines (CoViHM)" under grant number 200020_207918.

References

[1] M. Schwarz, M. Welzel, T. Kabdullayeva, A. Becker, B. Freisleben, et D. Heider. Mesa: automated assessment of synthetic dna fragments and simulation of dna synthesis, storage, sequencing and pcr errors. *Bioinformatics (Oxford, England)*, 36(11):3322–3326, 2020.

[2] George M Church, Yuan Gao, et Sriram Kosuri. Next-generation digital information storage in dna. *Science*, 337(6102):1628–1628, 2012.

[3] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M LeProust, Botond Sipos, et Ewan Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized dna. *nature*, 494(7435):77–80, 2013.

[4] Robert N Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, et Wendelin J Stark. Robust chemical preservation of digital information on dna in silica with error-correcting codes. *Angewandte Chemie International Edition*, 54(8):2552–2555, 2015.

[5] Melpomeni Dimopoulou, Eva Gil San Antonio, et Marc Antonini. A jpeg-based image coding solution for data storage on dna. Dans *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 786–790. IEEE, 2021.

[6] Luka Secilmis, Michela Testolina, Davi Lazzarotto, et Touradj Ebrahimi. Towards effective visual information storage on dna support. Dans *Applications of Digital Image Processing XLV*, volume 12226, pages 29–35. SPIE, 2022.

[7] Xavier Pic et Marc Antonini. Image storage on synthetic dna using autoencoders. *arXiv preprint arXiv:2203.09981*, 2022.

[8] Meinolf Blawat, Klaus Gaedke, Ingo Hütter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W. Pruitt, et George M. Church. Forward error correction for dna data storage. *Procedia Computer Science*, 80:1011–1022, 2016.

[9] Kodak Lossless True Color Image Suite (PhotoCD PCD0992), accessed: 13.03.2023. "`http://r0k.us/graphics/kodak/`".

[10] Zhou Wang, Eero P Simoncelli, et Alan C Bovik. Multiscale structural similarity for image quality assessment. Dans *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

Etude de la faisabilité d'une compensation efficace de la latence par extrapolation des images vidéo

H. KANJ[1] A. TRIOUX[1] M. CAGNAZZO[2] F.X. COUDOUX[1] P. CORLAY[1] M. KIEFFER[3]

[1] UMR 8520 - IEMN, DOAE, Univ. Polytechnique Hauts-de-France, CNRS, Univ. Lille, YNCREA, Centrale Lille, France
[2] LTCI, Telécom ParisTech, Institut Polytechnique de Paris, F-91123 Palaiseau Cedex, & DEI, University of Padova, Italy
[3] Univ. Paris-Saclay, CNRS, CentraleSupélec, L2S, F-91192 Gif-sur-Yvette
{hind.kanj, anthony.trioux, Francois-Xavier.Coudoux, patrick.corlay }@uphf.fr
marco.cagnazzo@telecom-paris.fr, michel.kieffer@l2s.centralesupelec.fr

Résumé

Les applications telles que la télé-conduite et la téléprésence reposant sur des services vidéo doivent garantir une interaction en temps réel avec une qualité d'expérience satisfaisante. La réduction du délai G2G (Glass-to-Glass), c'est à dire le délai entre l'acquisition et l'affichage d'une image vidéo sur un terminal distant, est essentielle pour ces applications. L'extrapolation d'images vidéo basée sur l'apprentissage profond a récemment été considérée pour réduire le délai G2G. Dans cet article, nous examinons l'efficacité de cette technique pour réduire la latence globale dans un système de transmission vidéo point à point. L'objectif est de déterminer le domaine de fonctionnement, les avantages et les inconvénients de cette approche. Pour cela, nous comparons le compromis latence-qualité pour deux méthodes de compensation de latence : la réduction du débit de codage et l'extrapolation. Les résultats montrent que les méthodes d'extrapolation peuvent fournir une réduction significative du délai G2G avec une perte de qualité acceptable, surtout pour les applications avec des contenus vidéo à faible information temporelle.

Mots clefs

Délai Glass-to-Glass, Transmission vidéo à faible latence, Extrapolation d'image, Réduction de débit, Qualité vidéo.

1 Introduction

Ces dernières années, les services vidéo ont été intégrés dans des applications émergentes et interactives telles que la téléprésence [1] ou la téléconduite à distance [2]. Pour garantir une qualité d'expérience satisfaisante (QoE) dans des contextes de téléprésence, ou un comportement sûr d'un système commandé à distance, il faut que les contenus visuels soient fournis à l'opérateur humain (ou à la machine) avec une bonne qualité et une latence réduite.
La latence dans ces applications est déterminée par le délai Glass-to-glass (G2G), c'est à dire le délai entre l'acquisition et l'affichage d'une image vidéo[3] comme illustré à la Fig. 1. Le délai G2G acceptable pour la visioconférence ou les jeux en ligne doit être inférieur à 100 ms pour être en dessous du seuil de perception humaine [4], et pour les applications interagissant avec des machines, le délai est encore plus faible (10-30 ms) [5]. Néanmoins, le délai G2G minimum réalisable (actuellement entre 50 et 400 ms [6]) est limité par les délais d'acquisition, de codage, de transmission, de décodage et de mise en mémoire tampon.
Diverses études ont essayé de réduire chaque source de latence. Pour réduire le délai d'acquisition, on utilise traditionnellement des caméras analogiques, car elles offrent une faible latence à cause de l'absence de mise en mémoire tampon et de traitement des données [7]. Dans le codage vidéo, la configuration Low Delay P (LDP) permet de réduire la latence du codage [8] car elle évite le délai de réorganisation des images. Une autre approche courante consiste à réduire le débit d'encodage [9] pour diminuer la quantité de données transmises par image, et par conséquent, la latence.

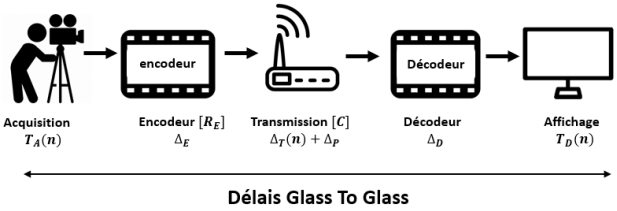


FIGURE 1 – *Latence G2G dans un schéma de transmission vidéo point à point*

Récemment, l'extrapolation des images vidéo a été considérée comme une approche alternative pour réduire le délai G2G et obtenir une latence faible à nulle. L'extrapolation vidéo exploite les techniques d'apprentissage profond en extrayant des caractéristiques profondes des images déjà acquises pour prédire les images futures. Si l'horizon d'extrapolation est suffisamment éloigné, l'image extrapolée peut être transmise à la place de l'image acquise puis affichée au niveau du récepteur, tandis que l'image correspondante est acquise au niveau de l'émetteur, ce qui entraîne une réduction drastique de la latence G2G. La méthode proposée par [10] n'a pas pris en compte l'impact des délais de codage et de transmission, ni le délai de l'extrapolation.
Lorsque la latence est réduite, la qualité de la vidéo reconstruite se dégrade. Par exemple, la réduction du débit cause de forts artefacts de codage, et les images extrapolées

diffèrent des images originales. Dans cet article, nous étudions ce compromis qualité-latence pour des scénarios tels que la téléconduite et la visioconférence. Nous cherchons à étudier la zone d'opération où l'extrapolation d'image est efficace et à identifier ses avantages et ses inconvénients par rapport à la réduction du débit d'encodage. Le reste de l'article est organisé comme suit : La Section 2 décrit le modèle utilisé pour évaluer la latence de G2G du schéma de transmission vidéo. La configuration de la simulation est détaillée et les résultats sont présentés et discutés dans la Section 3. Enfin, la Section 6 conclut ce travail et donne des perspectives.

2 Méthodologie

2.1 Modèle d'estimation du délai G2G

La Fig. 1 décrit le schéma de diffusion vidéo considéré. La vidéo est codée à un débit R_{E} et transmise via une liaison de capacité C, supposée constante et non affectée par R_{E}. L'analyse pourrait être facilement étendue à un canal dont la capacité varie dans le temps.

Les images vidéo sont acquises avec une période Δ_{F}. On suppose que l'acquisition de la n-ième image commence au temps $T_{\mathrm{A}}(n) = n \times \Delta_{\mathrm{F}}$. Le délai d'acquisition et de codage d'une image est supposé constant et égal à Δ_{E}. La taille de l'image encodée est $S(n)$. Une fois encodée, l'image est prête à être mise en paquets et protégée par codage canal. Étant donné qu'une image vidéo codée ne peut être transmise avant que les images précédentes ne soient complètement codées et transmises, l'image n commence à être transmise à l'instant :

$$T_{\mathrm{T}}(n) = \max[T_{\mathrm{A}}(n) + \Delta_{\mathrm{E}}, T_{\mathrm{T}}(n-1) + \Delta_{\mathrm{T}}(n-1)], \quad (1)$$

où $\Delta_{\mathrm{T}}(n-1) = S(n-1)/C$ est le délai de transmission de la $n-1$-ème image codée. Pendant la transmission, la n-ième image se propage dans le canal jusqu'au récepteur durant $\Delta_{\mathrm{P}}(n)$ (ce délai dépend de la distance et de la congestion du réseau). Quand l'image n atteint le récepteur, elle est décodée pendant Δ_{D}, puis affichée au temps $T_{\mathrm{D}}(n)$, par conséquent :

$$T_{\mathrm{D}}(n) = T_{\mathrm{T}}(n) + S(n)/C + \Delta_{\mathrm{P}} + \Delta_{\mathrm{D}}. \quad (2)$$

Le délai G2G est alors la différence entre le moment où une image est affichée au récepteur et le moment où elle a commencé à être acquise à l'émetteur.

$$\Delta_{\mathrm{G}}(n) = T_{\mathrm{D}}(n) - T_{\mathrm{A}}(n). \quad (3)$$

Dans cette étude, afin de respecter les contraintes de faible latence, nous avons utilisé la configuration d'encodage vidéo LDP. La latence doit être évaluée en faisant la moyenne 1) des images I et P, puis 2) des images I uniquement représentant la latence maximale due à leur grande taille. Cela permet de simuler l'effet d'un buffer tampon qui stocke les images vidéo et les lit avec une cadence constante permettant un affichage régulier sans pause.

2.2 Méthode de référence : Réduction du débit de codage

En réduisant le débit d'encodage vidéo, la quantité de données par image vidéo encodée à transmettre diminue, et le temps nécessaire à l'envoi d'une image est réduit. Si le débit d'encodage pour l'image n est réduit de R_{E} à $R'_{\mathrm{E}} = \alpha R_{\mathrm{E}}$, $\alpha \in]0, 1]$, la latence de transmission résultante sera telle que

$$R'_{\mathrm{E}}\Delta_{\mathrm{F}}/C < R_{\mathrm{E}}\Delta_{\mathrm{F}}/C. \quad (4)$$

Supposons qu'une vidéo codée à $R_{\mathrm{E}} = 10$ Mb/s est transmise sur un canal de capacité moyenne $C = 10$ Mb/s à 25 fps, le délai de transmission d'une image est de 40 ms. En considérant un facteur de réduction du débit d'encodage vidéo $\alpha = 1/10$, la taille moyenne de l'image est également divisée par 10 et le délai de transmission devient 4 ms.

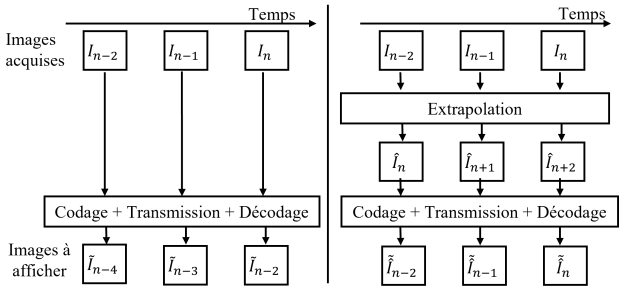


FIGURE 2 – *Illustration de la transmission vidéo sans extrapolation et avec extrapolation. $\hat{I}$ et $\tilde{I}$ indiquent respectivement l'image extrapolée et décodée.*

2.3 Méthode analysée : Extrapolation

L'extrapolation d'image est considérée comme un outil alternatif pour compenser la latence. Elle est applicable à l'encodeur ou au décodeur. Dans cet article, seule l'extrapolation à l'encodeur est considérée sachant que les deux approches ont un comportement similaire en termes de dégradation de la qualité [10].

Pour compenser le délai G2G, un extrapolateur est inclus dans la chaîne de transmission avant l'encodage, *i.e.*, au temps $n \times \Delta_{\mathrm{F}}$ pour la n-ième image. Ce dernier prend k images précédentes $(I_{n-k}, \ldots, I_{n-1})$ comme images de contexte pour produire une estimation de l'image I_{n+h}, où h représente l'horizon temporel d'extrapolation. Par exemple, supposons que le délai G2G soit de $2 \times \Delta_{\mathrm{F}}$, càd lorsque I_n est acquis au moment $T_{\mathrm{A}}(n) = n \times \Delta_{\mathrm{F}}$, l'image affichée au récepteur est I_{n-2}, si l'extrapolation n'est pas introduite (Fig. 2-a). En utilisant un extrapolateur avec un horizon temporel $h = 2$, à $t = (n-2) \times \Delta_{\mathrm{F}}$, I_n est prédite et envoyée au récepteur. Ainsi, l'image affichée au récepteur à $t = n \times \Delta_{\mathrm{F}}$ sera l'image prédite de I_n. Quand l'extrapolateur n'est capable de prédire qu'une seule image à l'avance, l'extrapolation d'horizon h nécessite h itérations d'extrapolation, donc le délai d'extrapolation est $h \times \Delta_{\mathrm{X}}$, avec Δ_{X} est le délai d'extrapolation. Alors, la version extrapolée de l'image I_n est prête à être transmise à :

$$T_{\mathrm{T}}(n) = \max[T_{\mathrm{A}}(n-h) + h\Delta_{\mathrm{X}} + \Delta_{\mathrm{E}}, T_{\mathrm{T}}(n-1) + \Delta_{\mathrm{T}}(n-1)]. \quad (5)$$

3 Configuration de la simulation

Nous proposons de comparer l'efficacité de ces méthodes en évaluant le compromis qualité-latence dans plusieurs scénarios. Nous considérons différentes séquences vidéo : *Stefan*, *Tennis*, et *Touch down pass* à 30 fps, *Soccer*, *Four*

people, *Johnny* à 60 fps de la collection Xiph [11] et *Bike 1*, *Bike 2*, *Road*, *Person* à 10 fps de la base de données Kitti [12]. On prend les 90 premières images de chaque séquence, redimensionnées à 640×448 pixels.
Le codec VTM 18.0 est utilisé pour encoder les images originales et extrapolées avec la configuration LDP [1] et trois débits d'encodage typiques : $R_E = \{1,5$ Mb/s, 800 kb/s, 400 kb/s$\}$ [13]. La taille du groupe d'images (GOP) ou le nombre d'images entre les I-frames successives est fixé à 32 images. Nous considérons des valeurs typiques du délai d'acquisition et d'encodage $\Delta_{\mathrm{E}} = 23$ ms et du délai de décodage $\Delta_{\mathrm{D}} = 5$ ms [2, 3].
Pour éviter l'accumulation de la latence due à l'augmentation du temps de transmission, la capacité du canal est prise supérieure à R_{E}, $C = \{3\,\text{Mb/s}, 6\,\text{Mb/s}, 10\,\text{Mb/s}, 20\,\text{Mb/s}\}$. Le délai de propagation est pris $\Delta_{\mathrm{P}} = 3$ ms qui est une valeur typique dans les réseaux d'accès 5G.
Parmi les techniques d'extrapolation [14], nous considérons le réseau SDC-Net [15], qui présente les meilleures performances [10]. Différentes hypothèses pour le délai d'extrapolation (qui dépend de la plateforme matérielle) sont considérées : $\Delta_{\mathrm{X}} \in \{0, 1/4, 1/2, 3/4\} \times \Delta_{\mathrm{F}}$ pour déterminer quand la méthode d'extrapolation est efficace.
Deux métriques objectives de la qualité vidéo sont prises en compte : le rapport signal à bruit (PSNR), et l'index de similarité structurelle (SSIM).

4 Résultats des simulations

Bien que les images extrapolées soient déformées, la structure et la position de l'objet dans la scène sont en grande partie préservées [10]. Par conséquent, le SSIM est une métrique plus appropriée pour évaluer les artefacts d'extrapolation car il ne repose pas sur des comparaisons pixel à pixel comme le PSNR.
La Fig. 3 montre la variation de PSNR et de SSIM en fonction de la latence moyenne et maximale pour la séquence *Four people* qui représente une vidéo à faible complexité temporelle. Des résultats supplémentaires peuvent être trouvés sur le lien suivant : [2]. Indépendamment de la qualité de départ de la vidéo, celle-ci diminue lorsqu'on utilise l'extrapolation pour atteindre environ 36 dB. Pour toutes les valeurs de R_{E}, l'approche d'extrapolation fournit un gain accru en termes de latence. Pour un canal de faible capacité, même si on gagne la même latence, la qualité obtenue avec la méthode d'extrapolation est meilleure. Par exemple, en considérant $C = 6$ Mb/s, malgré que $\Delta_{\mathrm{X}} = 3/4 \times \Delta_{\mathrm{F}}$, pour obtenir une latence maximale de 37.3 ms avec un horizon d'extrapolation $h = 5$, un PSNR de 36.3 dB et un SSIM de 0.97 sont obtenus. Par contre, l'utilisation de la réduction du débit de codage avec $\alpha = 1/8$ conduit à un PSNR de $33,36$ dB, un SSIM de $0,91$ et un gain de latence de $35,6$ ms. Ainsi, pour une compensation de latence similaire ($35 \sim 37$ ms), nous observons de meilleurs scores de qualité à la fois visuellement (Fig. 4) et objectivement (PSNR et SSIM). Cela montre que la compensation de latence par extrapolation est une méthode viable pour les applications où un contenu vidéo à faible complexité temporelle est transmis.

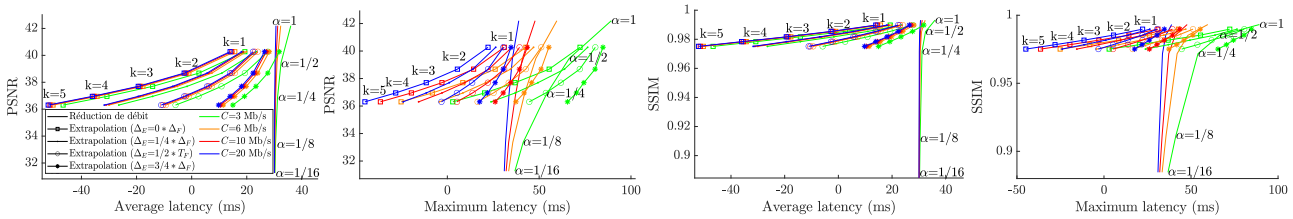
1. encoder_lowdelay_P_vtm.cfg
2. https://drive.google.com/drive/folders/12UJ_Il7yicPklxq1ymioKLj5PgcD06Dz?usp=share_link

5 Discussion

La réduction du débit d'encodage est une technique simple, bien connue et maîtrisée en termes de qualité et d'artefacts générés, offrant une plus grande flexibilité/granularité par rapport à l'extrapolation. En effet, la réduction de la latence par extrapolation dépend de l'horizon temporel h, alors que la réduction du débit d'encodage est contrôlée par le facteur de réduction α. Le choix d'un α approprié permet de réduire la latence de manière plus fine. Néanmoins, cette méthode ne permet de réduire la latence que de quelques ms, au prix d'une baisse significative de la qualité. De plus, un délai G2G nul (ou négatif) n'est pas réalisable avec cette approche car cela nécessiterait de ne pas transmettre de données. D'autre part, l'extrapolation permet de compenser une latence plus importante à condition que le délai d'extrapolation reste faible, e.g., en utilisant un processus non-itératif qui pourrait prédire directement l'image désirée. Cela induirait probablement une perte de qualité supplémentaire, cette idée sera considérée dans un travail futur.
En ce qui concerne la perte de qualité, l'extrapolation semble mieux convenir aux contenus à faible complexité temporelle. Un défi concernant la méthode d'extrapolation est le changement soudain de scène ou de cut, puisque l'image prédite est basée sur les images contextuelles précédentes, l'utilisation d'extrapolation pendant ces cuts conduirait à des informations inexactes. On peut proposer un mécanisme de commutation entre la réduction du débit de codage et l'extrapolation pour éviter un tel problème. Enfin, dans le cas de la multidiffusion, le délai G2G subit par différents récepteurs peut être variable. Lors de l'extrapolation à l'encodeur, l'horizon temporel est décidé au niveau de l'encodeur et, donc, h ne peut être choisi que pour optimiser une certaine mesure de performance globale parmi tous les utilisateurs, e.g, compenser un délai moyen ou le délai maximal. En outre, nous sommes en train d'étendre cette étude basée sur des paramètres statiques (débit binaire de codage, capacité du canal, *etc.*) pour proposer un algorithme de débit de codage adaptatif qui s'adapte aux conditions de fluctuations du canal pour fournir une meilleure qualité d'expérience en tenant compte de l'extrapolation. Les résultats de cette étude approfondie seront présentés lors de la conférence en cas d'acceptation de cet article.

6 Conclusion et perspectives

Deux méthodes de compensation de la latence vidéo sont comparées : la réduction du débit d'encodage et l'extrapolation des images. Cette étude illustre l'efficacité de l'extrapolation en considérant le compromis qualité-latence et détermine sa région d'opération. Les résultats montrent que l'extrapolation est plus performante que la réduction du débit en terme de compensation de latence et peut atteindre une latence G2G nulle ou négative, notamment lors de la transmission de contenus à faible information temporelle.



(a) PSNR de la séquence *Four people*

(b) SSIM de la séquence *Four people*

FIGURE 3 – *Evolution de la qualité en fonction du retard G2G moyen (à gauche) et maximum (à droite) pour $R_E = 800$ kB/s pour la séquence Four people : a) PSNR, b) SSIM.*

(a) Image originale

(b) latence moyenne = 35.6 ms

(c) latence moyenne = 37.3 ms

FIGURE 4 – *Comparaison visuelle de Four people ($R_E = 800$ kb/s, $C = 6$Mb/s) : (a) originale, (b) Réduction de débit ($\alpha = 1/8$, PSNR = 33.3 dB et SSIM = 0.91) et (c) Extrapolation (h=5, $\Delta_X = 3/4 \times \Delta_F$, PSNR = 36.3 dB et SSIM = 0.97).*

Néanmoins, la réduction du délai d'extrapolation est une étape nécessaire lorsque des canaux de faible capacité sont considérés. Actuellement, l'extrapolation est une technique prometteuse, mais elle est encore dans sa phase initiale concernant la qualité et le délai d'extrapolation. Cet article fournit un bon aperçu basé sur des paramètres statiques (capacité du canal, débit d'encodage, *etc.*) et sert de base indispensable à des travaux plus complexes visant à proposer des mécanismes adaptatifs tenant compte la variabilité de ces paramètres. Ces travaux seront présentés lors de la conférence.

Acknowledgments : Ce travail a été financé par le fond national ANR AAPG2020 dans le cadre du projet ZL-LVC (ANR-20-CE25-0014).

Références

[1] Mihir Mody, Pramod Swami, et Pavan Shastry. Ultra-low latency video codec for video conferencing. Dans *2014 IEEE CONECCT*, 2014.

[2] Oussama El Marai et Tarik Taleb. Smooth and low latency video streaming for autonomous cars during handover. *IEEE Netw.*, 34(6), 11. 2020.

[3] Christoph Bachhuber, Eckehard Steinbach, et al. On the minimization of glass-to-glass and glass-to-algorithm delay in video communication. *IEEE Trans. Multimed.*, 20(1), 1. 2018.

[4] Lothar Pantel et Lars C. Wolf. On the impact of delay on real-time multiplayer games. Dans *Proceedings of the 12th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, New York, NY, USA, 2002. Association for Computing Machinery.

[5] Sergiy Melnyk, Abraham Tesfay, et al. Reliable low latency wireless communication enabling industrial mobile control and safety applications. 4. 2018.

[6] Shree Krishna Sharma, Isaac Woungang, et al. Toward tactile internet in beyond 5G Era : Recent advances, current issues, and future directions. *IEEE Access*, 8, 3. 2020.

[7] Sven Ubik et Jiří Pospíšilík. Video camera latency analysis and measurement. *IEEE Trans. Circuits Syst. Video Technol.*, 31(1), 1. 2021.

[8] Soulef Bouaafia, Randa Khemiri, et al. Complexity analysis of new future video coding (fvc) standard technology. *Int. J. Digit. Multimed. Broadcast.*, 2021, 8. 2021.

[9] Ahmed Badr, Ashish Khisti, et al. Perfecting protection for interactive multimedia : A survey of forward error correction for low-delay interactive applications. *IEEE Signal Process. Mag.*, 34(2), 3. 2017.

[10] Melan Vijayaratnam, Marco Cagnazzo, et al. Towards zero-latency video transmission through frame extrapolation. Dans *2022 IEEE ICIP*, 10. 2022.

[11] Xiph.org media, URL https ://media.xiph.org/video/derf/.

[12] Andreas Geiger, Philip Lenz, et al. Vision meets robotics : The kitti dataset. *Int. J. Robot. Res.*, 32(11), 2013.

[13] Conditions for visual comparison of VCB, IVC and WVC codecs, iso/iec jtc1/sc29/wg11 mpeg2013/n13943, 2013.

[14] Qingming Huang, Zhongxiao Li, et al. Video frame prediction with dual-stream deep network emphasizing motions and content details. *Appl. Soft Comput.*, 125, 6. 2022.

[15] Fitsum A. Reda, Guilin Liu, et al. SDC-Net : Video prediction using spatially-displaced convolution. Dans *Computer Vision – ECCV 2018*, Cham, 2018. Springer International Publishing.

Multiple description video coding for real-time applications using HEVC

Trung Hieu Le[1] Marc Antonini[1] Marc Lambert[2] Karima Alioua[2]

[1] Laboratoire I3S - Université Côte d'Azur et CNRS, UMR 7271, Sophia Antipolis, France

[2]Lextan SAS, Gemenos, France

{thle, am}@i3s.unice.fr

Résumé

Remote control vehicles require the transmission of large amounts of data, and video is one of the most important sources for the driver. To ensure reliable video transmission, the encoded video stream is transmitted simultaneously over multiple channels. However, this solution incurs a high transmission cost. To address this issue, it is necessary to use more efficient video encoding methods that can make the video stream robust to noise. Moreover it should have a less complexity to adapt to the real time requirement. In this paper, we propose a low-complexity, low-latency 2-channel Multiple Description Coding (MDC) solution with an adaptive Instantaneous Decoder Refresh (IDR) frame period, which is compatible with the HEVC standard which adaptive redundancy adjustment. This method shows a better resistance to high packet loss rates with lower complexity.

Mots clefs

Multiple Description coding, HEVC, noisy channel, Error Correction, low latency

1 Introduction

Remote driving requires continuous video data transmission to allow the driver to perceive the environment. The vehicle's video is transmitted to the driver through a wireless channel, and the latency and frame quality must be within limits to guarantee safety. However, wireless networks are more susceptible to noise than wired networks and are thus less reliable. Forward Error Correction (FEC) has been implemented in the transport layer to improve the resilience of video transmission to losses, but it is designed to be robust only up to a certain limit of packet error rate and requires high complexity computation[1]. Furthermore, the wireless channel's characteristics are dynamic and vary over time, making it difficult to estimate the channel state accurately for FEC.

To mitigate interference in real-time video transmission, one approach is to use two separate wireless channels to send the same video sequence as a backup. However, this can result in a waste of bandwidth since the same information is being transmitted twice. Multiple Description Coding (MDC) can be used as a solution to address this issue. In the case of MDC for two channels, the MDC encoder produces two different descriptions, S1 and S2, of the video with bit rates R1 and R2, respectively, from the original video source. These two descriptions are then transmitted through two independent channels by two transmitters. If only one description is available at the MDC decoder, either S1 or S2, the side decoder will be used to produce the video sequence with distortion D1 or D2, respectively. However, if both descriptions are available at the MDC decoder, the central decoder merges them to construct the central reconstruction by removing redundant information and retaining the primary one, resulting in higher video quality with a smaller central distortion D0.

MDC has been studied in the past and can be classified into three categories based on how the redundancy information is added to each description. There have been many studies on MDC in the spatial domain, most of which are compatible with the H.264 encoding standard. In [2], Tilo *et* al proposed a method of MDC compatible with the H.264 standard based on adjusting the redundancy level of the different slices. However, this method required to calculate all the GOP to find the optimum bit allocation. In [3], the author used the Multiple-State Video Coding with Redundant Pictures method by shifting the I and P frames of Group of Picture (GOP) between the two descriptions. In [4], the authors proposed an H.264 MDC scheme based on block permutation and DCT coefficients splitting. However, These methods didn't include the redundancy control thus it cannot adapt in the redundancy level based on network condition.

However, recent studies have attempted to include an MDC scheme in the HEVC standard, which has higher performance and is more resilient to packet erasure, as demonstrated in [5]. Many of these studies, such as [6, 7], are based on visual saliency and transform domain splitting. However, these methods may not be suitable for real-time applications due to their use of Bidirectional Motion Compensation (B-Frame). In [8], authors propose a frame rate variation MDC scheme adapted for remote control vehicles, but this scheme does not exploit the temporal dependency between frames, resulting in lower compression performance.

In this article, a spatial domain multiple description codec based on the HEVC standard with Rate Distortion Optimization (RDO) at the CTU level for each frame independently is proposed. The scheme uses unidirectional temporal prediction, which allows for a reduction in transmission costs and meets real-time delay requirements. The proposed solution is presented in the first part of the article, and the second part shows its performance under a packet erasure channel.

2 Proposed Method

We propose a balanced MDC with two descriptions as described in Figure 1. Assume that the packet error distribution of both channels is independently and identically distributed (i.i.d). Therefore, for each frame of the sequence, the expected distortion at the decoder is expressed as :

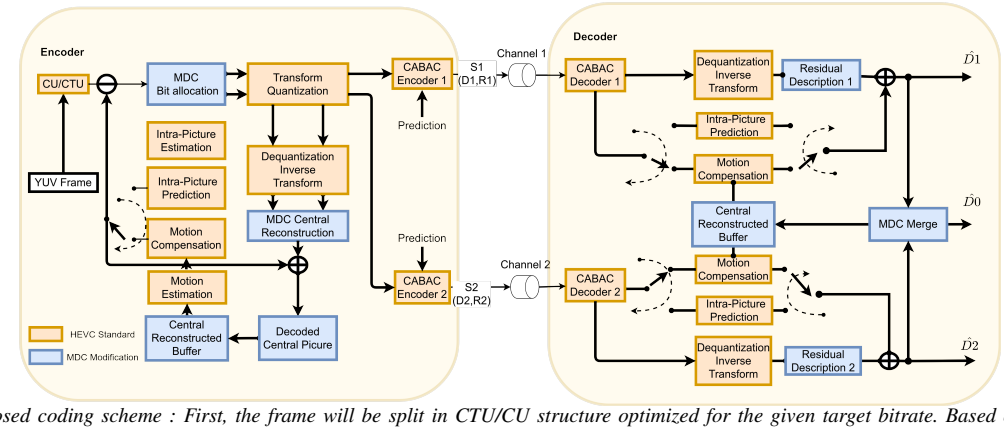


FIGURE 1 – *Proposed coding scheme : First, the frame will be split in CTU/CU structure optimized for the given target bitrate. Based on this structure and the residual produced by this step. The MDC bit allocation will allocate the QP value for each CTU using algorithm 1. Then, these QP values will feed the Transform and Quantization blocks at encoder side. After that the central reconstruction will be constructed by discarding the coarse quality CTU, this central reconstruction is stored in the Central Reconstructed Buffer. Finally MC process is performed on the central reconstruction.*

$$D_e = (1-p_e)\sum_{j=1}^{2} D_{p,j} + p_e(1-p_e)\sum_{j=1}^{2} D_{r,j} + p_e^2 D_{error} \quad (1)$$

Here, $D_{p,j}$ and $D_{r,j}$ are the total distortion of the principal CTUs and the redundant CTUs, respectively, in the same description j of the same frame. We define the following relationship :

$$D_{p,j} + D_{r,j} = \sum_{i=1}^{N} d_{i,j} \quad \forall j \in 1,2 \quad (2)$$

In this equation, $d_{i,j}$ is the quantization distortion of a CTU$_{i,j}$, where i is the CTU index in a frame, j is the description and N is the total number of CTUs in each frame. pe is the probability of packet error. The term D_{error} is the distortion when the two descriptions are lost simultaneously. Thus D_{error} is a constant and can be omitted from the cost function. The problem is to find the set of $QP_{i,j}$ to use for each CTU$_{i,j}$ in a frame which minimizes the expected distortion D_e under a frame target bit rate R_t. The optimal MDC rate-constrained optimization problem is then given by :

$$\begin{aligned} \min_{QP_{i,j}} \quad & D_e \qquad \forall j \in 1,2 \quad \forall i \in 1,...,N \\ \text{s.t.} \quad & R_j = \frac{R_t}{2}, \\ & QP_{min} \leq QP_{i,j} \leq QP_{max} \end{aligned}$$

This problem can be solved using the standard Lagrangian approach and minimizing the following cost function :

$$J_{\lambda_1,\lambda_2}(R_1,R_2) = D_e + \sum_{j=1}^{2} \lambda_j (R_j - R_t/2) \quad (3)$$

As the two descriptions are independent from each other, we can therefore establish :

$$J_{\lambda_1,\lambda_2}(R_1,R_2) = J_{\lambda_1}(R_1) + J_{\lambda_2}(R_2) \quad (4)$$

where $J_{\lambda_j}(R_j)$ contains only the terms of the corresponding description j and is given by :

$$J_{\lambda_j}(R_j) = (1-p_e)D_{p,j} + p_e(1-p_e)D_{r,j} + \lambda_j(R_j - R_t/2) \quad (5)$$

The solution to the optimization-constrained problem is given by the first-order conditions, leading to :

$$\begin{cases} \dfrac{\partial J_{\lambda_j}(R_j)}{\partial \lambda_j} = 0 \qquad \forall j \in \{1,2\} & (6) \\ \dfrac{\partial J_{\lambda_j(R_j)}}{\partial R_{i,j}} = 0 \qquad \forall j \in \{1,2\} \quad \forall i \in \{1,..,N\} & (7) \end{cases}$$

Due to the nonlinearity of the scalar quantizer in HEVC, minimizing the cost function is not straightforward. As a result, the rate-distortion function $d_{i,j}(R_{i,j})$ is not continuous and therefore not differentiable. To address this challenge, we approximate the rate-distortion relationship using the exponential function, given by :

$$d_{i,j}(R_{i,j}) = a_i e^{b_i R_{i,j}} \quad \forall i,j \quad (8)$$

where the parameters a_i and b_i are estimated using linear regression in each CTU of the residual frame produced by a pre-encoding process before encoding the two descriptions. Then, we can solve the problem using Algorithm 1, where $R_{i_{min}}$ and

Algorithm 1 Minimizing cost function (5)

Initialize $R_t, \lambda_{MAX}, \lambda_{MIN}, \epsilon, R_{i_{max}}, R_{i_{min}}$
while $|R_j - \frac{R_t}{2}| > \epsilon$ **do**
  $\lambda_j \leftarrow \frac{\lambda_{MAX}+\lambda_{MIN}}{2}$
  $\{R^*_{i,j}\} \leftarrow minimize(J_{\lambda_j}(R_j))$
  $\{R^*_{i,j}\} \leftarrow clip(R_{i_{max}}, R_{i_{min}}, R_{i,j})$
  if $R_j - \frac{R_t}{2} < 0$ **then**
    $\lambda_{MAX} \leftarrow \frac{\lambda_j+\lambda_{MIN}}{2}$
  else
    $\lambda_{MIN} \leftarrow \frac{\lambda_j+\lambda_{MAX}}{2}$
  end if
end while

$R_{i_{max}}$ are given by QP_{max} and QP_{min} repectively for each CTU$_{i,j}$. To prevent error propagation, the Instantaneous Decoder Refresh (IDR) frame allows the encoder to send an intra-frame signal to the decoder, clearing the Central Reconstructed Buffer. All frames can then be decoded from this IDR frame. Therefore, the encoder needs to select the appropriate amount of IDR frames to achieve the best coding quality concerning the channel noise. The study [9] has shown that the optimal IDR frame period under i.i.d. packet error distribution is given by :

$$T_{IDR} = \frac{1}{p_e} \quad (9)$$

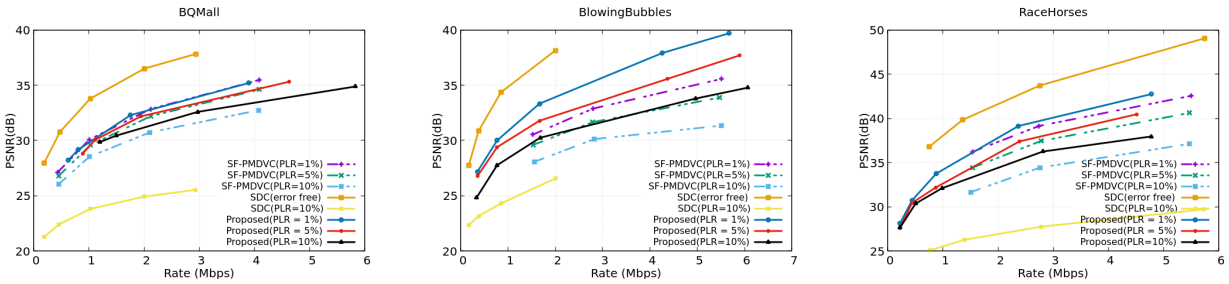


FIGURE 2 – *Comparison of the average Rate-Distortion with the method SF-PMDVC [7] under packet erasure : The experiment was conducted for three different packet loss rates : 0.1, 0.05, and 0.01. Each compressed stream was simulated three times over with different packet lost rate, and the average PSNR was computed. Our proposed MDC used LD-P configuration for encoding, while the referenced method used RA. SDC stands for Single Description Coding*

To decode the erroneous bitstream, if the principal CTU is lost, its redundant version will replace it. If two versions of the CTU are lost, the basic error concealment, which consists of replacing the block with the previous one, is applied.

3 Experimental Result

In this section, we evaluate the performance of our framework under a packet erasure channel. As mentioned earlier, the solution is implemented inside the HM codec [10]. To simulate the transmission, we use HEVC compressed streams with varying packet loss rates. Figure 2 shows that our proposed solution, which employs the LD-P configuration, outperforms the SF-PMDVC method with Random Access (RA) configuration for high packet error rates and high-motion sequences like BQMall, RaceHorses, and BlowingBubbles, while having a lower complexity encoding profile. Therefore, our proposed method is better suited for real-time applications.

4 Conclusion

In this study, we proposed a spatial-based multiple description encoding bit allocation and decoding solution that is adapted to HEVC standard. Our proposed MD coding scheme includes a bit allocation that distributes the redundancy between descriptions by adjusting the QP value for each CTU within a frame based on the channel characteristics and an IDR adaptation. This solution meets the requirements of low latency and good compression performance, making it suitable for use in remote control vehicles.

In our perspective, a more robust error handling mechanism with finer grain error detection at the CU level will enhance the performance of the system. Various methods such as those discussed in [11, 12], could be employed to improve the decoding performance. Additionally, a scheme of optimization with error mismatch propagation model should improve the performance of the system.

Appendix

This is the summary of the article that has been submitted for the IEEE ICIP 2023 conference and is currently awaiting the review process.

Références

[1] C. Soldani, G. Leduc, F. Verdicchio, et A. Munteanu. Multiple description coding versus transport layer fec for resilient video transmission. Dans *International Conference on Digital Telecommunications (ICDT'06)*, pages 20–20, 2006.

[2] Tammam Tillo, Marco Grangetto, et Gabriella Olmo. Redundant slice optimal allocation for H.264 multiple description coding. *IEEE Trans. Circuits Syst. Video Technol.*, 18(1) :59–70, 2008.

[3] Ivana Radulovic, Pascal Frossard, Ye-Kui Wang, Miska M. Hannuksela, et Antti Hallapuro. Multiple description video coding with H.264/AVC redundant pictures. *IEEE Trans. Circuits Syst. Video Technol.*, 20(1) :144–148, 2010.

[4] Chia-Wei Hsiao et Wen-Jiin Tsai. Hybrid multiple description coding based on H.264. *IEEE Trans. Circuits Syst. Video Technol.*, 20(1) :76–87, 2010.

[5] Kostas E. Psannis. HEVC in wireless environments. *J. Real Time Image Process.*, 12(2) :509–516, 2016.

[6] Muhammad Majid, Muhammad Owais, et Syed Muhammad Anwar. Visual saliency based redundancy allocation in HEVC compatible multiple description video coding. *Multim. Tools Appl.*, 77(16) :20955–20977, 2018.

[7] Feifeng Wang, Jing Chen, Huanqiang Zeng, et Canhui Cai. Spatial-frequency HEVC multiple description video coding with adaptive perceptual redundancy allocation. *J. Vis. Commun. Image Represent.*, 88 :103614, 2022.

[8] Mohamed Aymen Labiod, Mohamed Gharbi, François-Xavier Coudoux, Patrick Corlay, et Noureddine Doghmane. Cross-layer scheme for low latency multiple description video streaming over vehicular ad-hoc networks (vanets). *AEU - International Journal of Electronics and Communications*, 104 :23–34, 2019.

[9] Guy Côté et Faouzi Kossentini. Optimal intra coding of blocks for robust video communication over the internet. *Signal Process. Image Commun.*, 15(1-2) :25–34, 1999.

[10] High Efficiency Video Coding (HEVC) Test Model 16 (HM 16) Encoder Description Update 10 | MPEG.

[11] Trung Hieu Le, Marc Antonini, Marc Lambert, et Karima Alioua. Codage vidéo à description multiple basé sur hevc pour le pilotage de véhicules semi-autonomes. Dans *GRETSI*, 2022.

[12] Marie Andrée Agostini, Marc Antonini, et Michel Kieffer. Map estimation of multiple description encoded video transmitted over noisy channels. Dans *ICIP*, pages 3069–3072. IEEE, 2009.

Optimisation du décodage par liste de vidéos corrompues basée sur une architecture CNN

Y. Zhang[1,2] S. Coulombe[1] F-X. Coudoux[2] A. Trioux[2] P. Corlay[2]

[1] Dept. of Software and IT Engineering, École de technologie supérieure, Montreal, Canada

[2] UMR 8520 - IEMN, DOAE, Univ. Polytechnique Hauts-de-France, CNRS, Univ. Lille, YNCREA, Centrale Lille, France

{yujing.zhang.1@ens.etsmtl.ca, stephane.coulombe@etsmtl.ca}

{francois-xavier.coudoux, anthony.trioux, patrick.corlay}@uphf.fr

Résumé

Cet article présente une solution de décodage par liste optimisée pour des vidéos corrompues par des erreurs de transmission. Elle est basée sur l'évaluation de la qualité des images sans référence utilisant un réseau de neurones convolutif (CNN) qui gère efficacement les distorsions non uniformes. À l'issue d'un processus de décodage par liste, nous évaluons la qualité de chaque image candidate générée (sans référence) afin de sélectionner la meilleure. Lorsque l'erreur de transmission se produit dans une image intra, notre architecture a une précision de décision de plus de 98% contre 46% pour l'architecture CNN originale pré-entraînée. Pour les erreurs dans une image inter, c'est 79% contre 33%.

Mots clefs

Transmission vidéo, distorsions non uniformes, évaluation de la qualité des images, réseau de neurones convolutif.

1 Introduction

Nous assistons à un développement très rapide des applications impliquant la transmission de contenus vidéos. Cependant, les erreurs de transmission sur des réseaux sans-fil compromettent gravement la qualité visuelle des contenus vidéos reçus, ce qui se traduit par une mauvaise qualité d'expérience pour l'utilisateur final. Différentes approches existent dans la littérature pour réparer les paquets vidéos erronés reçus [1, 2, 3, 4, 5, 6]. Parmi celles-ci, nous nous intéressons aux approches de décodage par liste qui exploitent les paquets reçus corrompus. À partir de chaque paquet corrompu, la méthode génère plusieurs paquets *candidats*. Ces candidats représentent diverses tentatives de correction du paquet erroné. Le défi consiste à estimer sans référence au récepteur la qualité de chacun de ces candidats pour ensuite choisir le meilleur. Ce dernier correspondra idéalement à la version intacte originalement transmise.

Dans cet article, nous proposons donc un cadre d'optimisation du décodage par liste où une évaluation sans référence de la qualité visuelle permet d'identifier le meilleur candidat parmi une liste de plusieurs. Notre approche est basée sur l'usage d'un réseau de neurones convolutif (CNN) modifié afin de permettre la prise en compte de distorsions non-uniformes dues aux erreurs de transmission. Nos principales contributions sont :

1. Une nouvelle méthode d'évaluation de la qualité basée sur le CNN présenté dans [7], mais améliorée à plusieurs égards, dont une normalisation et une mesure de qualité locales, opérant par patch, pour supporter des distorsions non-uniformes dans les images.
2. Une nouvelle base de données constituée de vidéos encodées avec la norme High Efficiency Video Coding (HEVC) [8] et auxquelles nous avons injecté des erreurs de transmission. Cela mène à des images possédant des artéfacts non-uniformément distribués spatialement sur lesquelles notre système peut s'entrainer.
3. Un nouveau cadre d'optimisation du décodage par liste capable de sélectionner la vidéo ayant la meilleure qualité visuelle parmi plusieurs candidats.

À la section 2, nous présentons la méthode proposée. À la section 3, nous présentons nos résultats expérimentaux. Enfin, nous présentons nos conclusions à la section 4.

2 La méthode proposée

2.1 Cadre d'optimisation de décodage

À la Figure 1, nous proposons un cadre pour améliorer le processus de décodage par liste constitué de : 1) la génération d'une base de données d'images avec distorsions non uniformes, 2) l'apprentissage pour l'évaluation de la qualité (entraînement), et 3) la sélection du meilleur candidat. Le processus de génération de la base de données comprend différentes étapes : l'encodage vidéo par la norme HEVC, la génération d'erreurs de transmission et le décodage vidéo par liste, sans dissimulation d'erreur, pour obtenir les N candidats représentant des tentatives, pour la plupart infructueuses, de correction de la vidéo. Le processus d'entraînement à évaluer la qualité comprend la conversion d'images (du format YUV au format utilisé pendant l'apprentissage), la génération de patchs et l'apprentissage du réseau de neurones de manière supervisée en utilisant une métrique de qualité avec référence complète. Le choix du meilleur candidat est réalisé en identifiant le candidat avec la qualité la plus élevée.

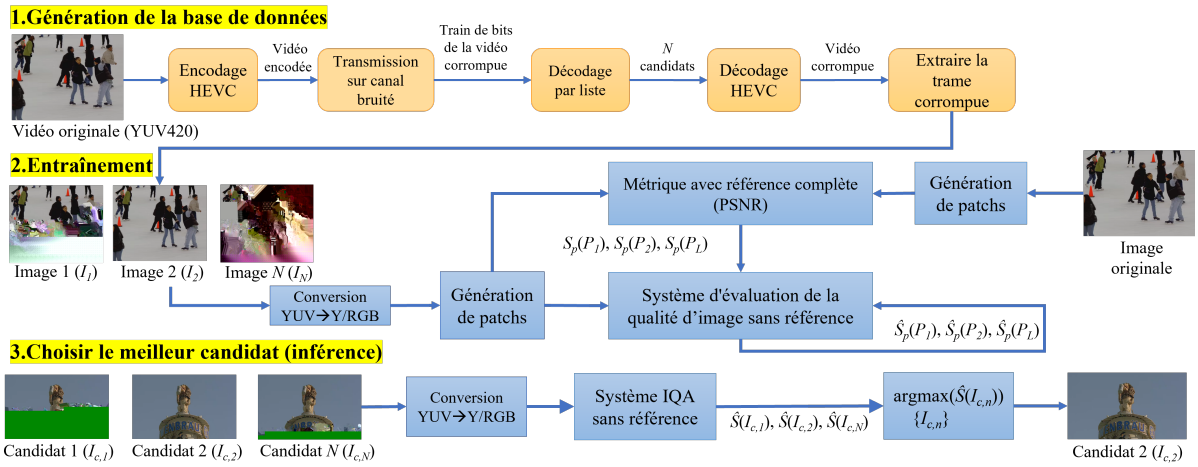


FIGURE 1 – *Cadre proposé pour optimiser le décodage par liste de vidéos corrompues lors de la transmission*

2.2 Méthode d'évaluation de la qualité visuelle améliorée basée sur le CNN

Plusieurs métriques basées sur les CNN [7, 9] séparent une image en plusieurs patchs de taille réduite et extraient les caractéristiques de chaque patch pour évaluer leur qualité. Les modèles basés sur les patchs attribuent souvent à tous les patchs de l'image le même niveau de qualité que l'image complète lors de l'apprentissage [10], ce qui donne de bons résultats pour les distorsions uniformes, mais n'est pas une approche souhaitable lorsque l'on considère des distorsions non uniformes. Nous proposons donc d'utiliser des scores locaux afin que chaque patch ait un score de qualité qui lui est propre. Cela peut aider le réseau de neurones à apprendre plus efficacement les distorsions locales.

Nous choisissons l'architecture CNN proposée dans [7] comme architecture de base et l'améliorons afin de l'adapter à notre objectif. L'architecture originale est un réseau neuronal à 5 couches, dont 1 couche convolutive, 2 couches de pooling et 2 couches entièrement connectées. Comme illustré à la Figure 2, nous utilisons, dans cet article, la structure du réseau suivante : $64\times64\times3-58\times58\times50-2\times50-800-800-1$. Au lieu d'effectuer la simulation uniquement sur le canal de luminance comme dans [7], nous étendons nos expériences à trois canaux R, G, B. L'architecture de base [7] utilise une méthode de normalisation du contraste local. Supposons que la valeur d'intensité d'un pixel à l'emplacement (i,j) soit $v(i,j)$, les auteurs calculent alors sa valeur normalisée $v_n(i,j)$ comme suit :

$$v_n(i,j) = \frac{v(i,j)-\mu(i,j)}{\sigma(i,j)+C}, \text{ avec}$$

$$\mu(i,j) = \frac{1}{(2W+1)^2}\sum_{p=-W}^{p=W}\sum_{q=-W}^{q=W} v(i+p,j+q)$$

$$\sigma(i,j) = \sqrt{\frac{1}{(2W+1)^2}\sum_{p=-W}^{p=W}\sum_{q=-W}^{q=W}[v(i+p,j+q)-\mu(i,j)]^2}$$

où C est une constante positive qui empêche la division par zéro. La taille de la fenêtre de normalisation est de $(2W+1)\times(2W+1)$ pixels avec $W=3$. Cependant, cette méthode pose un problème lorsqu'elle est appliquée à des patchs uniformes. Le problème est que le décodeur initialise chaque patch au format YUV à (0,0,0) lorsqu'il démarre le décodage d'une image. Considérant un seul canal, par exemple Y, nous ne pouvons pas faire la distinction entre un patch uniforme bien reçu dont la valeur est normalisée à 0 et un patch erroné qui est uniforme parce qu'il a été initialisé à 0 par le décodeur. Cette situation est problématique quand elle survient dans la base de données d'entraînement, car le réseau neuronal devient confus pendant l'apprentissage. En effet, à l'issue de la normalisation, un patch uniforme et un patch en erreur deviennent identiques et entrent dans les couches du CNN avec des scores de référence différents pour l'apprentissage. Pour éviter ce problème, nous améliorons la normalisation locale en séparant ces deux situations. Lorsque nous détectons $\sigma(i,j)=0$ dans les patchs d'entrée, nous calculons $\mu(i,j)$. Si $\mu(i,j)\neq 0$, nous forçons $v_n(i,j)$ à être égal à $\epsilon\neq 0$ après normalisation. Nous utilisons l'Eq.(1) sur chaque canal d'une image en format RGB où nous forçons la valeur à (0,0,0) lorsque le décodeur récupère YUV à (0,0,0) suite à une erreur.

$$v_n(i,j) = \begin{cases} 0, & \text{si } \sigma(i,j)=0 \text{ et } \mu(i,j)=0 \\ \epsilon, & \text{si } \sigma(i,j)=0 \text{ et } \mu(i,j)\neq 0 \end{cases} \quad (1)$$

Notre réseau neuronal est tout d'abord entraîné sur des patchs non chevauchants de 64×64 pixels, correspondant à un coding tree unit (CTU) en HEVC [8], provenant d'images haute définition. Pour l'entraînement, nous attribuons à chaque patch un score de qualité Peak signal-to-noise ratio (PSNR), calculé entre le patch corrompu et le patch correspondant dans l'image originale avant encodage (métrique avec référence). Pour les tests, nous utilisons la moyenne des scores de patchs prédits pour chaque image afin d'obtenir le score $\hat{S}(I)$ de qualité au niveau image :

$$\hat{S}(I) = \frac{1}{L}\sum_{l=0}^{L-1}\hat{S}_p(P_l)$$

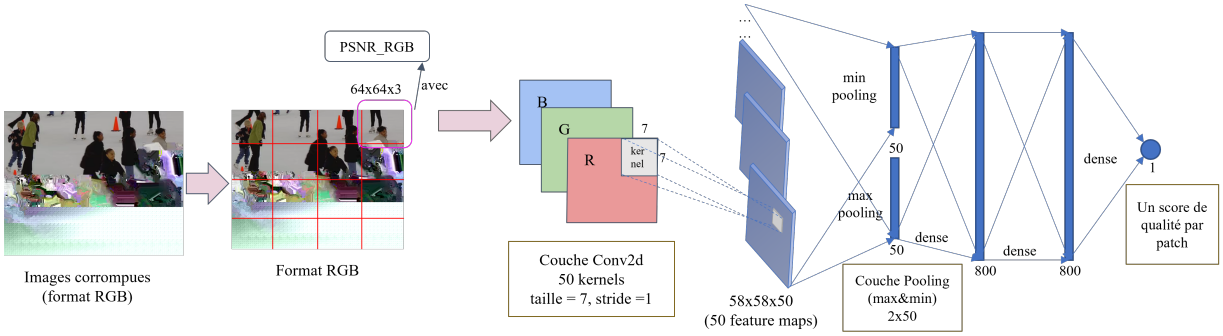


FIGURE 2 – *Architecture CNN proposée pour évaluer la qualité des images avec distorsions non-uniformes*

où $\hat{S}_p(P_l)$ indique le score de qualité prédit pour le patch P_l par le CNN (métrique sans référence), et L est le nombre total de patchs dans l'image. Le fait d'utiliser des petits patchs en entrée élargit considérablement l'échantillon d'apprentissage pour le CNN et évite le problème de manque de données rencontré lors de l'utilisation d'un ensemble d'images complètes. Nous utilisons la fonction de perte de [7], la descente de gradient stochastique et la rétropropagation sont utilisées pendant l'entraînement. Nous utilisons un ensemble de validation pour éviter un ajustement excessif et conservons les paramètres du modèle qui génèrent la valeur de *Spearman Rank Order Correlation Coefficient* la plus élevée sur l'ensemble de validation.

3 Résultats expérimentaux

3.1 Base de données vidéo utilisée

Toutes les séquences originales utilisées dans nos expériences proviennent de [11, 12]. Les vidéos collectées sont au format YUV avec une résolution de 1920×1024. Nous extrayons les 10 premières images de chaque vidéo pour les encoder avec la norme HEVC. Parmi les différentes valeurs du pas de quantification (QP) possibles, nous avons choisi 37 qui correspond à une valeur fréquemment utilisée. Nous supposons que chaque image encodée est contenue dans un seul paquet vidéo. La première image de la vidéo encodée est une image intra (I), et les 9 images suivantes sont des images inter (P). Pour simuler la combinaison d'une erreur de transmission suivie d'un décodage par liste où les bits sont inversés à différents endroits, des positions de bits inversés sont choisies en fonction de l'équation $p = \alpha \times M$, où $\alpha = \{0.1, 0.2, ..., 0.9, 0.99\}$ et M est la taille de chaque paquet. Ainsi, pour chaque séquence, nous avons 11 candidats à chaque fois (dont 1 est le candidat sans erreur). Nous obtenons 990 images corrompues à partir de 90 images de référence et finalement 475 200 patchs pour l'apprentissage, avec une taille de patch de 64×64 pixels. Nous avons séparé notre base de données en ensembles d'entraînement et de test, avec une répartition de 60% et 40%, respectivement. Chaque patch est associé à un score PSNR dans l'intervalle [0, 50] dB, qui est normalisé dans l'intervalle [0, 1] pendant l'apprentissage. Sur la base de résultats empiriques de simulation, nous fixons $\epsilon = -0.013$ dans l'Eq.(1).

3.2 Évaluation des performances

Nous entraînons et testons le modèle original CNN et notre version améliorée sur la nouvelle base de données dédiée. Afin de mieux évaluer les performances des modèles, nous définissons plusieurs métriques. Comme indiqué ci-dessous, $\overline{S}_{\text{intact}}$ indique le PSNR moyen, par rapport aux versions originales, de toutes les images intactes, qui sont compressées mais reçues sans erreur de transmission. Il est calculé sur YUV selon [8] et noté PSNR_{YUV}. $\overline{S}_{\text{système}}$ représente le PSNR moyen, par rapport aux versions originales, de toutes les images sélectionnées par une méthode donnée. $\overline{S}_{\text{diff}}$ est, pour une méthode, la différence absolue entre la qualité moyenne des images sélectionnées et celle des images intactes. N correspond au nombre de séquences originales considérées. K est le nombre de candidats, $I_{\text{c},i}$, parmi lesquels choisir pour chaque image corrompue.

$$\overline{S}_{\text{intact}} = \frac{1}{N}\sum_{n=0}^{N-1} \text{PSNR}_{\text{YUV}}(I_{\text{original},n}, I_{\text{intact},n}),$$

$$\overline{S}_{\text{système}} = \frac{1}{N}\sum_{n=0}^{N-1} \text{PSNR}_{\text{YUV}}(I_{\text{original},n}, I_{\text{système},n}),$$

$$\text{où } I_{\text{système}} = \underset{\{I_{\text{c},i},\ 0\le i<K\}}{\arg\max}\ \hat{S}(I_{\text{c},i}),\quad \overline{S}_{\text{diff}} = \left|\overline{S}_{\text{intact}} - \overline{S}_{\text{système}}\right|$$

Méthodes	Précision	$\overline{S}_{\text{intact}}$ (dB)	$\overline{S}_{\text{système}}$ (dB)	$\overline{S}_{\text{diff}}$ (dB)
CNN_Y_G pré-entraîné [7]	45.6%	39.18	28.69	10.49
CNN_Y proposé	93.0%		38.39	0.79
CNN_RGB proposé	96.5%		38.88	**0.30**
CNN_Y_NL proposé	**98.2%**		38.60	0.58
CNN_RGB_NL proposé	96.5%		38.88	**0.30**

TABLEAU 1 – *Performances sur les images codées en intra*

Méthodes	Précision	$\overline{S}_{\text{intact}}$ (dB)	$\overline{S}_{\text{système}}$ (dB)	$\overline{S}_{\text{diff}}$ (dB)
CNN_Y_G pré-entraîné [7]	33.3%	38.62	32.99	5.63
CNN_Y proposé	60.0%		30.19	8.43
CNN_RGB proposé	66.7%		36.49	2.13
CNN_Y_NL proposé	77.0%		36.55	2.07
CNN_RGB_NL proposé	**79.0%**		36.71	**1.91**

TABLEAU 2 – *Performances sur les images codées en inter*

Les Tableaux 1 et 2 présentent les résultats expérimentaux obtenus en utilisant des images codées respectivement en *intra* et en *inter*, en comparant la méthode CNN originale à différentes configurations de simulation. Pour les résultats

sur image *inter*, l'erreur a directement frappé l'image inter en question, et ne correspond pas à une propagation d'erreur survenue dans l'image intra précédente. Les meilleurs résultats sont indiqués en gras.

La première ligne de chaque tableau, CNN_Y_G pré-entraîné, utilise le modèle déjà entraîné de l'article [7], qui applique le même score pour chaque patch de la composante Y de l'image (score global), et teste avec notre base de données avec distorsions non uniformes. CNN_Y proposé indique la solution proposée où on utilise un score différent par patch de luminance et où nous ré-entraînons et testons sur notre base de données (comme toutes les méthodes proposées). Nous pouvons constater l'intérêt d'utiliser un score local par patch et de ré-entraîner sur la base de données proposée puisque la précision passe de 46% à 93% sur les images intra et de 33% à 60% sur les images inter. Nous croyons que ce score local permet de mieux apprendre les caractéristiques des distorsions provenant d'erreurs de transmission. CNN_RGB proposé indique que les images utilisées initialement au format YUV sont converties au format RGB pour l'entraînement et l'inférence. Une méthode avec le suffixe _NL indique une configuration qui applique notre méthode de normalisation locale améliorée (Eq.(1)). Pour les images codées en intra, l'utilisation de la normalisation locale permet d'obtenir une meilleure précision et une moins grande différence de qualité lorsque le CNN utilise le canal Y. Cependant, aucune amélioration n'est obtenue lorsque RGB est utilisé. Pour les images inter, l'utilisation de la normalisation locale permet d'améliorer significativement les performances tant pour Y que pour RGB avec une précision qui passe de 60% à 77% pour Y et de 68% à 79% pour RGB. Finalement, bien que les performances soient similaires pour Y et RGB en intra, l'usage de RGB performe mieux en inter. Nous croyons qu'il a l'avantage de pouvoir identifier les distorsions de couleurs. On note que la précision est beaucoup plus faible pour les images inter que pour les intra. En effet, les erreurs de transmission dans les images inter n'engendrent pas de pertes aussi importantes de qualité que sur les images intra, ce qui rend plus difficile l'apprentissage du modèle.

Nous pouvons voir l'intérêt du score local, du ré-entraînement sur notre base de données et la normalisation locale. Néanmoins, les performances ne sont pas optimales et plusieurs travaux sont envisagés pour les améliorer. Par exemple, on pourrait penser à modifier la taille des patchs pour pouvoir détecter les discontinuités aux frontières des CTUs HEVC, rendues visibles suite à la présence d'erreurs de transmission. Aussi, nous pourrions adapter notre système en opérant directement en YUV plutôt qu'en Y ou RGB pour détecter les distorsions de couleur tout en évitant des conversions supplémentaires. Les derniers résultats de ces améliorations seront présentés lors de la conférence.

4 Conclusion

Nous avons développé une architecture d'évaluation de la qualité d'image reposant sur un CNN existant, modifié pour devenir sensible à des distorsions non uniformes, rencontrées lors de transmission avec erreurs. Cette méthode peut être utilisée pour extraire la meilleure reconstruction d'une liste d'images candidates générées à l'issue d'un processus de correction d'erreurs. Sur cette base, nous présentons également un cadre de simulation pour simuler le processus de génération d'images candidates, d'évaluation de la qualité de l'image et de sélection de la meilleure image. Notre architecture possède une précision de décision de plus de 98% lorsque l'erreur de transmission est localisée dans une image intra et d'environ 80% en inter.

Références

[1] Yao Wang et Qin-Fan Zhu. Error control and concealment for video communication : A review. *Proceedings of the IEEE*, 86(5) :974–997, 1998.

[2] W-Y Kung et al. Spatial and temporal error concealment techniques for video transmission over noisy channels. *IEEE transactions on circuits and systems for video technology*, 16(7) :789–803, 2006.

[3] Xijin Liu et al. Exploiting error-correction-CRC for polar SCL decoding : A deep learning-based approach. *IEEE Transactions on Cognitive Communications and Networking*, 6(2) :817–828, 2020.

[4] Jinzhi Lin et al. Joint source-channel decoding of polar codes for HEVC-based video streaming. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(4), mar 2022.

[5] Vivien Boussard et al. Table-free multiple bit-error correction using the CRC syndrome. *IEEE Access*, 8 :102357–102372, 2020.

[6] Galina Sabeva et al. Robust decoding of H.264 encoded video transmitted over wireless channels. Dans *2006 IEEE Workshop on Multimedia Signal Processing*, pages 9–13, 2006.

[7] Le Kang et al. Convolutional neural networks for no-reference image quality assessment. Dans *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[8] Vivienne Sze et al. *High Efficiency Video Coding (HEVC) : Algorithms and Architectures*. Springer Publishing, 2014.

[9] Simone Bianco et al. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12 :355–362, 2018.

[10] Junyong You et Jari Korhonen. Transformer for image quality assessment. Dans *2021 IEEE international conference on image processing (ICIP)*, pages 1389–1393. IEEE, 2021.

[11] Xiph.org video test media [derf's collection].

[12] Margaret H. Pinson. The consumer digital video library [best of the web]. *IEEE Signal Processing Magazine*, 30(4) :172–174, 2013.

Transformer-Based Image Compression Without Positional Encoding

Bouzid Arezki, Fangchen Feng, Anissa Mokraoui
L2TI, Université Sorbonne Paris Nord
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
{bouzid.arezki, fangchen.feng, anissa.mokraoui}@univ-paris13.fr

Abstract

In this paper, we address the image compression problem and introduce the Swin Non-Positional Encoding (SwinNPE) transformer. SwinNPE improves the efficiency of the SwinT transformer while reducing the number of model parameters. We generalize the Swin cell and propose the Swin convolutional block, which can better handle the local correlation between image patches. Additionally, the Swin convolutional block can capture the local context between tokens without relying on positional encoding, reducing the model complexity. Preliminary results show that SwinNPE outperforms state-of-the-art CNN-based architectures in terms of the trade-off between bit-rate and distortion, achieving results comparable to SwinT with 16% less computational complexity on the Kodak dataset.

Key Words

Image Processing, Image Compression, Transformer, Transform Coding, Attention Mechanism.

1 Introduction

Transform coding is a widely used approach for image compression and forms the basis for many popular coding standards, such as JPEG. Codecs based on transform coding typically comprise three components for lossy compression : transform, quantization, and entropy coding. These components have all been improved using deep neural networks through end-to-end training, as demonstrated by various works [1, 2, 3, 4, 5, 6].

As one of the first works, the authors of [1] proposed a CNN-based two-level hierarchical variational autoencoder with hyper-prior as the entropy model. This architecture consists of two pairs of encoders/decoders, one for the generative model and another for the hyper-prior model.

Recently, transformers [7] have had great success in the computer vision area including neural image compression. The authors of [8] incorporated the attention mechanism into the image compression framework by introducing self-attention in the hyper-prior model. The more sophisticated Swin block [9] is also used in [10] in both the generative and the hyper-prior model to adopt shift window-based attention to restrict the attention in local windows. Indeed, with the attention mechanism that can better handle global context compared to convolutional neural networks, transformers have the ability to adapt the receptive field depending on the task conversely to CNNs where the kernel size is fixed. This better understanding of global information allows for capturing long-range dependencies in image compression applications.

Positional encoding is a vital component of transformers. The original ViT transformer [7] breaks down images into non-overlapping series of patches mapping each patch to a token. The standard transformer layers are then used to read the whole sequence of tokens at once. The positional encoding, therefore, plays a crucial role to maintain the sequence order, and different variations are proposed for better modeling the positional information of the sequence and maintaining the local context [11, 12, 13]. In the context of image compression, the benefits of positional encoding have been demonstrated in terms of Rate-Distortion (RD) performance in works such as [8, 10]. In particular, the authors of [8] have shown that a 2D diamond-shaped relative position encoding is useful and has particular advantages.

Despite its many advantages, using positional encoding in transformers can increase the dimensionality of embeddings, leading to higher computational costs during training and limiting the flexibility of the models. Recently, the authors of [14] demonstrated that the positional encoding can be abandoned in the attention module for image classification without any drop in performance. This was achieved by introducing convolution in the tokenization process of patches and in the self-attention block to maintain local spatial information. It is claimed that this combination of convolution and the attention mechanism benefits from both the advantages of convolutional neural networks and transformers.

In this paper, we present a new image compression framework called SwinNPE. It is based on our proposed *convolutional Swin block* which combines patch convolution and shift window-based attention in Swin without positional encoding. We believe that this framework can better capture spatial contextual information. Our preliminary experiments show that SwinNPE achieves comparable results to the SwinT architecture [10], without the need for positional encoding and with fewer parameters.

2 Proposed framework

The proposed SwinNPE uses the same architecture as in [10], which is shown in Figure 1. Specifically, the in-

put image x is first encoded by the generative encoder $y = g_a(x)$, and the hyper-latent $z = h_a(y)$ is obtained. The quantized version of the hyper-latent $\hat{z}$ is modeled and entropy-coded with a learned factorized prior to passe through $h_s(\hat{z})$ to obtain μ and σ which are the parameters of a factorized Gaussian distribution $P(y|\hat{z}) = \mathcal{N}(\mu, diag(\sigma))$ to model y. The quantized latent $\hat{y} = Q(y - \mu) + \mu$ is finally entropy-coded and sent to $\hat{x} = g_a(\hat{y})$ to reconstruct the image $\hat{x}$. We use the classical strategy of adding uniform noise to simulate the quantization operation which makes the operation differentiable. The channel-wise autoregressive block [2, 3] is designed to learn the auto-regressive prior which factorizes the distribution of the latent as a product of conditional distributions incorporating prediction from the causal context of the latents [4, 5, 6].

The generative and the hyper-prior encoder, g_a and h_a, are built with the patch merge block and the convolutional Swin block. The patch merge block contains the *Depth-to-Space* operation [10] for down-sampling, a normalization layer, and a linear layer to project the input to a certain depth C_i. In g_a, the depth C_i of the latent representation increases as the network gets deeper which allows for getting a more abstract representation of the image. The size of the latent representation decreases accordingly. In each stage, we down-sample the input feature by a factor of 2.

The proposed *convolutional Swin block* is a generalization of the Swin cell [9]. As shown in Figure 2, we use convolutions instead of position-wise linear projections to project the K, Q, and V matrices in the multi-head attention block. This makes the attention module more sensitive to spatial context. Instead of using hand-crafted positional encoding, we let the convolution layer capture the positional information. In this paper, we use depth-wise separable convolution [14] due to its parameter efficiency. More specifically, the depth-wise separable convolution first applies a 2D convolution in each feature channel independently. The outcome is then concatenated and passed through another convolution layer, such convolution reduces the number of parameters and computation while increasing representational efficiency where it deals not just with spatial dimension but with depth dimension already. It's important to note that the proposed block is not limited to convolution operations. Different forms of convolution [15, 16] are possible, making the proposed convolutional Swin block particularly flexible. Compared to the convolutional attention block in [14], we keep the shift window structure which allows cross-window connections.

The generative and the hyper-prior decoder, g_s and h_s, are built with the patch split block and the convolutional Swin block. In the patch split block, we reverse the merging sequence and use *Space-to-Depth* operation [10] for up-sampling.

3 Experiment and Analysis

3.1 Experiment configuration

This section presents an assessment of the SwinNPE architecture and a comparison of its image compression results against state-of-the-art approaches. The SwinNPE was trained on the CLIC2020 training set for 3.3 million steps. During training, each batch consisted of eight randomly cropped images with a size of 256×256 pixels.

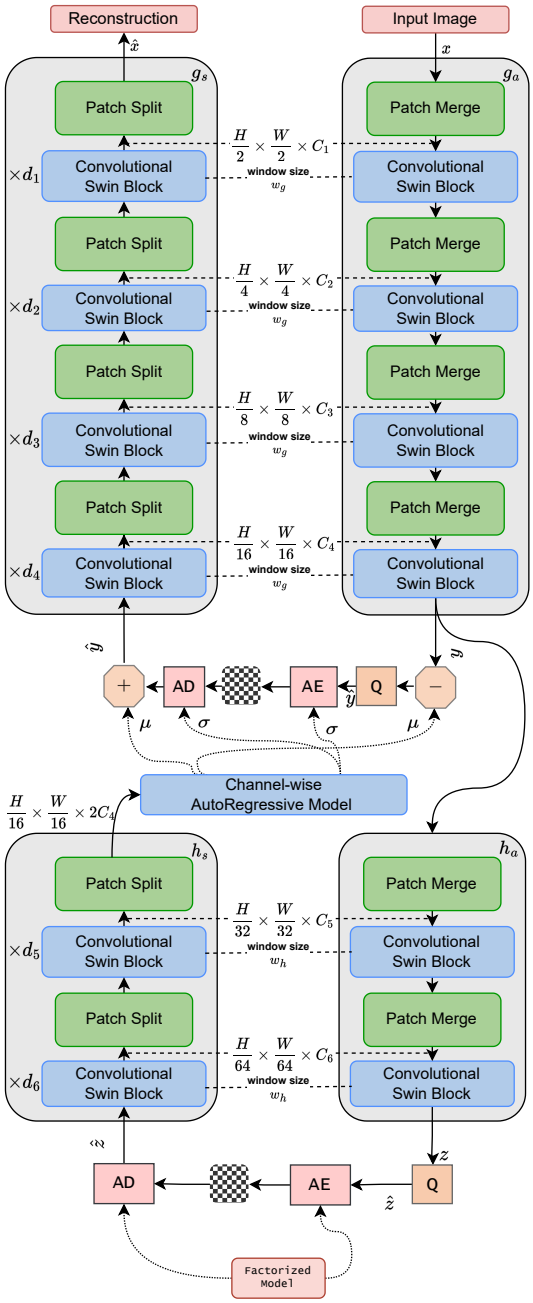


FIGURE 1 – *Network architecture of our proposed SwinNPE.*

The SwinNPE's performance was evaluated on the Kodak [17] dataset and we center-cropped all images to mul-

tiples of 256 to avoid padding. We choose the following loss function to optimize the trade-off between the bit-rate R and the quality of reconstruction D which corresponds to the Mean Squared Error (MSE) in RGB color space :

$$L = D + \beta R, \tag{1}$$

with $\beta \in \{0.003, 0.001, 0.0003, 0.0001\}$.
The learning rate starts at 10^{-4} and the hyper-parameters of the architecture shown in Figure 1 are as follows. $(d_1, d_2, d_3, d_4, d_5, d_6) = (2, 2, 6, 2, 5, 1)$, $(w_g, w_g) = (8, 8)$, $(w_h, w_h) = (4, 4)$, and $(C_1, C_2, C_3, C_4, C_5, C_6) = (128, 192, 256, 320, 192, 192)$.
For the autoregressive model, we use the model proposed in [6] with 10 slices. The kernel size in all convolutional Swin blocks for depth-wise separable convolution is set to 3.

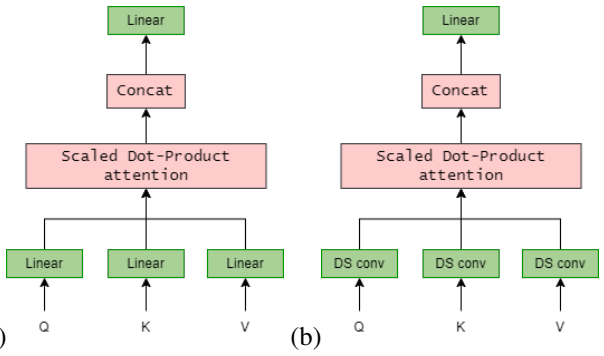


FIGURE 2 – *(a) The attention mechanism scheme for multi-head attention (b) The attention mechanism scheme for convolutional Swin. DS conv means depthwise separable convolution.*

3.2 Analysis

We compare our proposed SwinNPE with the results of two transformers-based architectures [10, 8] and some of the most used CNN-based image compression architectures and standard codecs on the Kodak dataset [17]. The rate-distortion curves of different methods are shown in Figure 3. We summarize the number of parameters of the tested transformer-based architectures in Table 1 where we also illustrate the Bijonteguard metric [18] using the SwinT-CHARM as the reference.
From Figure 3, we can clearly see that the SwinNPE outperforms all of the tested CNN-based architectures in terms of the bit-rate/distortion tradeoff. It is particularly interesting to notice that our proposed approach obtains almost the same results as Entroformer [8] (orange dashed line in Figure 3) with much less model parameters (see Table 1). Specifically, the saving bit-rate of SwinNPE is 5.46% less than SwinT-CHARM (optimal saving bit-rate) which is at the same level as Entroformer with 4.33% more bit-rate saving compare to SwinT-CHARM. We argue that it is due to the fact that the convolutional layer in the proposed convolutional Swin block can capture the local contextual information. With fewer parameters, the proposed SwinNPE has results comparable to SwinT-CHARM. We emphasize that our proposed architecture is particularly advantageous compared to SwinT-based architecture without positional encoding [1] validating the advantages of combining convolutions and transformers for image compression.

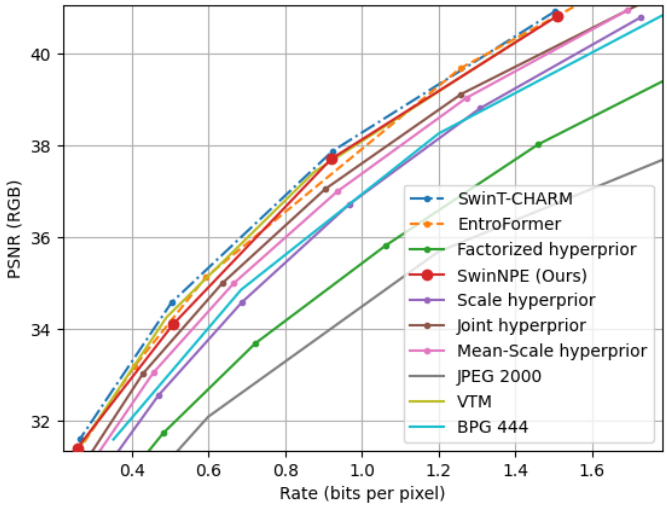


FIGURE 3 – *SwinNPE achieves nearly the same results as Entroformer [8] and SwinT-CHARM [10] that relying on Positional encoding and better RD performance than CNNs-based methods Factorized [19], Scale [1], Mean-Scale [4], Joint hyperprior [4] and standard codecs on the Kodak image set.*

4 Conclusion

In this paper, we propose SwinNPE, a transformer-based image compression model built with convolutional Swin blocks without positional encoding. SwinNPE achieves comparable results to state-of-the-art methods while using fewer model parameters and outperforming CNN-based architectures. The proposed convolutional Swin block allows for better exploitation of spatial context without the need for positional encoding, resulting in greater flexibility and fewer parameters.
For future work, it would be interesting to explore the use of different convolution operations and sizes in the proposed SwinNPE model. This could allow for more accurate modeling of complex spatial relationships and patterns, leading to improved performance in image compression. Additionally, incorporating the convolution operation into the patch merge/split module could benefit from the advantages of CNN. The proposed SwinNPE model with convolutional Swin blocks provides a promising direction for the development of efficient and effective transformer-based models for image compression.

1. The results are shown in the ablation studies in [10].

Network	#Param. (M)	Positional encoding	Bijonteguard Metric	
			$\Delta PSNR$	$\% \Delta rate$
SwinT-CHARM [10]	32	Positional Relative Encoding 2D	0	0%
Entroformer [8]	142.7	Positional Relative Encoding 2D + Diamond	-0.228	4.33%
SwinNPE (Ours)	**27**	-	-0.311	5.46%

TABLE 1 – *Performance comparison using Bijonteguard metric [18] where ΔPSNR measures the average PSNR difference and % Δ rate the average rate saving in percent between SwinT-CHARM [10] (selected as the reference network) and another given network .*

Références

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, et Nick Johnston. Variational image compression with a scale hyperprior. Dans *6th Inter. Conf. on Learning Representations (ICLR)*, 2018.

[2] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, et David Zhang. Learning convolutional networks for content-weighted image compression. Dans *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3214–3223, 2018.

[3] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, et Luc Van Gool. Conditional probability models for deep image compression. Dans *2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 4394–4402, 2018.

[4] David Minnen, Johannes Ballé, et George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. Dans *Advances in Neural Information Processing Systems*, 2018.

[5] Jooyoung Lee, Seunghyun Cho, et Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. Dans *International Conference on Learning Representations*, 2019.

[6] David Minnen et Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. Dans *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343, 2020.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, et Neil Houlsby. An image is worth 16x16 words : Transformers for image recognition at scale. Dans *Inter. Conf. on Learning Representations*, 2021.

[8] Yichen Qian, Ming Lin, Xiuyu Sun, Tan Zhiyu, et Rong Jin. Entroformer : A transformer-based entropy model for learned image compression. Inter. Conf. on Learning Representations (ICLR), 02 2022.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, et B. Guo. Swin transformer : Hierarchical vision transformer using shifted windows. Dans *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society.

[10] Yinhao Zhu, Yang Yang, et Taco Cohen. Transformer-based transform coding. Dans *Inter. Conf. on Learning Representations*, 2022.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, et Illia Polosukhin. Attention is all you need. Dans I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et R. Garnett, éditeurs, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[12] Peter Shaw, Jakob Uszkoreit, et Ashish Vaswani. Self-attention with relative position representations. Dans *Proceedings of the 2018 Confe. of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468. Association for Computational Linguistics, Juin 2018.

[13] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, et Chunhua Shen. Conditional positional encodings for vision transformers. Dans *The Eleventh Inter. Conf. on Learning Representations*, 2023.

[14] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, et Lei Zhang. Cvt : Introducing convolutions to vision transformers. Dans *Proceedings of the IEEE/CVF Inter. Conf. on Computer Vision (ICCV)*, pages 22–31, October 2021.

[15] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, et Yichen Wei. Deformable convolutional networks. Dans *Proceedings of the IEEE inter. conf. on computer vision*, pages 764–773, 2017.

[16] Lu Chi, Borui Jiang, et Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33 :4479–4488, 2020.

[17] Kodak. Kodak test images. `http://r0k.us/graphics/kodak/`, 1999.

[18] Gisle Bjøntegaard. Calculation of average psnr differences between rd-curves. 2001.

[19] Johannes Ballé, Valero Laparra, et Eero P Simoncelli. End-to-end optimized image compression. *5th Inter. Conf. on Learning Representations (ICLR)*, 2017.

Étude comparative des méthodes de prédiction de l'échelle de débit basées sur l'apprentissage pour le streaming vidéo adaptatif

A. Telili[1], W. Hamidouche[1], S. Ahmed Fezza[2], L. Morin[1]

[1] Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
[2] National Higher School of Telecommunications and ICT, Oran, Algeria

Ce travail a été réalisé dans le cadre du projet DEEPTEC finance par la Région Bretagne.

Une version en anglais de 5 pages de cet article a été acceptée et publiée dans la conférence Picture Coding Symposium (PCS) 2022

https://ieeexplore.ieee.org/document/10018038

Exploring Temporal Consistency in Image-Based Rendering for Immersive Video Transmission

S. Lingadahalli Ravi[2], F. Henry[1], L. Morin[2], M. Gendrin[1]

[1]Orange Labs, 35510 Cesson Sévigné, France
[2]INSA Rennes – IETR, 35000 Rennes, France
slingada@insa-rennes.fr

This paper has been accepted and presented at the 10th European Workshop on Visual Information Processing held on 11th-14th September 2022, in Lisbon, Portugal.

https://ieeexplore.ieee.org/document/9922680

Sécurité

Analyse d'images secrètes bruitées

E. Reinders[1] B. Jansen van Rensburg[1,2] P. Puteaux[3] W. Puech[1]

[1] LIRMM, Univ. Montpellier, CNRS, Montpellier, France
[2] Stratégies, Rungis, France
[3] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Résumé

La protection de données multimédia est un sujet d'actualité crucial. L'un des moyens d'y parvenir est d'appliquer sur ces données des méthodes de partage de secret. Les parts ainsi générées et distribuées peuvent cependant être compressées par des algorithmes de compression avec pertes, tels que JPEG. Dans cet article, nous analysons l'impact d'une compression JPEG sur une part dans la reconstruction d'une image secrète. Pour cela, nous nous appuyons sur une méthode développée en 2022 par Bertojo et Puech [1] pour la correction d'images secrètes bruitées, reconstruites à partir de parts compressées avec JPEG.

Mots clefs

Sécurité multimédia, Partage d'image secrète, Compression JPEG, Reconstruction en présence de bruit.

1 Introduction

Dans un monde en constante évolution numérique, il est devenu courant de sauvegarder et de partager des contenus multimédia (images, vidéos et données 3D) sur différents appareils ou réseaux informatiques. La démocratisation de ces pratiques entraîne inévitablement des questionnements quant à la sécurisation de ces données multimédia. Des techniques comme le chiffrement, basé sur une ou plusieurs clés utilisées pour permuter et substituer les données à sécuriser, ou encore le tatouage, basé sur l'insertion de données cachées pour vérifier l'intégrité par exemple, ont vu le jour. Ces données multimédia représentent souvent de gros volumes d'informations à traiter pour les réseaux informatiques, et doivent par conséquent être compressées. Pour les images, le standard de compression est JPEG (*Joint Photographic Experts Group*) [2]. Une alternative au chiffrement ou au tatouage, est le partage de secret, qui permet de partager un secret entre n personnes, et de le reconstruire avec k parmi ces n personnes. Ainsi, en 1979, Shamir [3] et Blakley [4] ont proposé chacun leur propre schéma de partage de secret. La particularité de ces schémas est que, contrairement à la cryptographie, il n'y a plus de clés à échanger, mais des parts du secret.

Naor et Shamir [5] ont proposé en 1994 le premier schéma de partage d'image binaire, basé sur un schéma de cryptographie visuelle. Appliqué au partage d'image secrète, les parts s'apparentent à des images. En 2002, Thien et Lin [6] ont amélioré ce partage, afin de pouvoir l'appliquer à des images en niveaux de gris. Chaque valeur de pixels de l'image constitue un secret et doit appartenir à un corps fini $\mathbf{GF}(251)$, induisant un traitement spécifique sur les pixels de valeur supérieure à 250. La méthode de Yang *et al.* [7] a permis ensuite de s'affranchir de ce problème, en travaillant sur un corps fini $\mathbf{GF}(2^8)$.Même s'ils représentent une grande évolution dans le partage d'image secrète, les travaux effectués dans ce domaine ont pour inconvénient principal de ne pas prendre en compte des parts bruitées lors de la reconstruction. Il est en effet possible que certaines d'entre elles subissent des transformations, comme une compression JPEG. Bertojo et Puech [1] ont proposé une première analyse et une correction d'images secrètes bruitées à cause de l'utilisation de parts compressées avec JPEG lors de la reconstruction.

Dans cet article, nous analysons en détail la dégradation sur une image secrète reconstruite, induite par l'utilisation d'une part compressée avec JPEG. En section 2, nous analysons l'impact d'une part compressée avec JPEG sur la reconstruction d'une image secrète. En section 3, nous présentons des résultats expérimentaux issus d'une telle reconstruction. Enfin, en section 4, nous concluons sur le travail présenté, et discutons des perspectives d'amélioration.

2 Analyse

Dans cette section, nous analysons le modèle du bruit impactant une image secrète quand celle-ci est reconstruite à partir de k parts dont certaines sont bruitées. Le bruit subi par certaines parts peut être généré pendant une transmission sur des réseaux bas débit par exemple, ou à cause d'une compression avec pertes telle que JPEG.

2.1 Partage d'image avec Shamir

Pour pallier les problèmes induits par le chiffrement, Shamir a décrit en 1979 une méthode de partage de secret [3]. Sur la base de n utilisateurs, il a proposé de générer n parts d'un secret $\mathcal{S}$, afin qu'il soit possible de reconstruire le secret à partir de k parts, avec $1 < k \leq n$. Pour cela, Shamir suggère de partager des couples de valeurs $S_i = (x_i, y_i = f(x_i))$, avec $i \in \{1, ..., n\}$, les coordonnées d'un point d'un polynôme $f(.)$ de degré $k-1$. Ce polynôme est construit de telle sorte que le secret soit reconstruit pour $x = 0$, avec $f(0) = \mathcal{S}$. Shamir se base sur le partage d'une donnée entière non signée, définie sur un

corps fini $\mathbb{F}_p$ de cardinalité p. L'utilisation d'un tel anneau $\mathbb{Z}/p\mathbb{Z}$ permet une approximation sûre lors de la reconstruction du secret par la suite. Par conséquent, le polynôme de degré $k-1$ est généré à partir de $k-1$ coefficients aléatoires a_j, avec $j \in \{1, ..., k-1\}$, tels que $a_j < p$ et $a_0 = \mathcal{S}$:

$$f(x) = \mathcal{S} + \sum_{j=1}^{k-1} a_j \times x^j \mod p. \quad (1)$$

Afin de pouvoir reconstruire le secret $\mathcal{S} = a_0$ avec l'interpolation de Lagrange, au minimum k parts $S_i = (x_i, y_i)$ sont nécessaires :

$$f(x) = \sum_{i=1}^{k} y_i \times \prod_{u=1, u\neq i}^{k} \frac{x - x_u}{x_i - x_u} \mod p, \quad (2)$$

avec :

$$f(0) = a_0 = \sum_{i=1}^{k} y_i \times \prod_{u=1, u\neq i}^{k} \frac{x_u}{x_i - x_u} \mod p. \quad (3)$$

Le partage de secret peut s'étendre à du partage d'image secrète, en considérant une image secrète comme une matrice de pixels secrets. Les secrets deviennent alors les niveaux de gris (NDG), codés sur 8 bits, des pixels de l'image à partager. Pour ce faire, nous considérons un polynôme par pixel de l'image secrète. Soit une image secrète I, composée de $W \times H$ pixels $p_{w,h}$, avec $(w, h) \in [\![0, W[\![\times [\![0, H[\![$ leur position, nous générons alors n parts, sous la forme de couples de valeurs $S_i = (x_i = i, y_i = I_i)$, avec $i \in \{1, ..., n\}$ et I_i l'image part de même taille que l'image secrète. Ces parts peuvent se voir plus simplement comme des images I_i indexées de 1 à n. Pour I_i, à partir d'un pixel $p_{w,h}$ de I et de l'équation 1, nous générons n pixels $p_{w,h}(x_i)$ suivant :

$$p_{w,h}(x) = p_{w,h} + \sum_{j=1}^{k-1} a_{j,w,h} \times x^j \mod 251. \quad (4)$$

Nous prenons $p = 251$, car il s'agit du plus grand nombre premier inférieur à 2^8. Par conséquent, soit les NDG entre 251 et 255 ne sont pas traités et restent en clair dans les images parts, soit un pré-traitement de l'image secrète est nécessaire afin de ramener les valeurs des NDG entre 0 et 250.

2.2 Analyse du bruit induit dans une part par une compression JPEG

Une part I_i peut être bruitée, soit pendant une transmission bas débit ou sans fil, soit lors d'une compression avec pertes. Dans cette section nous analysons en particulier l'impact du bruit sur une telle part compressée avec JPEG [2], standard en matière de compression d'images. Lors d'une compression JPEG d'une image, après un changement d'espace couleur et d'éventuels sous-échantillonnages, les pixels, par bloc de 64, subissent une transformation DCT (*Discrete Cosine Transform*). Cette étape génère des coefficients DCT $F(u, v)$ pour chaque bloc, quantifiés selon des coefficients de quantification $q(u, v)$, dépendant d'un facteur de quantification QF compris entre 1 et 100 :

$$F'(u, v) = \left[\frac{F(u, v)}{q(u, v)}\right], \quad (5)$$

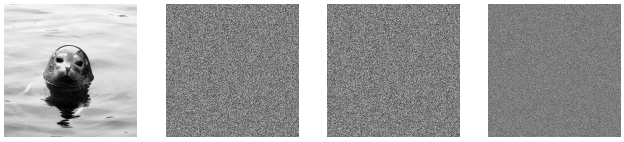
avec $u, v \in \{0, \ldots, 7\}$ et $[.]$ correspondant à l'entier le plus proche. C'est durant la phase de quantification que la perte par compression JPEG a principalement lieu. Un codage entropique est alors appliqué sur les coefficients quantifiés $F'(u, v)$, suivant un ordre zig-zag. Dans le cas $QF = 100$, l'image à compresser est faiblement impactée car tous les coefficients de quantification $q(u, v)$ valent 1, soit :

$$F'(u, v) = [F(u, v)]. \quad (6)$$

Afin d'analyser les effets d'une compression JPEG avec $QF = 100$, nous avons compressé 1000 parts générées à partir d'images de la base BOWS-2 [8]. Nous remarquons alors au décodage, que sur les 1000 images compressées, en moyenne $90,79\%$ des pixels restent inchangés ($0,04\%$ d'écart type), $4,59\%$ sont modifiés de +1 niveau de gris et $4,62\%$ de -1 niveau de gris (pour un écart type de $0,03\%$ dans les deux cas), soit un taux d'erreur $\tau = 9,21\%$. À partir d'un pixel $p_{w,h}(x_i)$ de la part I_i, nous obtenons alors un pixel bruité $p^*_{w,h}(x_i)$:

$$p^*_{w,h}(x_i) = p_{w,h}(x_i) + N_{w,h}, \text{ avec } N_{w,h} = \begin{cases} +1 \\ 0 \\ -1 \end{cases}, \quad (7)$$

tel que $Prob(N_{w,h} = 0) = 100 - \tau$, $Prob(N_{w,h} = +1) = Prob(N_{w,h} = -1) = \frac{\tau}{2}$.



(a) (b) (c) (d)

FIGURE 1 – *Compression JPEG avec QF = 100 sur une image de BOWS-2 [8] : a) Image secrète I, b) Part I_i, c) I_i compressée avec JPEG : I^*_i, d) Différence entre I_i et I^*_i.*

La figure 1 illustre le bruit induit dans une part par une compression JPEG. À partir de l'image secrète illustrée en figure 1.a, n parts sont générées, dont une est illustrée en figure 1.b. Ces parts sont semblables visuellement à des images de bruit aléatoire (PSNR avec l'image originale secrète de 6,74 dB, pour un SSIM de 0,01). Si la part illustrée figure 1.b est compressée avec JPEG (QF=100), nous obtenons une part bruitée très similaire (PSNR avec la figure 1.b de 58,35 dB, pour un SSIM de 0,99), comme illustrée en figure 1.c. En observant les différences entre la part originale et la part bruitée, illustrées figure 1.d, nous remarquons que la majorité des pixels n'est pas impactée par cette compression (NDG = 128), et que $9,18\%$ sont impactés de $+1$ (NDG = 255) ou de -1 (NDG = 0).

2.3 Impact d'une part bruitée sur la reconstruction d'une image secrète

En section 2.2, nous avons constaté que la compression JPEG avec $QF = 100$ d'une part introduisait un bruit. Utiliser des parts bruitées pour la reconstruction d'une image secrète risque donc d'induire des erreurs au moment de la reconstruction. Supposons que k' parts bruitées selon l'équation 7, avec $k' \leq k$, soient utilisées pour reconstruire

une image secrète. À la place de l'image secrète originale I, nous obtenons alors une image secrète bruitée I^*, composée de pixels $p^*_{w,h}$ reconstruits à partir de polynômes approximés $p^*_{w,h}(x)$:

$$\begin{aligned} p^*_{w,h} &= p^*_{w,h}(0) \\ &= \sum_{m=1}^{k'} p^*_{w,h}(x_m) \prod_{u=1,u\neq m}^{k} \frac{x_u}{x_u - x_m} \\ &+ \sum_{m=k'+1}^{k} p_{w,h}(x_m) \prod_{u=1,u\neq m}^{k} \frac{x_u}{x_u - x_m} \mod 251. \end{aligned} \tag{8}$$

Le bruit de reconstruction $\Delta^{err}_{w,h}$ est donc la différence entre $p_{w,h}(0)$ et $p^*_{w,h}(0)$:

$$\begin{aligned} \Delta^{err}_{w,h} &= p_{w,h}(0) - p^*_{w,h}(0) \\ &= \sum_{m=1}^{k'} N_{w,h}(x_m) \prod_{u=1,u\neq m}^{k} \frac{x_u}{x_u - x_m} \mod 251, \end{aligned} \tag{9}$$

avec $N_{w,h}(x_m)$ le bruit appliqué à la part I_m.
Dans cet article, nous limitons cette analyse à une seule image part bruitée ($k' = 1$) et une reconstruction avec 3 parts ($k = 3$). Nous définissons la part bruitée I_m, en position m, et les deux parts positionnées de part et d'autre de I_m à la distance $\pm\delta$: $I_{m-\delta}$ et $I_{m+\delta}$. Le bruit de reconstruction Δ^{err} appliqué sur l'image reconstruite I^*, représenté par l'équation 9, se ramène alors à :

$$\Delta^{err} = N_{w,h} \times \left(\frac{1 - m^2}{\delta^2}\right) \mod 251, \tag{10}$$

avec $N_{w,h}$ le bruit subi par l'image part I_m, comme défini dans l'équation 7.

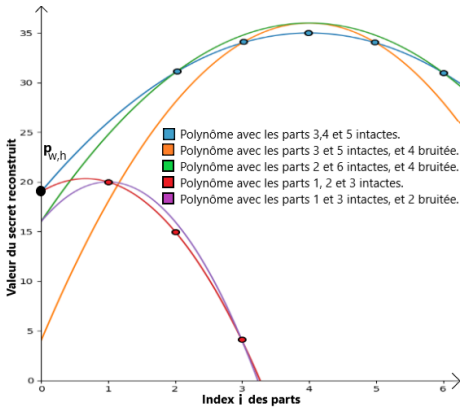


FIGURE 2 – *Impact de la valeur de m et de δ dans la reconstruction du secret.*

Prenons $p_{w,h} = 19$. Dans la figure 2, nous construisons un polynôme $f(.)$ de degré deux (courbe bleue), à partir des parts $p_{w,h}(3) = 34$, $p_{w,h}(4) = 35$ et $p_{w,h}(5) = 34$. Supposons désormais la part $p^*_{w,h}(4) = 36$ bruitée de $+1$, le polynôme orange est alors obtenu. D'après l'équation 10, nous nous retrouvons avec un scénario de reconstruction $m = 4$, $\delta = 1$ et $N_{w,h}(4) = 1$. La valeur $p^*_{w,h}$ reconstruite est de 4, soit $\Delta^{err}_{w,h} = 15$. La courbe verte correspond à une reconstruction avec $p^*_{w,h}(4) = 36$ et $p_{w,h}(2) = p_{w,h}(6) = 31$, soit $m = 4$, $\delta = 2$ et $N_{w,h}(4) = 1$. Dans ce cas, $\Delta^{err}_{w,h} = 3$.

Nous remarquons que quand δ croit, alors le bruit de reconstruction s'atténue. Nous construisons un autre polynôme pour le même secret $p_{w,h} = 19$, mais en utilisant les parts $p_{w,h}(1) = 20$, $p_{w,h}(2) = 15$ et $p_{w,h}(3) = 4$ (courbe rouge). Supposons que nous nous retrouvons à nouveau dans une reconstruction incluant une part bruitée, avec $m = 2$, $\delta = 1$ et avec le même bruit tel que $p^*_{w,h}(2) = p_{w,h}(2) + 1$. Nous obtenons alors la courbe violette, et nous remarquons que reconstruire à partir d'une part bruitée en position plus faible (m), atténue la valeur du bruit à la reconstruction du secret. Pour un m et un δ donnés, nous obtenons alors cinq scénarios possibles de bruit pour $p^*_{w,h}$:

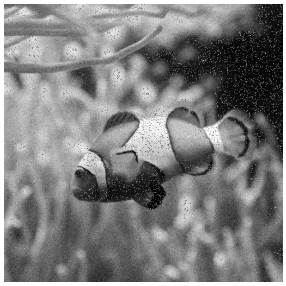
$$\begin{cases} p_{w,h} & \text{si } N_{w,h} = 0, \\ p_{w,h} + \Delta^{err}_{w,h} & \text{si } p_{w,h} < 251 - \Delta^{err}_{w,h} \text{ et } N_{w,h} = 1, \\ p_{w,h} + (\Delta^{err}_{w,h} - 251) & \text{si } p_{w,h} \geq 251 - \Delta^{err}_{w,h} \text{ et } N_{w,h} = 1, \\ p_{w,h} - (\Delta^{err}_{w,h} - 251) & \text{si } p_{w,h} < \Delta^{err}_{w,h} \text{ et } N_{w,h} = -1, \\ p_{w,h} - \Delta^{err}_{w,h} & \text{si } p_{w,h} \geq \Delta^{err}_{w,h} \text{ et } N_{w,h} = -1. \end{cases} \tag{11}$$

Nous nous retrouvons avec un bruit à la reconstruction de type **poivre et sel mod 251**.

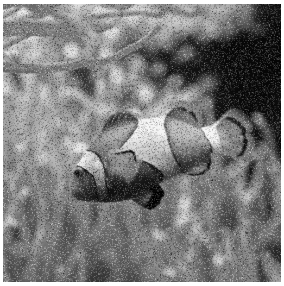
3 Résultats

(a) Image secrète
PSNR = +∞, SSIM = 1,0

(b) Δ^{err}=3 (m=8,δ=4)
PSNR = 43,89 dB, SSIM = 0,98



(c) Δ^{err}=48 (m=7,δ=1)
PSNR = 20,22 dB, SSIM = 0,38



(d) Δ^{err}=120 (m = 11,δ=1)
PSNR = 16,03 dB, SSIM = 0,14

FIGURE 3 – *Reconstructions d'une image secrète, issue de BOWS-2 [8], pour différentes valeurs de* Δ^{err} *(*$\tau = 9,21\%$*) : a) Image secrète originale, b) c) d) Images secrètes reconstruites pour différentes valeurs de* Δ^{err}.

La figure 3 illustre la reconstruction d'une image secrète selon différentes configurations possibles de m et δ, avec en figure 3.a la reconstruction sans part bruitée (image secrète originale). Nous remarquons figure 3.b que pour des valeurs faibles de bruit Δ^{err} à la reconstruction ($\Delta^{err} = 3$), l'image reconstruite est visuellement très similaire à l'image secrète originale. Seules quelques variations vi-

sibles s'opèrent au moment d'ajouter (ou retrancher) 3 à un pixel bruité de valeur d'intensité supérieure à 248, ou inférieure à 3. D'après les scénarios possibles de bruit, présentés dans l'équation 11, nous nous retrouvons dans le cas d'un scénario de bruit décrit ligne 3 ou 4, entraînant de fortes variations entre pixels originaux et bruités. Pour une configuration induisant un bruit Δ^{err} plus important à la reconstruction de l'image secrète (illustré figure 3.c et 3.d), un pixel reconstruit bruité $p^*_{w,h}$ est plus fortement impacté pour un scénario de bruit décrit ligne 2 ou 5. Un scénario de bruit décrit ligne 3 ou 4 a, pour cette configuration, une plus grande probabilité d'apparition, mais ne va pas faire autant varier la valeur du pixel bruité $p^*_{w,h}$, que pour une configuration de bruit Δ^{err} plus faible. Cela a pour conséquence directe de limiter l'apparition du phénomène de forte variation entre pixels originaux et bruités dans l'image reconstruite.

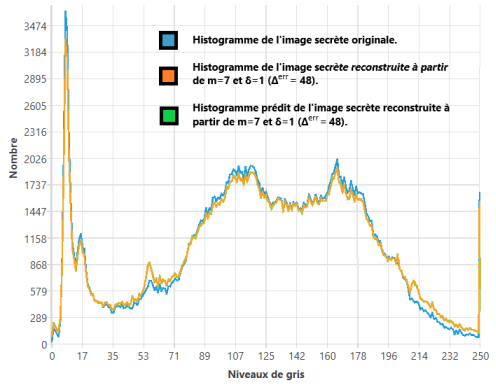


FIGURE 4 – *Exemple de prédiction d'histogramme d'une image secrète reconstruite pour la configuration m = 7, δ = 1 ($\Delta^{err} = 48$) et τ = 9,21%.*

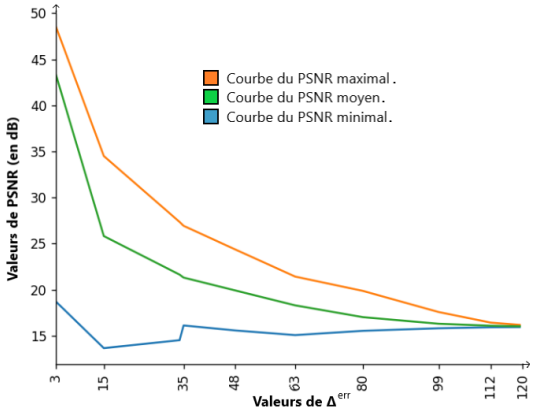


FIGURE 5 – *Variation du PSNR en fonction de la valeur Δ^{err} sur 200 images de BOWS-2 [8] pour un τ = 9,21%.*

À partir de m et δ, nous connaissons donc le bruit appliqué à la reconstruction d'une image secrète. Nous pouvons par conséquent prédire l'histogramme de l'image reconstruite. En effet, comme illustré figure 4, partant de l'histogramme de l'image secrète (courbe bleue), supposons que $\frac{\tau}{2}\%$ des pixels de I^* sont bruités de $+\Delta^{err}_{w,h}$, $\frac{\tau}{2}\%$ de $+(251 - \Delta^{err}_{w,h})$ et $(100 - \tau)\%$ restent inchangés : nous obtenons la courbe verte. Celle-ci correspond assez fidèlement à l'histogramme réellement obtenu après reconstruction sur une telle configuration (courbe orange).
Sur la figure 5, les courbes bleue,verte et orange, représentent respectivement les valeurs de PSNR minimales, moyennes et maximales obtenues avec les valeurs de $\Delta^{err} = \{3, 15, 34, 35, 48, 63, 80, 99, 112, 120\}$, sur 200 images de BOWS-2 [8]. Nous remarquons que plus la valeur de Δ^{err} augmente, et plus la qualité de l'image reconstruite, par rapport à l'image secrète originale, diminue.

4 Conclusion

Dans cet article, nous avons proposé une analyse de la reconstruction d'une image secrète à partir d'une part compressée avec JPEG ($QF = 100$), et de deux autres parts intactes ($k = 3$). Une compression JPEG ($QF = 100$) impacte la part compressée, induisant des erreurs de reconstruction dans l'image secrète. Dépendant des index m et δ des parts utilisées au moment de reconstruire ce secret, le bruit induit à la reconstruction (bruit **poivre et sel mod 251**) varie et impacte plus ou moins fortement l'image secrète reconstruite. Nous avons également démontré qu'il est possible de prédire l'histogramme d'une image reconstruite bruitée, dès lors que nous connaissons la configuration de bruit appliquée, en fonction des parts utilisées pour la reconstruction. Cette prédiction d'histogramme peut être utile pour nous guider dans la reconstruction de l'image secrète.
En perspectives, il serait intéressant de chercher à débruiter ces images reconstruites, par emploi de réseaux de neurones entraînés sur ce type de bruit par exemple.

Références

[1] L. Bertojo et W. Puech. Correction of Secret Images Reconstructed from Noised Shared Images. Dans *IEEE IPTA*, pages 1–6, 2022.

[2] G. K. Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1) :XVIII–XXXIV, 1992.

[3] A. Shamir. How to share a secret. *Communications of the ACM*, 22(11) :612–613, 1979.

[4] G. R. Blakley. Safeguarding cryptographic keys. *Proceedings of the National Computer Conference*, 48 :313–317, 1979.

[5] M. Naor et A. Shamir. Visual cryptography. Dans *Workshop on the Theory and Application of of Cryptographic Techniques*, pages 1–12. Springer, 1994.

[6] C. C. Thien et J. C. Lin. Secret image sharing. *Computers & Graphics*, 26(5) :765–770, 2002.

[7] C. N. Yang. New visual secret sharing schemes using probabilistic method. *Pattern Recognition Letters*, 25(4) :481–494, 2004.

[8] P. Bas et T. Furon. Image database of BOWS-2. http ://bows2.eclille.fr/.

Méthode Jointe de Tatouage et Compression Draco pour les Objets 3D

B. Jansen van Rensburg[1,2] A. G. Bors[3] W. Puech[1] *

[1] LIRMM, Univ. Montpellier, CNRS, Montpellier, France
[2] Stratégies, Rungis, France
[3] University of York, York, Royaume-Uni

Résumé

De nos jours, les objets 3D jouent un rôle de plus en plus important dans différents domaines, surtout avec le développement du métaverse. Ces objets 3D sont parfois très grands et nécessitent énormément ressources pour le stockage et la transmission. La compression des objets 3D peut résoudre ce problème. Les objets 3D doivent également être sécurisés pendant leur stockage et leur transmission. Des mesures de sécurité, comme les droits d'auteurs et l'authentification sont donc essentielles. Dans cet article, nous proposons une méthode jointe de tatouage et compression des objets 3D basée sur Draco. Pour cela, nous proposons d'intégrer une étape de tatouage pendant la compression Draco, méthode de compression 3D proposée par Google, qui devient actuellement le standard industriel. Notre méthode conjointe permet d'obtenir une grande capacité d'insertion pouvant être utilisée pour la protection des droits d'auteurs par exemple.

Mots clefs

Tatouage 3D, sécurité multimédia, compression Draco, sécurité des objets 3D, compression des objets 3D.

1 Introduction

De nos jours, les objets 3D sont utilisés dans de nombreux domaines de la vie quotidienne. Ils sont souvent stockés et partagés en ligne au cours de leur existence. La plupart du temps, ces objets 3D sont constitués de millions de sommets, et donc très consommateurs en termes de temps et de ressources. Dans un environnement industriel, ces objets 3D sont des biens importants, et permettent de représenter des données confidentielles. Il est donc essentiel que ces objets 3D soient à la fois compressés et sécurisés pendant leurs transmissions et leurs archivages.

Il existe deux catégories principales de méthodes de sécurisation des objets 3D. La première étant le chiffrement, qui rend les objets 3D illisibles avec une clé secrète afin de protéger leur confidentialité visuelle. La deuxième est l'insertion des données cachées ou tatouage, qui insère des données dans un objet 3D de façon invisible. Le tatouage peut être utilisé pour insérer les droits d'auteurs, des données avec une haute capacité, ou assurer l'intégrité d'un objet 3D.

Pour la compression d'objets 3D, en 1999, Rossignac a proposé Edgebreaker, une méthode qui compresse la connectivité d'un objet 3D [1]. Cette méthode parcourt les triangles d'un objet 3D avec un depth-first spiralling spanning tree. En 2014, Google a proposé le schéma Draco pour la compression des objets 3D, qui devient le standard industriel [2]. En 2019, Cao *et al.* a établi un état de l'art sur la compression des nuages des points [3]. Concernant le tatouage d'objets 3D, en 2007, Cho *et al.* ont proposé une méthode statistique, où les données sont insérées dans la distribution des normales des sommets [4]. En 2010, Wang *et al.* ont créé un référence pour le tatouage 3D [5]. En 2013, Bors et Luo ont proposé une méthode de tatouage qui minimise la distorsion de la surface [6]. En 2018, Zhang *et al.* ont proposé une méthode de tatouage réversible basée sur l'expansion et le tri des erreurs de prédictions [7].

Les méthodes jointes de tatouage et compression sont des challenges en image [8] et en 3D [9, 10]. En effet, ces deux codages modifient les mêmes domaines de la représentation des données 3D et donc s'interfèrent entre eux. En 2009, Abdallah *et al*. ont développé une méthode jointe de tatouage et compression où les données sont insérées dans le domaine spectral d'un sous-maillage Laplacien [9]. En 2011, Lee *et al.* ont décrit une méthode jointe de tatouage et compression progressive où les données sont insérées dans chaque niveau de détail d'un objet 3D [10]. Alors que Jansen van Rensburg *et al.* ont proposé une méthode de crypto-compression basée sur Draco [11], à ce jour il n'existe aucune méthode de tatouage pour les objets 3D compressés avec Draco.

Dans cet article, nous proposons une méthode jointe de tatouage et compression pour les objets 3D basée sur Draco. Nous intégrons une étape de tatouage entre l'étape de quantification des sommets et l'étape de prédiction des sommets, pendant la phase d'encodage de Draco. L'extraction des données cachées est effectuée pendant la phase de décodage de Draco, avant que les sommets ne soient reconstruits.

Cet article est organisé dans la manière suivante. En section 2, nous décrivons la méthode jointe de tatouage et compression Draco. En section 3, nous présentons des ré-

*Nous remercions le GdR ISIS, CNRS, pour avoir financé cette collaboration.

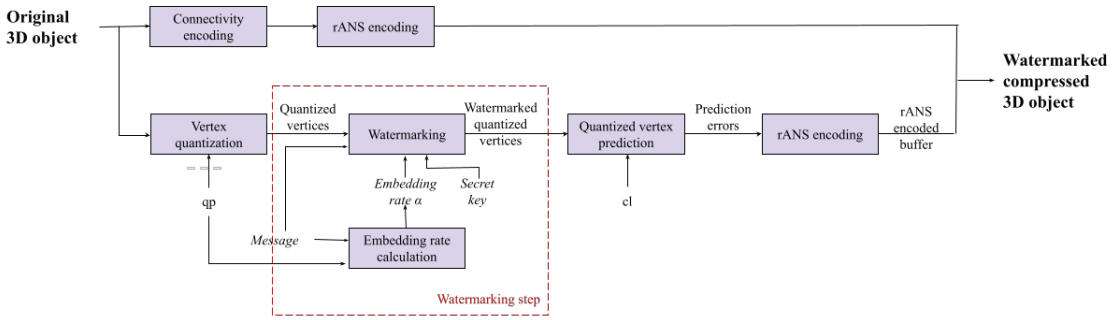


FIGURE 1 – *Vue globale de la phase d'encodage de la méthode jointe de tatouage et compression.*

sultats expérimentaux. Enfin, en section 4, nous concluons nos travaux.

2 La Méthode Jointe Proposée

Dans cette section, nous détaillons la méthode jointe de tatouage et compression des objets 3D. La figure 1 présente une vue globale de la phase d'encodage de la méthode proposée. L'étape de tatouage est intégrée dans la phase d'encodage de la géométrie après que les sommets soient quantifiés. Après l'étape de tatouage, l'étape de prédiction des sommets quantifiés permet de préserver les données cachées. En section 2.1, nous décrivons la méthode de compression des objets 3D par Draco. En section 2.2, nous décrivons la phase de tatouage qui est effectuée pendant la phase d'encodage de Draco. En section 2.3, nous détaillons l'étape d'extraction des données qui est effectuée pendant la phase de décodage de Draco.

2.1 Compression des objets 3D par Draco

Comme illustré dans la figure 1, la méthode de compression Draco a deux phases principales d'encodage : l'encodage de la connectivité, basée sur la méthode Edgebreaker [1], et l'encodage de la géométrie. Ceux deux phases sont effectuées séparément. Dans cet article, nous proposons d'intégrer une étape de tatouage dans la phase d'encodage de la géométrie, sans modifier la phase d'encodage de la connectivité.

Dans la phase d'encodage de la géométrie, les coordonnées x, y et z de chaque sommet sont quantifiées selon le paramètre de quantification de Draco $qp \in [0, 30]$. A part pour $qp = 0$, signifiant qu'il n'y a pas de quantification, chaque coordonnée c, composée d'un point flottant de 32 bits, est transformée en un entier non-signé c' de qp bits. Une étape de prédiction des sommets est alors effectuée sur les coordonnées quantifiées $c' \in [0, 2^{qp}]$ selon le paramètre de compression de Draco $cl \in [0, 10]$. Nous remarquons que la valeur de qp est un compromis entre le taux de compression et la qualité de l'objet 3D reconstruit. Le système d'encodage entropique range Asymmetric Numeral System (rANS) [12] est appliqué après l'encodage de la connectivité et l'encodage de la géométrie, afin de fournir un objet 3D compressé au format Draco.

2.2 Tatouage dans la phase d'encodage

Nous proposons d'intégrer une étape de tatouage dans la phase d'encodage de Draco. Ceci ne modifie pas l'encodage de la connectivité, qui est effectué séparément. Plus précisément, le tatouage est intégré entre l'étape de quantification des sommets et l'étape de prédiction des sommets. Premièrement, comme illustré en figure 1, la longueur du message à insérer est calculée afin de déterminer le taux d'insertion par coordonnée qui est limité par la taille de la coordonnée quantifiée.

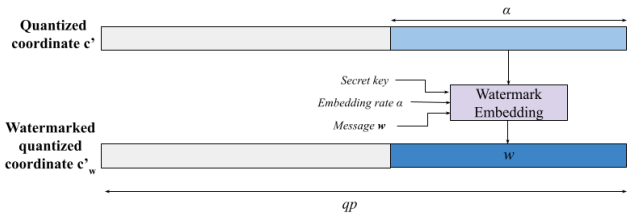


FIGURE 2 – *Etape de tatouage pour une coordonnée quantifiée c'.*

La figure 2 présente l'étape de tatouage pour une coordonnée quantifiée c', qui a une taille de qp bits. Chaque c' est tatouée avec un taux d'insertion de α bits par substitution des LSB, selon une clé secrète définissant l'ordre d'insertion :

$$c'_w = \left\lfloor \frac{c'}{2^\alpha} \right\rfloor \times 2^\alpha + w, \tag{1}$$

où c'_w est la coordonnée tatouée, et w est le message de α bits à insérer dans c'.

Après l'étape de tatouage, les coordonnées quantifiées et marquées subissent une étape de prédiction des sommets. Les erreurs de prédiction sont encodées avec la méthode d'encodage entropique rANS. Nous obtenons alors un objet 3D tatoué et compressé dans le format Draco.

2.3 Phase d'extration et de décodage

La figure 3 présente une vue globale de la phase de décodage de la méthode jointe de tatouage et compression. Il est possible d'extraire le message à deux instants différents, soit après la prédiction et la reconstruction des som-

mets pendant la phase de décodage de Draco, soit dans le domaine spatial de l'objet 3D reconstruit après le décodage de Draco.

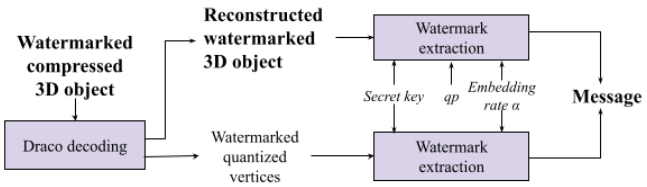


FIGURE 3 – *Vue globale de la phase de décodage de la méthode jointe de tatouage et compression.*

D'abord, l'objet 3D compressé et marqué est décodé avec le décodeur rANS afin de reconstruire les erreurs de prédiction des sommets. Les sommets tatoués sont alors reconstruits, et le message peut être extrait. Le message w est extrait en lisant les α LSB de chaque coordonnée reconstruite, où α est le taux d'insertion :

$$w = c'_w \bmod 2^\alpha, \tag{2}$$

où c'_w est une coordonnée reconstruite et tatouée, de taille de qp bits.

Après l'extraction du message, les sommets sont déquantifiés afin de retrouver des valeurs flottantes et de compléter le décodage de Draco permettant d'obtenir un objet 3D tatoué. A partir du paramètre de quantification qp, le message peut également être extrait de l'objet 3D tatoué.

3 Résultats Expérimentaux

Dans cette section, nous présentons des résultats expérimentaux de notre méthode jointe de tatouage et compression Draco. Dans un premier temps, nous appliquons notre méthode à l'objet 3D *Bunny* de la base de Stanford [13], comme illustré en figure 4. Notons que α est le taux d'insertion en bits par coordonnée quantifiée ($\alpha = 0$ indique qu'il n'y a aucun tatouage).

La figure 4 présente les résultats visuels de la méthode jointe de tatouage et compression proposée appliquée sur *Bunny* pour plusieurs combinaisons de taux d'insertion α et de paramètre de quantification qp. Le taux d'insertion α doit forcément être plus petit que qp, car qp est la taille d'une coordonnée. Quand qp augmente, α peut aussi augmenter, car plus de LSB sont rendus disponibles pour le tatouage. Quand la valeur de α se rapproche de celle de qp, la dégradation visuelle augmente. Nous observons en figure 4 que la dégradation n'est pas visible quand $(qp - \alpha) \geq 7$, ce qui confirme les résultats présentés par [14].

La figure 5(a) présente le RMSE moyen avec l'écart type entre les objets 3D originaux de la base de Stanford [13] et les objets 3D correspondants décodés et tatoués avec la méthode proposée en fonction du taux d'insertion α pour plusieurs valeurs de qp. Nous remarquons que l'écart type est négligeable et donc le RMSE ne varie pas énormément entre les différents objets 3D. Nous pouvons observer que

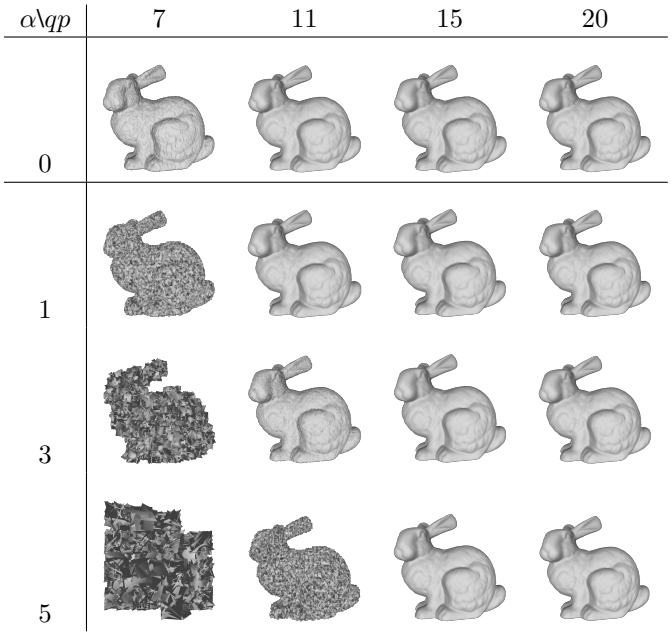
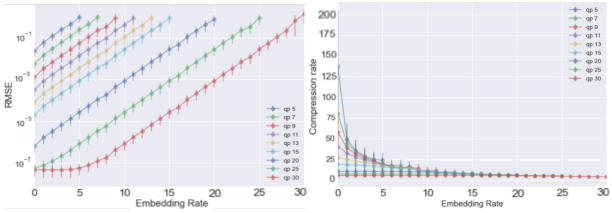


FIGURE 4 – *Résultats visuels de la méthode proposée appliquée à Bunny, pour plusieurs taux d'insertion α (bits par coordonnée quantifiée) et valeurs de paramètre de quantification de Draco qp.*

le RMSE ne dépend pas uniquement du paramètre de quantification qp, mais de la relation entre qp et α, *ie.* le nombre de MSB non modifié. En effet, quand $(qp - \alpha) \geq 7$, le RMSE est négligeable car il a toujours un ordre de 10^{-3}, quelque soit la valeur de qp. Ceci confirme les résultats visuels illustrés en figure 4 pour l'objet 3D *Bunny*.

La figure 5(b) présente les taux de compression des objets 3D de la base de Stanford [13] tatoués et compressés avec la méthode proposée en fonction du taux d'insertion α pour plusieurs valeurs de qp. Nous remarquons que $\alpha \in [0, qp]$, où $\alpha = 0$ correspond aux objets 3D compressés sans tatouage. Nous pouvons observer que quand qp augmente, le taux de compression diminue. Ceci est lié à la méthode originale de Draco, illustrée quand $\alpha = 0$. Ce comportement est prédictible, car qp est la taille d'une coordonnée après la quantification. La perte du taux de compression devient également moins significative, donc le changement du taux de compression entre les objets 3D est mineur, particulièrement quand qp augmente.

Comme pour la méthode de compression originale de Draco, un compromis est à trouver entre le taux de compression et la qualité de l'objet 3D reconstruit. La figure 6 présente le taux de compression moyen en fonction du RMSE moyen pour plusieurs valeurs de $qp \in [0, 30]$ pour la base de Stanford [13]. Nous pouvons observer que le changement du paramètre de quantification qp a une grande influence sur le taux de compression et une influence mineure sur le RMSE. Par conséquent, la valeur de qp doit être choisie selon le taux d'insertion souhaité. Par exemple, si le taux d'insertion souhaité est $\alpha = 1$ bit par coordonnée



(a) RMSE (b) Taux de compression

FIGURE 5 – *RMSE moyen et taux de compression obtenu pour la base de Stanford [13] tatouée et compressée avec la méthode proposée en fonction du taux d'insertion α pour plusieurs $qp \in [1, 30]$.*

(donc 3 bits par sommet), et l'utilisateur souhaite minimiser la distorsion, alors $qp = 15$ est recommandé, car le RMSE moyen est 5.22×10^{-5}, pour un taux de compression de 18.93. Si, par exemple, le taux d'insertion souhaité est $\alpha = 3$ bits par coordonnée (donc 9 bits par sommet), un bon compromis entre le taux de compression et la qualité est $qp = 11$. En effet avec $qp = 11$, paramètre de défaut de Draco donné par Google, nous avons un taux de compression de 24.38 et un RMSE moyen de 3.47×10^{-3}. En conclusion, pour ces deux exemples ($qp = 15, \alpha = 1$ et $qp = 11, \alpha = 3$), nous observons en figure 4 que ces objets 3D reconstruits ont une haute qualité sans dégradation visuelle.

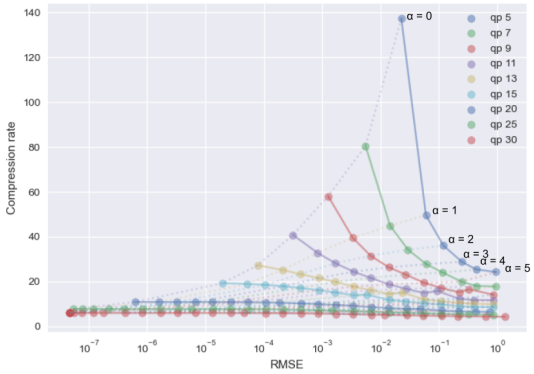


FIGURE 6 – *Taux de compression moyen en fonction du RMSE pour plusieurs qp et α, appliqué à la base de Stanford [13].*

4 Conclusion

Dans cet article, nous proposons une méthode jointe de tatouage et compression Draco des objets 3D. Une étape de tatouage est intégrée dans la phase d'encodage de géométrie de Draco. Le message caché peut être extrait soit pendant la phase de décodage de Draco, soit après la reconstruction de l'objet 3D décodé et marqué. Nos résultats expérimentaux montrent un taux d'insertion élevé qui peut être utilisé, par exemple, pour insérer des droits d'auteur. Nous avons proposé une valeur optimale des paramètres dans des scénarios différents. Nous avons également testé notre méthode sur une grande base de données. Dans le futur, nous souhaitons maximiser la qualité de l'objet 3D reconstruit ou augmenter la capacité d'insertion en conservant le taux de compression original de Draco.

Références

[1] J. Rossignac. 3D compression made simple : Edgebreaker with ZipandWrap on a corner-table. Dans *Proceedings International Conference on Shape Modeling and Applications*, pages 278–283, 2001.

[2] Google. Draco 3D graphics compression, 2014.

[3] C. Cao, M. Preda, et T. Zaharia. 3D point cloud compression : A survey. Dans *The 24th International Conference on 3D Web Technology*, pages 1–9, 2019.

[4] J.-W. Cho, R. Prost, et H.-Y.Jung. An Oblivious Watermarking for 3-D Polygonal Meshes Using Distribution of Vertex Norms. *IEEE Transactions on Signal Processing*, 55 :142 – 155, 02 2007.

[5] K. Wang, G. Lavoué, F. Denis, A. Baskurt, et X. He. A Benchmark for 3D Mesh Watermarking. Dans *2010 Shape Modeling International Conference*, pages 231–235, 2010.

[6] A. G. Bors et M. Luo. Optimized 3D Watermarking for Minimal Surface Distortion. *IEEE Transactions on Image Processing*, 22(5) :1822–1835, 2013.

[7] Q. Zhang, X. Song, T. Wen, et C. Fu. Reversibility improved data hiding in 3D mesh models using prediction-error expansion and sorting. *Measurement*, 135 :738–746, 2019.

[8] D. Goudia, M. Chaumont, W. Puech, et N. Said. Tatouage et Compression Conjoint dans JPEG2000 avec un Algorithme de Quantification Codée par Treillis (TCQ). Dans *CORESA*, 2010.

[9] E. Abdallah, A. Hamza, et P. Bhattacharya. Watermarking 3D models using spectral mesh compression. *Signal, Image and Video Processing*, 3 :375–389, 10 2009.

[10] H. Lee, Ç. Dikici, G. Lavoué, et F. Dupont. Joint reversible watermarking and progressive compression of 3D meshes. *The Visual Computer*, 27 :781–792, 2011.

[11] B. Jansen van Rensburg, W. Puech, et J.-P. Pedeboy. The First Draco 3D Object Crypto-Compression Scheme. *IEEE Access*, 10 :10566–10574, 2022.

[12] J. Duda, K. Tahboub, N. J. Gadgil, et E. J. Delp. The use of asymmetric numeral systems as an accurate replacement for Huffman coding. Dans *Picture Coding Symposium (PCS)*, pages 65–69, 2015.

[13] M. Levoy, J. Gerth, B. Curless, et K. Pull. The Stanford 3D scanning repository. *URL http ://graphics.stanford.edu/data/3Dscanrep/*, 5(10), 2005.

[14] B. Jansen van Rensburg, W. Puech, et J.-P. Pedeboy. Draco-Based Selective Crypto-Compression Method of 3D objects. Dans *IEEE IPTA 2022*, pages 1–6, 2022.

Stéganographie robuste et sans erreur dans des images JPEG en utilisant les sorties des codeurs JPEG

J. Butora, P. Puteaux, P. Bas

Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Résumé d'un article soumis dans la revue internationale
IEEE Transactions on Dependable and Secure Computing (TDSC),
dont une version préliminaire est disponible sur arXiv :

https://arxiv.org/pdf/2211.04750.pdf

Applications

Détection d'anomalies dans des vidéos acquises par drone pour la maintenance préventive de lignes électriques

Guillaume Fourret[1,3] Christophe Fiorio[1] Gérard Subsol[1] Marc Chaumont[1,2] Samuel Brau[3]

[1] Équipe ICAR, LIRMM, Univ. Montpellier, CNRS, Montpellier, France
[2] Univ. Nîmes, France
[3] Drone Geofencing, Nîmes, France

{guillaume.fourret, fiorio, gerard.subsol, marc.chaumont}@lirmm.fr
samuel.brau@drone-geofencing.com

Résumé

Les lignes électriques sont constituées de plusieurs composants susceptibles de se déteriorer au fur et à mesure de leurs utilisation. Pour détecter des possibles problèmes sur ces objets et ainsi prévenir des pannes coûteuses sur le réseau, les drones aériens sont de plus en plus utilisés car ils permettent d'inspecter rapidement de grandes distances, et d'avoir un bon angle de vue sur les différents composants et leurs défauts. Cependant, l'analyse des vidéos de vol par des experts est laborieuse. De plus, le nombre d'anomalies différentes pouvant provoquer une coupure du réseau est grand et faire une solution automatique dédiée pour chacune n'est pas envisageable, notamment à cause du faible nombre d'exemples pour chacune d'entre elles. De ce fait, nous nous sommes tournés vers une solution non supervisée et générique permettant d'attribuer un score d'anomalie aux objets sans nécessiter d'informations à priori sur l'apparence des défauts.

Mots clefs

Détection d'objets, Détection d'anomalies, Deep Learning, Drone aérien, Vidéo.

1 Introduction

Les coupures dues à des défauts des composants des lignes électriques coutent très cher aux fournisseurs d'électricité. De nombreuses causes de dysfonctionnement sont possibles, allant de la végétation alentour jusqu'à la rupture des composants. La maintenance préventive est donc essentielle pour ces entreprises afin d'anticiper ces pannes.

L'inspection de lignes électriques par drone est utilisée couramment mais elle nécessite une inspection visuelle par des techniciens et pourrait être automatisée [1, 2, 3]. Par exemple, l'analyse automatique d'un des composants principaux, l'isolateur, a été étudiée notamment dans le cas de l'absence d'un disque [4, 5]. Mais ces méthodes sont dédiées à l'analyse d'un objet connu à l'avance et surtout d'une anomalie bien identifiée. Or, une anomalie peut avoir un aspect différent en fonction de la condition de prise de vue (distance, éclairage...). Des méthodes récentes de détection d'anomalies plus génériques utilisent des réseaux de neurones de classification entraînés sur des très grandes bases de données d'objets présentant ces anomalies [6]. Or, les anomalies sont, par définition, rares et la collecte d'une base de données permettant un entraînement devient une tâche compliquée, c'est pourquoi des solutions non-supervisées ont été proposées. Ces algorithmes s'intéressent à la classification de données extrêmes (aussi appelées "outliers") qui se démarquent par rapport à une distribution que suivent la majorité des données. Par exemple, dans [7], un descripteur local est calculé pour chaque disque d'une chaîne d'isolateurs, et la détection d'une anomalie se fait de manière non supervisée en utilisant le Local Outlier Factor (LOF) pour comparer les descripteurs entre eux.

2 Description de notre méthode

Nous proposons dans cet article une chaîne de traitement visant à proposer un score d'anomalie à tous les objets détectés d'une ligne, sans s'intéresser à une anomalie précise. Pour notre méthode non-supervisée, nous posons l'hypothèse qu'une anomalie est par définition rare. La majorité des objets que nous allons détecter sur une ligne électrique sera donc en bon état, et les anomalies se démarqueront. Notre idée générale est donc de comparer après le vol tous les objets d'une même classe entre eux pour voir lesquels se distinguent. Une étape préalable de détection des objets est donc indispensable. Pour cela, nous utilisons un réseau détecteur d'objets afin d'obtenir les imagettes des objets durant le vol. Ensuite, nous utilisons un réseau CNN afin d'extraire un vecteur caractéristique de chaque imagette, permettant de décrire l'objet de manière efficace. Enfin, nous comparons ces vecteurs pour trouver lesquels se démarquent.

2.1 Détection d'objets

Nous nous sommes tournés vers Yolov5 [8] car c'est un détecteur assez précis, léger et rapide en inférence pour tourner potentiellement sur des cartes embarquées dans les

drones.

2.2 Extraction des vecteurs caractéristiques

Nous calculons pour toutes les imagettes résultantes de l'étape de détection un vecteur caractéristique servant à décrire l'apparence de l'objet. Celui-ci est obtenu en utilisant la dernière couche convolutionnel d'un réseau de neurones VGG16 [9], pré-entraîné sur ImageNet. Les réseaux ResNet [10] et EfficientNet [11] ont également été testés comme extracteur de caractéristiques, mais VGG16 a été retenu car il nous donne expérimentalement les meilleurs résultats de discrimination des anomalies.

2.3 Score d'anomalie

À partir de la séquence de vol, nous obtenons une série de vecteurs caractéristiques $\boldsymbol{X}_i$ extraits des imagettes venant de Yolo. Certains de ces vecteurs représentent potentiellement un objet présentant une anomalie. Nous utilisons le Local Outlier Factor [12] (LOF) comme algorithme de détection d'anomalies non supervisé.
Pour cela, on utilise la notion de "densité local atteignable" (local reachability density) par rapport à ses k plus proches voisins définie comme suit (1) :

$$lrd_k(\boldsymbol{X}_i) = 1/(\frac{\sum_{\boldsymbol{X}_j \in N_k(\boldsymbol{X}_i)} d_k(\boldsymbol{X}_i, \boldsymbol{X}_j)}{|N_k(\boldsymbol{X}_i)|}) \qquad (1)$$

$d_k(\boldsymbol{X}_i, \boldsymbol{X}_j)$ est la reachability-distance qui représente le maximum entre la distance euclidienne de $\boldsymbol{X}_i$ et $\boldsymbol{X}_j$, et la distance euclidienne entre $\boldsymbol{X}_j$ et son k-ème plus proche voisin. $N_k(\boldsymbol{X}_i)$ représente l'ensemble des k plus proches voisins de $\boldsymbol{X}_i$.
A partir de cette densité local atteignable, nous pouvons calculer le Local Outlier Factor d'un vecteur (2) qui correspond au ratio entre la densité local atteignable de $\boldsymbol{X}_i$ et celle de ses voisins dans $N_k(\boldsymbol{X}_i)$.

$$LOF_k(\boldsymbol{X}_i) = \frac{\sum_{\boldsymbol{X}_j \in N_k(\boldsymbol{X}_i)} lrd_k(\boldsymbol{X}_j)}{|N_k(\boldsymbol{X}_i)| \times lrd_k(\boldsymbol{X}_i)} \qquad (2)$$

Cet algorithme sert à donner un score d'anomalie en fonction de la densité local d'un vecteur. Plus un vecteur se retrouve isolé (peu de voisins et/ou voisins éloignés), plus ce score LOF sera éloigné des autres vecteurs.

3 Expérimentation et résultats

Pour la 1re étape de détection, nous avons sélectionné le modèle Yolov5m6, que nous avons entrainé sur 143 images d'isolateurs basse tension en utilisant une carte GPU Nvidia RTX3070 Laptop. Ces images de résolution 4096*2160 proviennent de 2 vidéos de missions de maintenance préventive par drone prises, en conditions réelles.
Les hyperparamètres et les poids utilisés sont ceux du pré-entrainement sur ImageNet réalisé par Ultralytics [1]. Nous avons fixé un batch size de 4, une résolution d'image de 1280*720, et l'entraînement dure 85 époques. La figure 1 présente des résultats de ces détections sur une image de test.

FIGURE 1 – *Exemple de détection d'isolateurs par Yolov5m6.*

Pour avoir une meilleure idée de la qualité de la détection de Yolov5m6, nous avons créé une interface graphique permettant de voir facilement les résultats sur de nouvelles vidéos de tests, comme le montre la figure 2. Le graphe affiché montre le nombre d'isolateurs détectés à chaque image de la vidéo. On peut donc voir sur une vidéo où sont les moments d'intérêts (ici le drone a survolé 3 poteaux qui supportaient respectivement 3, 7 et 3 isolateurs, et perdait de vue 2 isolateurs pendant quelques secondes sur le 2ème poteau).
Nous avons alors utilisé des algorithmes de suivi comme Deepsort [13] et Strongsort [14]. Ces algorithmes permettent de suivre et d'assigner un id à chaque objet détecté dans une séquence d'images, en se basant sur une estimation de leur mouvement par filtre de Kalman et sur la ressemblance de texture grâce à une corrélation entre des vecteurs caractéristiques calculés sur les imagettes. Cela nous a permis d'améliorer la détection d'isolateur en baissant le seuil de confiance à partir duquel nous considérons les détections de Yolo comme valides. Nous filtrons ensuite les potentiels faux positifs créés, en prenant en compte uniquement les objets suivis apparaissant sur plus de 15 images d'affilés (voir figure 3).

FIGURE 2 – *GUI affichant les résultats de l'analyse d'un drone survolant 3 poteaux, supportant respectivement 3,7, et 3 isolateurs. Le graphe montre en abscisse la timeline de la vidéo et en ordonnée le nombre d'objets détectés.*

1. https ://github.com/ultralytics/yolov5

FIGURE 3 – *Graphe de détection filtré par tracking. Les "tracklets" (ensemble des bounding box d'un objet identifié) apparaissant moins de 15 frames sont considérées comme des faux positifs et ne sont donc pas pris en compte. On voit que par rapport à la figure 2, le nombre d'isolateurs détectés est plus stable lors du survol des poteaux.*

Nous effectuons ensuite l'extraction de vecteurs caractéristiques par VGG16 et le calcul du LOF.

Nous avons lancé cette chaîne de traitement sur une vidéo de vol fournie par Enedis, où un isolateur présente une fissure. Nous avons représenté sur le graphe de la figure 4 les scores triés par ordre décroissant (et non pas par ordre d'apparition des objets dans la vidéo) afin de repérer plus facilement les objets se démarquant. Nous voyons en rouge les scores d'anomalies correspondant aux images de cet isolateur défectueux. On remarque donc que les scores des images de cet isolateur font partie des plus distincts et se démarquent. A noter que les valeurs négatives de LOF sur les graphes viennent de l'implémentation par sklearn qui donne l'opposé de la valeur de LOF.

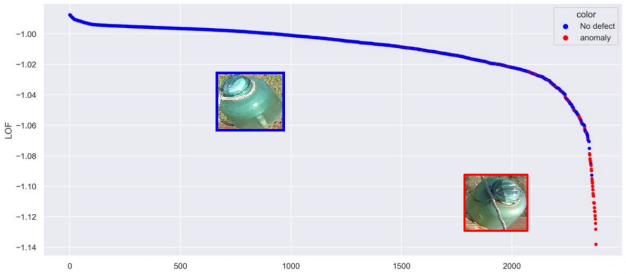


FIGURE 4 – *Score LOF (k=400) associé aux imagettes des isolateurs détectés par YOLO pendant un vol. Les imagettes sont triées par ordre décroissant de leur score d'anomalie. En rouge, les imagettes de l'isolateur défectueux*

Pour tester l'aspect générique de la méthode, les figures 5 et 6 montrent les résultats de tests sur des poteaux en bois dont certains étaient rongés, et sur des poteaux en acier dont le couvercle avait disparu [2]. On remarque que ces objets en défaut se retrouvent également avec des scores LOF se démarquant des objets sans défauts.

2. N'ayant pas de base de données pour entraîner le réseau à les détecter, les imagettes utilisées ont été découpées manuellement depuis les vidéos de vol.

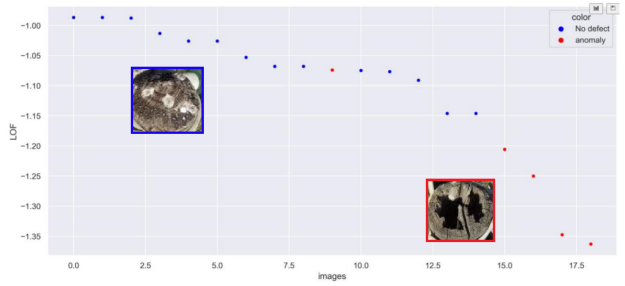


FIGURE 5 – *Score LOF (k=2) pour des poteaux en bois, trié par ordre décroissant. Les poteaux rongés ayant une texture différente, ils obtiennent globalement un score d'anomalie différent.*

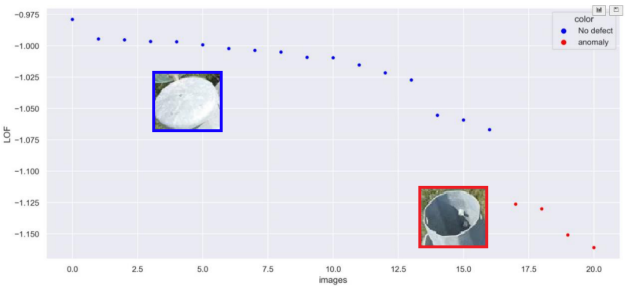


FIGURE 6 – *Score LOF (k=2) pour des poteaux en acier, trié par ordre décroissant. L'absence de couvercle sur certains poteaux affectant beaucoup l'aspect visuel, leurs scores d'anomalies se démarquent beaucoup.*

Nous avons implémenté cette fonctionnalité dans notre interface graphique afin de faciliter l'analyse des résultats (figure 7). L'interface affiche pour chaque id créé par le tracking les frames où l'objet identifié apparait durant la vidéo. La couleur (du vert au rouge) dépend de la valeur du score d'anomalie calculée par le Local Outlier Factor.

FIGURE 7 – *En abscisse on présente la "timeline" de la vidéo. Chaque objet détecté est représenté par une barre colorée en fonction de sa normalité (vert=normal, rouge =anormal). L'utilisateur peut alors cliquer sur la barre pour visualiser l'image représentative de la tracklet considérée afin de l'étudier.*

4 Conclusion et perspectives

Nous avons proposé une méthode d'analyse de défauts visuels générique. Nous collaborons actuellement avec Enedis afin de pouvoir récolter de plus amples exemples d'anomalies dans le but de pouvoir évaluer correctement l'efficacité de notre méthode. Cependant, de par la rareté des anomalies, cette base est difficile à constituer. Cela nous permettrait de quantifier l'impact du choix de la valeur de l'hyperparamètre k du LOF. L'introduction du Self-Supervised Learning (SSL) dans le domaine de la détection d'anomalie est également une piste prometteuse [15].
La première étape de détection d'objets par Yolo n'a été entrainé que sur une centaine d'exemples. Afin de rendre toute la méthode le plus générique possible, nous voulons explorer la piste du Few-Shot Learning. Le Few-Shot dans le cadre de la classification est assez répandu, mais beaucoup moins pour de la détection [16] à cause de la difficulté supplémentaire de devoir localiser l'objet parmis beaucoup de zones potentielles. Cette chaîne de traitement est une aide de diagnostic pour les techniciens qui devront valider ou non les résultats. Nous voulons donc aussi nous pencher sur l'active/incremental learning afin de prendre en compte les retours des experts pour améliorer continuellement notre modèle [17].

Remerciements

Nous remercions l'Association Nationale de la Recherche et de la Technologie ainsi que Drone Geofencing pour le financement de la thèse CIFRE. Nous remercions également Enedis pour ses vidéos de vol.

Références

[1] Van Nhan Nguyen, Robert Jenssen, et Davide Roverso. Automatic autonomous vision-based power line inspection : A review of current status and the potential role of deep learning. *International Journal of Electrical Power & Energy Systems*, 99 :107–120, 2018.

[2] Xinyu Liu, Xiren Miao, et al. Data analysis in visual power line inspection : An in-depth review of deep learning for component detection and fault diagnosis. *Annual Reviews in Control*, 50 :253–277, 2020.

[3] Diana Sadykova, Damira Pernebayeva, Mehdi Bagheri, et Alex James. IN-YOLO : Real-time detection of outdoor high voltage insulators using UAV imaging. *IEEE Transactions on Power Delivery*, 35(3) :1599–1601, 2019.

[4] Xian Tao, Dapeng Zhang, Zihao Wang, et al. Detection of Power Line Insulator Defects Using Aerial Images Analyzed With Convolutional Neural Networks. *IEEE Transactions on Systems, Man, and Cybernetics : Systems*, 2018.

[5] Xuefeng Li, Hansong Su, et Gaohua Liu. Insulator defect recognition based on global detection and local segmentation. *IEEE Access*, 8 :59934–59946, 2020.

[6] Van Nhan Nguyen, Robert Jenssen, et Davide Roverso. Intelligent Monitoring and Inspection of Power Line Components Powered by UAVs and Deep Learning. *IEEE Power and Energy Technology Systems Journal*, 6(1) :11–21, 2019.

[7] Markus Oberweger, Andreas Wendel, et Horst Bischof. Visual recognition and fault detection for power line insulators. Dans *Proc. 19th Comput. Vis. Winter Workshop*, pages 1–8, 2014.

[8] Glenn Jocher, Ayush Chaurasia, et Alex Stoken et al. ultralytics/yolov5 : v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, Février 2022.

[9] K Simonyan et A Zisserman. Very deep convolutional networks for large-scale image recognition. pages 1–14. Computational and Biological Learning Society, 2015.

[10] Kaiming He, Xiangyu Zhang, et al. Deep Residual Learning for Image Recognition. Dans *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Mingxing Tan et Quoc Le. EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks. Dans Kamalika Chaudhuri et Ruslan Salakhutdinov, éditeurs, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 de *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019.

[12] Markus M. Breunig, Hans-Peter Kriegel, et al. LOF : Identifying Density-Based Local Outliers. *SIGMOD Rec.*, 29(2) :93–104, May 2000.

[13] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, et Ben Upcroft. Simple online and realtime tracking. Dans *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, Sep. 2016.

[14] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, et Hongying Meng. StrongSORT : Make DeepSORT Great Again. *IEEE Transactions on Multimedia*, pages 1–14, 2023.

[15] Loïc Jézéquel, Ngoc-Son Vu, Jean Beaudet, et Aymeric Histace. Anomaly Detection via Learnable Pretext Task. Dans *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1178–1185, 2022.

[16] Anh-Khoa Nguyen Vu, Nhat-Duy Nguyen, et al. Few-shot object detection via baby learning. *Image and Vision Computing*, 120 :104398, 2022.

[17] Daochen Zha, Kwei-Herng Lai, Mingyang Wan, et al. Meta-AAD : Active Anomaly Detection with Deep Reinforcement Learning. Dans Claudia Plant, Haixun Wang, Alfredo Cuzzocrea, Carlo Zaniolo, et Xindong Wu, éditeurs, *20th IEEE International Conference on Data Mining, ICDM 2020, Sorrento, Italy, November 17-20, 2020*, pages 771–780. IEEE, 2020.

Vers un outil d'inspection temps réel de l'état d'avancement d'un chantier de construction par RA

M. Baubriaud[1 2] S. Derrode[1] R. Chalon[1] K. Kernn[2]

[1] LIRIS, CNRS, UMR5205, Centrale Lyon, F-69130 Ecully, France

[2] SPIE Building Solutions, F-69320 Feyzin, France

{mathis.baubriaud, stephane.derrode, rene.chalon}@ec-lyon.fr

Résumé

Parmi les technologies clés de l'industrie 4.0, la Réalité Augmentée (RA) et le Building Information Modeling (BIM) sont deux des technologies les plus prometteuses pour assister les experts à inspecter un chantier de construction. Pour cette tâche, la pratique actuelle reste chronophage, laborieuse et manuelle, nécessitant une extraction exhaustive des données à partir de dessins et d'autres bases de données. Les systèmes basés sur des appareils mobiles pourraient accélérer et faciliter le processus d'inspection. À cette fin, un nouvel outil de RA pour assister les opérateurs lors des activités d'inspection de chantier en intérieur est proposé. La méthode introduite détecte automatiquement si les équipements du génie climatique et électrique (gaines, ventilation, chauffage, etc.) existent en réalité, conformément aux maquettes numériques 3D. Des expérimentations sur le terrain ont été menées afin d'estimer l'ensemble optimal de paramètres de la méthode et d'évaluer sa convivialité.

Mots clefs

Réalité Augmentée, suivi de progression, BIM, inspection.

1 Introduction

Le suivi d'avancement des chantiers de construction est essentiel, car il donne aux gestionnaires les informations nécessaires pour agir rapidement et judicieusement [1]. Un suivi d'avancement inefficace entraîne une perte de contrôle du chantier, occasionnant des dépassements en temps et en coûts. La saisie manuelle des données requise par les techniques conventionnelles de supervision des progrès est laborieuse, chronophage et sujette à l'erreur humaine [2]. Les inspections des travaux intérieurs - par exemple, l'installation de chauffage, de ventilation et de climatisation (CVC) - peuvent s'avérer encore plus difficiles pour les inspecteurs en raison du niveau de détail et de l'interdépendance des tâches [3]. Par conséquent, il devient souhaitable d'automatiser ces tâches.

Le BIM est un ensemble de politiques, de processus et de technologies en interaction ayant pour but de gérer les données essentielles d'un chantier, sous forme numérique, tout au long du cycle de vie de celui-ci [4]. Une étape cruciale pour se rapprocher de cet idéal réside dans la numérisation des maquettes en 3D.

La RA est l'une des techniques les plus prometteuses pour améliorer le transfert d'informations du monde numérique vers le monde physique de manière non intrusive [5]. La RA peut assister les opérateurs sur leur lieu de travail, en temps réel et pendant les opérations manuelles, en réduisant les erreurs humaines et la dépendance à la mémoire de l'opérateur, car elle a la capacité d'intégrer des informations virtuelles dans la perception du monde réel par l'utilisateur.

Des efforts ont été déployés vers des systèmes pour automatiser le suivi des progrès à l'aide de scanners laser, de photogrammétrie, de vidéogrammétrie et d'identification par radiofréquence (RFID) en lien avec des plateformes BIM [6]. De plus, certaines études ont tenté d'appliquer la RA au domaine de l'ingénierie architecturale [7, 8]. Cependant, la plupart de ces systèmes sont statiques, limités aux chantiers de construction extérieurs ou à des équipements spécifiques et ne fonctionnent pas en temps réel.

Cet article présente une méthode basée sur un appareil mobile pour la surveillance automatisée des progrès en temps réel d'un chantier. Les sections suivantes comprennent la mise en œuvre, les tests et les résultats de la méthode imaginée. Pour terminer, une discussion sur les limites et les futurs efforts de recherche est proposée.

2 Mise en œuvre de l'outil

La majorité des appareils qui prennent en charge les applications de RA sont des appareils RGBD, qui fournissent une reconstruction 3D de l'environnement. Les lunettes Microsoft Hololens 2 [1] embarquent deux capteurs de profondeurs, un accéléromètre, et un gyroscope, exploités pour fournir une cartographie spatiale de l'environnement. Il s'agit d'un appareil portable qui ne limite pas l'utilisation des mains, ce qui réduit les risques pour la sécurité et offre une expérience interactive et intuitive pour les utilisateurs.

1. https://www.microsoft.com/fr-fr/hololens/hardware

Cet article présente une méthode de comparaison automatisée et en temps réel entre le volume reconstruit capturé à partir de Hololens et la maquette 3D tel que planifiée. L'analyse est pensée pour fonctionner pendant que l'inspecteur navigue dans le bâtiment en construction. Une vue d'ensemble de la solution est présentée dans la Fig. 1

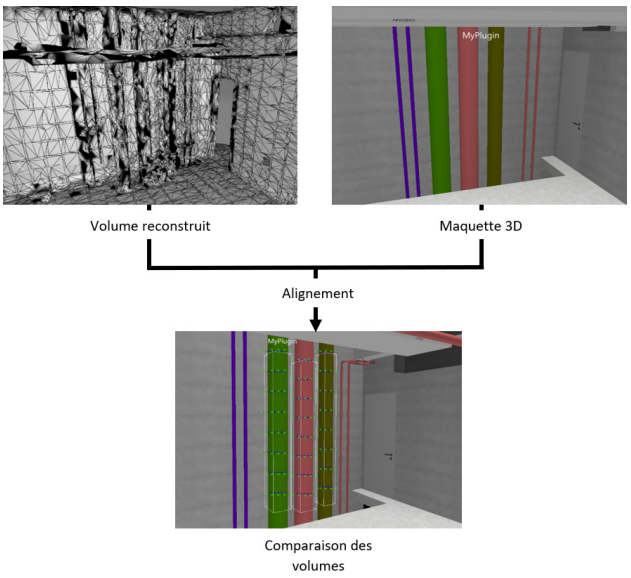


FIGURE 1 – *Vue d'ensemble de la solution.*

Le processus se compose de deux étapes principales, décrites dans les sous-sections suivantes. La première est un contrôle de visibilité des objets depuis le point de vue de la caméra dans la maquette 3D et la seconde permet la comparaison entre le volume reconstruit et les objets 3D considérés visibles. Le résultat final est la classification des objets 3D comme « construits » ou « non construits ». Ce travail s'inscrit dans une dynamique d'amélioration de l'application NEXT-BIM [2]. Ce logiciel permet la visualisation des maquettes 3D sur les lunettes ainsi que l'alignement semi-automatique du volume reconstruit.

2.1 Contrôle de visibilité des objets 3D

La première étape consiste à effectuer un contrôle de visibilité. Cette étape détecte quels objets de la maquette 3D sont visibles depuis la caméra de l'appareil mobile. La comparaison est effectuée pour les éléments visibles au fur et à mesure du déplacement de la caméra. L'emplacement de la caméra, son orientation et ses caractéristiques intrinsèques dans le monde et dans la maquette 3D sont récupérés automatiquement depuis les différents capteurs des lunettes. Un objet doit satisfaire quatre conditions pour être défini comme visible :

— L'objet est du type cible que nous cherchons à analyser ;
— L'objet est dans le frustum de la caméra, qui correspond à la zone qui apparaît à l'écran ;
— La distance entre l'objet et la caméra est inférieure à 5m, la qualité du maillage renvoyé par les lunettes devenant inexploitable au-delà ;
— L'objet n'est pas occulté.

Pour chaque objet rendu dans la maquette 3D appartenant au type recherché, un simple calcul de la distance de celui-ci à la caméra permet de faire une présélection. Ensuite, plusieurs rayons, d'origines uniformément réparties dans l'objet et pointant vers la caméra, sont générés. Le nombre de ces rayons est proportionnel aux dimensions de l'objet. Si au moins 2 de ces rayons atteignent la caméra, l'objet est estimé visible. En outre, la position des points non masqués permet d'extraire uniquement la portion visible de l'objet 3D en question.

2.2 Analyse de présence d'un objet 3D

La maquette 3D ainsi que le volume reconstruit se compose de sommets reliés entre eux, formant un maillage triangulaire dense. Une description de l'algorithme élaboré pour l'analyse de présence est présentée dans l'Alg. 1. Dans un premier temps, une phase de prétraitement de la donnée est réalisée contenant 3 sous-étapes, illustrées dans la Fig. 2. Une projection des points 3D de l'objet sur l'écran 2D permet d'obtenir $projPts$ 2D. L'enveloppe convexe 2D de $projPts$ est calculée à l'aide d'une implémentation de la méthode du parcours de Graham [9]. Cette enveloppe représente maintenant la portion qu'occupe l'objet sur l'écran de l'utilisateur et l'Axis Aligned Bounding Box $aabb$ est calculée pour simplifier la forme du polygone.

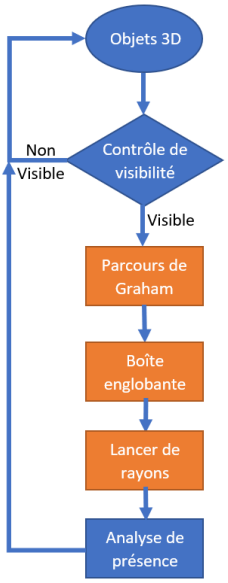


FIGURE 2 – *Prétraitement à l'analyse de présence.*

À l'aide des paramètres de cette surface 2D, de la taille maximale de l'objet 3D obj_size et de la taille de l'écran $screen\_size$, un nombre de rayons $nRay[x, y]$, à générer le long de l'axe des abscisses x et des ordonnées y est défini puis ajusté [3] pour chaque objet à l'aide de la formule (1) :

$$nRay[x]* = obj\_size * \frac{max(aabb[x]) - min(aabb[x])}{screen\_size[x]} \quad (1)$$

Ces rayons, d'origines uniformément réparties dans l'enveloppe, sont projetés vers l'objet 3D. Cette méthode permet de récupérer un échantillon significatif de la forme 3D de l'objet tel qu'il est vu par l'utilisateur. Si un

2. https://next-bim.com/

3. de manière équivalente pour l'axe y

rayon atteint bel et bien l'objet, la position 3D de l'impact ainsi que la normale à la surface de l'objet est enregistrée. Si le nombre de rayons atteignant l'objet dépasse un seuil $reachedRayTreshold$ proportionnellement au nombre envoyé depuis l'enveloppe, alors l'analyse se poursuit, sinon l'objet est considéré de forme trop complexe pour pouvoir être étudié (au vu de la précision des capteurs Hololens). Pour chacun des points 3D enregistrés, un nouveau rayon ayant pour vecteur directeur la normale à l'objet et partant dans les deux directions avec une longueur $segmentSize$ de 10 cm est créé. L'objectif est ainsi de détecter si une éventuelle intersection avec le volume reconstruit est identifiée et si la normale au point d'intersection est cohérente avec la forme de l'objet. Selon un seuil $intersectMeshTreshold$ du rapport nombre d'intersections par le nombre de rayons générés, on considérera l'objet virtuel comme existant ou non existant dans la scène réelle.

Algorithm 1 Analyse de présence.

Input : $aabb, nRay, rayPositions$
Output : $presence$
Data : $nbrHits = 0, reachedRayTreshold = 0.25, segmentSize = 10$
if $size(rayPositions) > reachedRayTreshold$
 then
  for p **in** $rayPositions[]$ **do**
   $hit = GetRayIntersection(p.pos+ segmentSize * p.norm, p.norm)$
   if $hit.HitDistance < segmentSize * 2$ **then**
    //Il y a une intersection avec le maillage dans l'intervalle de 20cm
    $nbrHits++$
   end if
  end for
  $intersectMeshTreshold = 0.5$
  if $nbrHits > intersectMeshTreshold * size(rayPositions)$ **then**
   return $presence = 1$
  else
   return $presence = 0$
  end if
end if

Les seuils ont été déterminés par des expériences qui sont présentées dans la section suivante.

3 Expérimentation et résultats

L'environnement de développement de NEXT-BIM, le framework Microsoft Hololens et Microsoft Visual Studio 2022 ont permis la validation et les tests de la méthode proposée. L'algorithme est rédigé dans le langage de programmation C#. Des expériences ont été menées avec des données de différents chantiers de construction CVC, plomberie et électrique.

La solution implémentée sur les lunettes délivre une performance de 10 à 30 images par seconde, inversement proportionnelle à la quantité d'objets d'intérêts à analyser dans une scène. Une étude qualitative de l'outil a été conduite, avec, comme critères, sa convivialité, sa pertinence et son acceptabilité. Un échantillon significatif de chantiers tertiaires de construction en intérieur fut sélectionné. Durant chaque visite, un expert sur place, externe au projet, testait la solution, et son ressenti était recueilli à l'aide d'un questionnaire. Différents points positifs et cas d'usages, mais également des voies d'améliorations en découlent.

L'outil apparaît aux premiers abords convaincants, les résultats sont immédiats et une aide visuelle à l'aide de couleur, verte ou rouge en fonction du bilan de l'analyse automatique permet de répertorier immédiatement la présence ou non des objets. Néanmoins, on observe rapidement des limitations matérielles importantes qui seront mentionnées dans la conclusion de cet article.

Dans l'optique de réduire le facteur humain et matériel, l'utilisation d'un outil développé par NEXT-BIM permet d'enregistrer les relevés des capteurs des lunettes. Ainsi, il est possible de rejouer une scène en temps réel, reproduisant les mouvements exacts de l'utilisateur dans le bâtiment et limitant les résultats d'analyse aux seules constantes de l'algorithme proposé. Un exemple de l'interface de cet outil est présenté en Fig. 3.

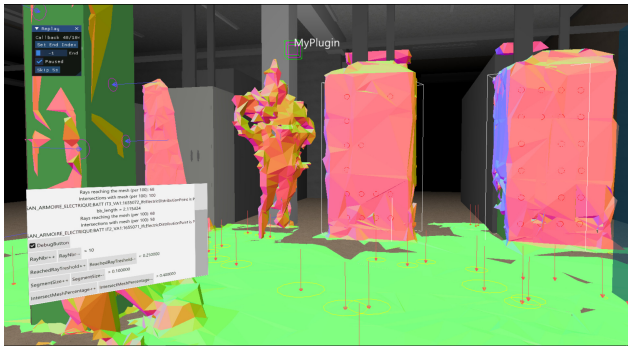


FIGURE 3 – *Exemple d'interface d'expérimentation.*

Les métriques pour évaluer les résultats sont l'exactitude, la précision et le rappel. L'exactitude indique à quel point l'algorithme est correct dans l'ensemble. La précision spécifie le taux d'éléments retrouvé rapporté au nombre d'éléments total et le rappel indique le taux d'éléments retrouvé rapporté au nombre d'éléments construit. La vérité terrain est définie en étiquetant manuellement tous les objets comme existant ou non. Une étude de l'impact des constantes de l'algorithme révèle que les deux constantes dont la sensibilité est la plus haute sont la taille des segments $segmentSize$ qui vérifie l'intersection avec le maillage Hololens et le seuil $intersectMeshTreshold$ nécessaire pour que l'objet soit considéré présent. Le Tabl. 1 présente les résultats obtenus dans le cas d'une pièce d'un datacenter en construction contenant 10 armoires élec-

triques dans la maquette 3D dont 5 correctement placés à la date de l'inspection. Malgré le nombre de données réduites

Seuil			Taille des segments 7	10	13
	25%	Exactitude	60	90	80
		Précision	**100**	83	71
		Rappel	60	**100**	**100**
	50%	Exactitude	70	**100**	90
		Précision	**100**	**100**	**100**
		Rappel	70	**100**	83
	75%	Exactitude	50	80	90
		Précision	**100**	**100**	**100**
		Rappel	40	60	83

TABLEAU 1 – *Pourcentage d'exactitude, précision et rappel pour différents seuils et tailles de segments.*

lors de ce test, on observe bien des tendances. Lorsque les segments sont trop courts, de moins nombreuses intersections sont relevées entre la maquette 3D et le volume reconstruit ; inversement des faux positifs sont observés lorsqu'ils sont trop longs. Les expériences et l'évaluation des résultats suggèrent que les paramètres optimaux de seuil et de taille de segment pour la solution proposée sont respectivement de 50% et 10 cm. Toutefois, ces valeurs ne peuvent pas être généralisées pour tous les types d'objets 3D. Des résultats différents pourraient être obtenus dans d'autres cas d'usage.

4 Conclusion et perspectives

Les pratiques actuelles d'inspection de chantier sont principalement manuelles et demandent une main-d'œuvre importante. Cet article présente une solution mobile de suivi des progrès automatisés en temps réel. Pendant que l'inspecteur navigue à l'intérieur d'un bâtiment en construction, l'outil développé analyse les objets de la maquette 3D et un algorithme basé sur les lancers de rayons permet de les comparer avec le volume reconstruit par les lunettes. Des expérimentations ont permis d'optimiser les paramètres de la solution. Néanmoins, de plus amples tests doivent être conduits pour affiner ses constantes, en particulier dans des scénarios différents et avec un plus grand nombre de données. La méthode actuelle est prometteuse, mais présente comme défaut principal de reposer uniquement sur des données de profondeurs, peu fiables et précises. De plus, il existe parfois de grosses disparités entre le BIM et le chantier à tel point que des décalages de plusieurs dizaines de centimètres peuvent être observés. Dans ces cas, un algorithme est en cours de conception pour calculer ces décalages. Également, l'usage d'un casque de RA dans un environnement complexe tel qu'un chantier soulève des questions de sécurité et de prise en main de la technologie. D'autres travaux sont en cours pour pallier ces manques en utilisant par exemple la caméra des lunettes. Bien que la solution proposée ait été développée pour Microsoft Hololens, elle pourrait s'appliquer à tout appareil dont les données d'entrée et de sortie sont similaires (c'est-à-dire un maillage construit grâce au capteur de profondeur de temps de vol (ToF) des lunettes).

En conclusion, de nombreux cas d'usage ont été imaginés et le potentiel d'un tel outil est justifié. Par conséquent, des développements supplémentaires doivent être menés.

Remerciements

La recherche menant à ces résultats a été financée par une thèse co-encadrée entre l'École Centrale Lyon et la société SPIE Building Solutions. L'entreprise NEXT-BIM est remerciée pour son soutien dans la mise en place de la solution dans leur environnement de travail et l'utilisation de leurs outils.

Références

[1] V. K. Reja, K. Varghese, et Q. P. Ha. Computer vision-based construction progress monitoring. *Automation in Construction*, 138 :104245, juin 2022.

[2] Jochen Teizer. Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites. *Advanced Engineering Informatics*, 29(2) :225–238, avril 2015.

[3] B. Koo et M. Fischer. Feasibility study of 4D CAD in commercial construction. *Journal of Construction Engineering and Management*, 126(4) :251–260, juillet 2000. Publisher : American Society of Civil Engineers.

[4] Rafael Sacks, Charles Eastman, Ghang Lee, et Paul Teicholz. *BIM Handbook : A Guide to Building Information Modeling for Owners, Designers, Engineers, Contractors, and Facility Managers*. Wiley, août 2018.

[5] F. De Pace, F. Manuri, et A. Sanna. Augmented Reality in Industry 4.0. *American Journal of Computer Science and Information Technology*, 06(01), 2018.

[6] T. Omar et M. L. Nehdi. Data acquisition technologies for construction progress tracking. *Automation in Construction*, 70 :143–155, octobre 2016.

[7] C. Park, D. Lee, O. Kwon, et X. Wang. A framework for proactive construction defect management using BIM, augmented reality and ontology-based data collection template. *Automation in Construction*, 33 :61–71, août 2013.

[8] M. Kopsida et I. Brilakis. Real-time volume-to-plane comparison for mixed reality–based progress monitoring. *Journal of Computing in Civil Engineering*, 34(4) :04020016, juillet 2020.

[9] X. Kong, H. Everett, et G. Toussaint. The Graham scan triangulates simple polygons. *Pattern Recognition Letters*, 11(11) :713–716, novembre 1990.

Session posters

Actual Fabric Digitalization

Thu Ha DO, Minh Chau HUYNH, Xuyuan TAO, Pascal BRUNIAUX,
Ludovic KOEHL, Kim Phuc TRAN, Xianyi ZENG
Univ. Lille, ENSAIT, ULR 2461 - GEMTEX - Génie et Matériaux Textiles, F-59000 Lille, France

Abstract

3D computer-aided design systems have emerged as promising techniques for garment learning processing, virtual shopping, and fashion shows within the fashion and garment industries. However, the effective application of these digital systems requires precise characterization of fabrics, garment patterns, and human body shapes that accurately reflect the appearance and behavior of the garments in a digital environment. Creating a 3D digital garment involves inputting the corresponding digital fabric properties. Nevertheless, obtaining these measurements can be complex, often necessitating the involvement of well-trained technicians. In this study, our focus is on a simplified and automated technique for digitizing real fabrics. Specifically, we aim to find the most relevant digital fabric in the database of a 3D software by employing image processing and machine learning techniques to drape images.

Mots clefs

Machine learning, Drape parameter, Garment design

1 Introduction

Industry 4.0 represents a significant challenge and opportunity for technology companies, as it has the potential to contribute to the advancement of humanity. Clothing is a fundamental human need, and thus, competition among companies in the garment industry must adapt to this revolution with flexibility and responsiveness in order to satisfy a broad range of potential market demands. In textile industry, the choice of fabrics plays a crucial role in determining the wearer's comfort and garment style by their various mechanical properties. Specifically, the drape of fabric is one of the most important mechanical properties that can significantly impact on interactions of a garment with the body. It is related to comfort feeling of the wearer and aesthetics of the design style (e.g. well fitted or not). Various studies have been conducted to develop drapemeters with the aim of simplifying the measurement process, (presented in [1, 2, 3]), improving the accuracy of fabric properties, reducing the reliance on operator expertise, and proposing alternative fabric drape parameters.

Creating a 3D digital garment requires inputs of the corresponding digital fabric properties. These properties can be directly measured using physical instruments, such as Kawabata Evaluation System (KES) [4] and Fabric Assurance by Simple Testing (FAST) [5]. However, these measurements are rather complex and require interventions of well-trained technicians. In this situation, to facilitate the creation of a 3D garment, it is imperative to select a suitable digital fabric already existing in an extensive fabric database linked to the 3D software (for example Lectra, Toray-Acs, Gerber, Investronica, Optitex, etc...), in which the technical parameters (drape parameters, optical parameters, and mechanical parameters) of the representative fabrics are complete [6].

In this study, we focus on a simplified and automated technique for digitizing real fabric through the use of image processing and machine learning techniques based on a software fabric database. First, for a real fabric, the user (designer) extracts its drape image with the use of a simple drape meter (Cusick Drapemeter [7]) and extracts drape parameters by using image analysis. Then, clustering fabric in the software database to choose the most relevant group. The closest digital fabric present in the fabric software database is predicted using machine learning concerning drape image features as the input. The rest of the paper is organized as follows. Section 2 presents the method to predict the closest fabric including three main steps based on image processing and machine learning techniques. Then the results are discussed in Section 3. Finally, the conclusion is briefly discussed in Section 4.

2 Methodology

In this section, the process of digitalizing actual fabric is presented. The concept is a new fabric that has no information about mechanical properties can be assigned with the most relevant fabric found in a software database, (e.g. Lectra database). The overall progress includes three main steps. In the first step, a drape image of a real fabric is processed by extracting contour information using image processing techniques to estimate five drape parameters (explained in Section 2.2). Then, digital fabrics in the software database are clustered into different groups based on their drape parameters and the number of nodes in the real fabric to determine the closest group in Section 2.3. Finally, a number of classification machine learning models are applied to identify the closest digital fabric based on the output of previous steps in Section 2.4. In addition, the software database of fabric is also introduced in Section 2.1.

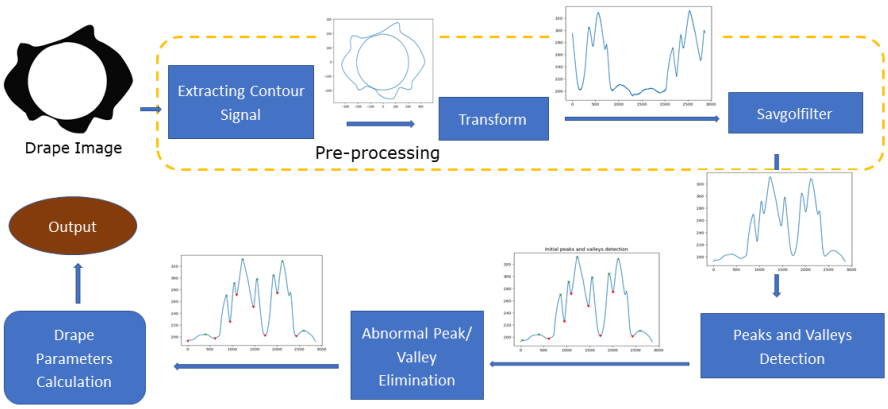


FIGURE 1 – *The process of Drape Parameter Estimation*

2.1 Fabric database preparation

In the 3D CAD software, the technical parameters of a fabric are considered as inputs to the garment simulation system. These technical parameters include a number of basic parameters (e.g. thickness, weight), optical parameters (e.g. texture (weft and warp structure) and color), mechanical parameters (e.g. bending, shearing, tensile), etc. For instance, in the Lectra Modaris 3D Fit CAD software, the database is composed of 111 digital fabrics with drape images (extracted by the use of a simple drape meter (Cusick Drapemeter [7])) and 23 technical parameters (i.e. drape shape, number of fabric, AA, AD, MP, MV, NoP, weight, commercial name, composition, thickness, weave, warp/weft texture, warp bending, weft bending, drape coefficient, number of plies, CisT, CisC, FlexT, FlexC, Colors, Patterns), which is capable of covering almost all ranges of fabrics used in garment design.

2.2 Drape Parameter Estimation based on Image processing

This section outlines the methodology for calculating drape parameters, which comprise the average amplitude (AA), average distance (AD), maximum peak (MP), minimum valley (MV), and number of peaks (NoP). The parameters can be determined by analyzing the contour signal which is extracted from drape image. First, the definition of the parameters is introduced, followed by technical image processing pipeline.

Drape parameter definition. The initial presentation of the extracted contour is in the Polar coordinate system, then the contour signal is transformed to the Cartesian coordinate system to facilitate the analysis, as illustrated in Figure 2. The number of peaks (NoP) in the signal can be determined by counting the number of signal periods, which is equivalent to the number of valleys. Specifically, in Figure 2, the orange, red, blue, and black arrow corresponds to the AD, MP, MV, and AA, respectively. **MP** and **MV** is the distance from zero to the highest and lowest point, respectively. **AD** is the average distance from zero or the mean

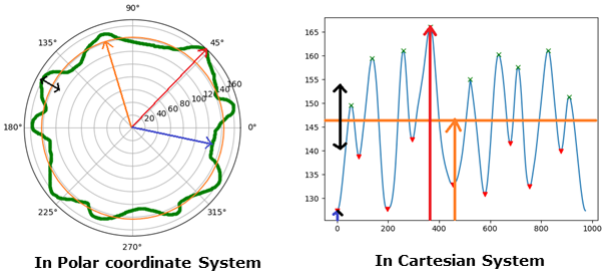


FIGURE 2 – *Contour signal in different space*

of the contour signal. **AA** is the average amplitude of each node in the signal. The unit of these parameters is pixel. The AA and AD from contour signal to the center point of circle are computed using the following formula 1 :

$$AA = \frac{1}{n}\sum_{i=1}^{n}\frac{p_i - v_i}{2}$$
$$AD = \frac{1}{n}\sum_{i=1}^{n}\frac{p_i + v_i}{2} \quad (1)$$

where n : number of peaks (number of valleys); p_i, v_i : dimension of peak and dimension of valley i, respectively. In addition, the maximum peak and minimum valley are also considered in dimension of signal.

Drape Parameters Estimation Process. The process of estimating drape parameters includes four steps, namely *Pre-processing*, *Peaks and valleys Detection*, *Abnormal Elimination*, and *Drape parameters Calculation*, as depicted in Figure 1.

The ***pre-processing*** step plays a critical role in extracting the raw contour information that forms the basis of drape parameter computation. The raw contour signal is extracted in Polar coordinates, achieved through the detection of changes in color or intensity. The next step involves transforming the raw signal into the Cartesian system, taking

into consideration the function of the distance from the center point to the contour line.

Peaks and Valleys Detection : After extracting the clean contour signal, an algorithm for peaks detection (presented in [8]) is applied to detect the number and position of peaks and valleys.

Abnormal peak/valley Elimination : A node consists of a peak and a valley. Therefore, the process of finding the number of nodes can be considered as the process of detecting peaks and valleys. In some cases, the detected peak or valley may not be considered as a node in the context of the whole signal. For example, some peaks or valleys may be considered as leakage or abnormal rather than true peaks or valleys. To address this, our algorithm detects pairs of peaks and valleys whose distance is less than a certain defined threshold and considers them as abnormal peaks or valleys. Additionally, we assume that the number of peaks and valleys is equal and that a peak is always followed by a valley. Consequently, any peak or valley that does not satisfy this condition will be removed.

Drape Parameters Calculation : Upon the removal of the outlier noise, the drape parameters are calculated in accordance with the procedures outlined in Section 2.2.

2.3 Clustering Process

The idea is to cluster the existing digital fabrics (in software database) into different clusters based on the similarity of their drape parameters to identify the group that is the closest to the real fabric (closet group), as shown in Figure 3. To simplify the process, we reduce the number of digital fabrics based on the estimated NoP parameter (calculated in Section 2.2). Suppose the estimated NoP parameter is N. The idea is to choose all digital fabrics that have NoP = [N-1, N, N+1]. Then, these fabrics are clustered using the K-means algorithm and Principal Component Analysis (PCA).

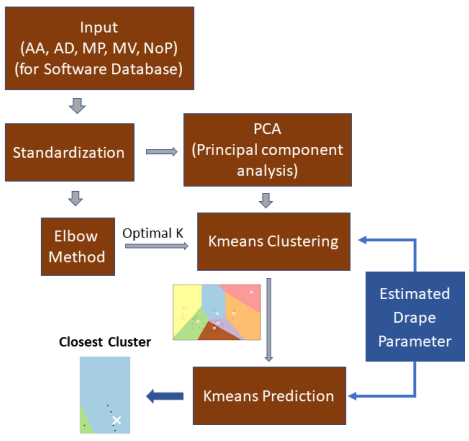


FIGURE 3 – *The process of Clustering software database*

K-means (presented in [9]) is an unsupervised machine learning algorithm that divides data into a specified number of clusters, where it partitions a set of fabrics based on the four attributes, namely AA, AD, MP, and MV. The selection of an appropriate value for the number of clusters (k) is crucial in K-means clustering, and the "elbow method" is commonly employed to determine the optimal value of k. Due to the limitation of the number of fabrics in the database, it is necessary to determine the ideal number of clusters. By using elbow method, we found that k=6 is the optimal choice (the elbow of the curve). In addition, we applied Principle Component Analysis (PCA) to reduce the data of high dimensions to a plan before applying K-means. Then, the selected digital fabrics are clustered into different k groups (6 clusters), as demonstrated in Figure 3. After the clustering, the closest group of digital fabrics can be obtained by comparing the distance between the centroid of each cluster with the real fabric base on its estimated drape parameters.

2.4 Prediction Process

The main concept behind the prediction process is to analyze the characteristics of both real fabric (based on its drape parameter - calculated in Section 2.2) and the digital fabrics in the closest cluster derived in Section 2.3 (they are taken as leaning data for prediction), in order to identify the most similar fabric existing in the digital software database. The prediction task is carried out using the Min Euclidean Distance technique and five machine learning techniques based on the previous learning data, including K-Nearest Neighbors (KNN) [10], Random Forests [11], Naive Bayes [12], and Decision Tree [13] based on drape parameters, as shown in Figure 5. The results achieved include the name of the predicted fabric and its mechanical properties.

3 Results

The process of testing a real fabric (collected from our partners) is illustrated in Figure 4. Five machine learning models are applied to predict the results. It should be noted that the achieved results (predicted numbers of digital fabrics) may differ in some cases. In such cases, we can provide all the results and the users can decide by themselves which one is the most relevant according to their preference or experience. If the user's experience is not available, the majority rule can also be used to select the most relevant digital fabric. For example, if five learning models deliver the fabric n°90 and another n°95, we will naturally take n°90 as the most relevant fabric.

The achieved results may be affected by differences in the distance between the camera and fabric, as well as variations in the size of the image captured. These factors can influence the accuracy and consistency of the obtained results and should be taken into consideration when analyzing the data. In Figure 2.4, the drape parameters of the digital and real fabrics are different but their contour shapes are almost the same. We consider that they are very similar fabric samples. Our objective is to find the most similar fabric in the software database, and we can enhance our performance by extending the size of the database sample.

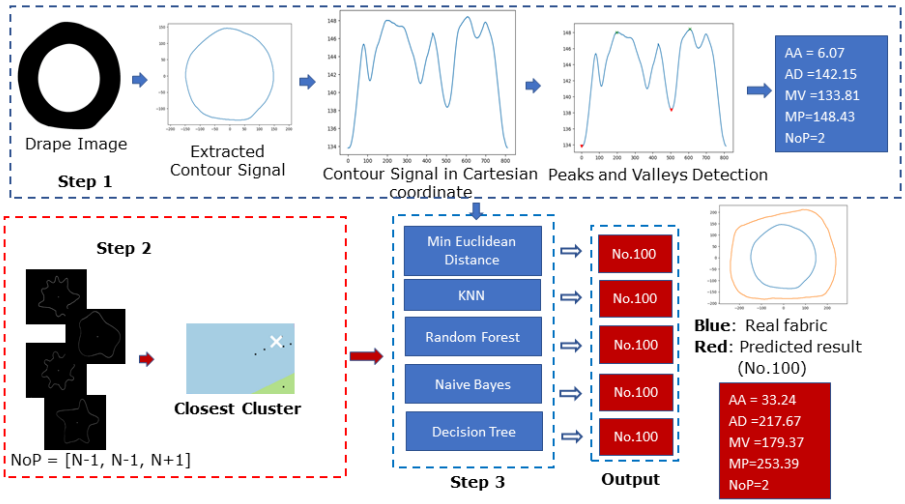


FIGURE 4 – *Testing in a real fabric*

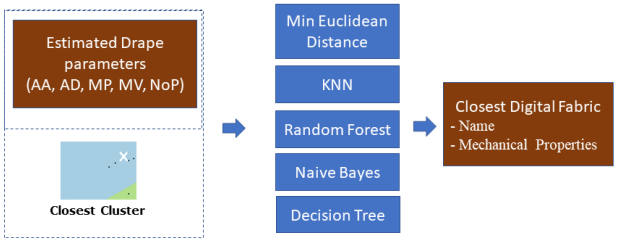


FIGURE 5 – *The process of Predicting the closest fabric based on Classification models*

4 Conclusion

The process of digitizing fabric can be accomplished manually, although it is important to acknowledge that this approach is inherently subjective and reliant upon the opponent of user/designer. With the objective of establishing a more precise and consistent approach, we initially developed an objective method to digitalize the digital fabric that most closely resembles the real fabric from the software database. Our objective is to identify the most accurate digital match for a given fabric within the software database. Our effort is try to find the best possible digital match for a given fabric within the software database. This process can be refined further to attain even greater accuracy. Further expansion of the digital fabric collection in the software database could potentially enhance the accuracy of our method.

Références

[1] A Giwa, EO Achukwu, et MW Shebe. Measurement of fabric drape using digital image processing. *AFRICAN JOURNAL OF NATURAL SCIENCES (AJNS) ISSN 1119-1104*, 13, 2015.

[2] Behera BK, Ajit Kumar Pattanayak, et Rajesh Mishra. Prediction of fabric drape behaviour using finite element method. *Journal of Textile Engineering*, 54(4) :103–110, 2008.

[3] Reham Sanad, Tom Cassidy, et TLV Cheung. Fabric and garment drape measurement-part 1. *Journal of Fiber Bioengineering and Informatics*, 5(4) :341–358, 2012.

[4] S Kawabata. Standardization and analysis of hand evaluation, and standardization committee of the textile machinery society of japan. *Osaka, Japan*, 1975.

[5] Pier Giorgio Minazio. Fast–fabric assurance by simple testing. *International Journal of Clothing Science and Technology*, 7(2/3) :43–48, 1995.

[6] Yong-Jin Liu, Dong-Liang Zhang, et Matthew Ming-Fai Yuen. A survey on cad methods in 3d garment design. *Computers in industry*, 61(6) :576–593, 2010.

[7] GE Cusick. 21—the measurement of fabric drape. *Journal of the Textile Institute*, 59(6) :253–260, 1968.

[8] Girish Palshikar et al. Simple algorithms for peak detection in time-series. Dans *Proc. 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*, volume 122, 2009.

[9] John A Hartigan et Manchek A Wong. Algorithm as 136 : A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1) :100–108, 1979.

[10] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2) :1883, 2009.

[11] Leo Breiman. Random forests. *Machine learning*, 45 :5–32, 2001.

[12] Kevin P Murphy et al. Naive bayes classifiers. *University of British Columbia*, 18(60) :1–8, 2006.

[13] J Ross Quinlan. Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2) :339–346, 1990.

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.

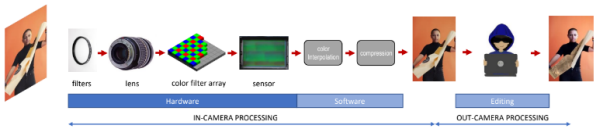


FIGURE 1 – *Système d'acquisition image numérique [8]*

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*.

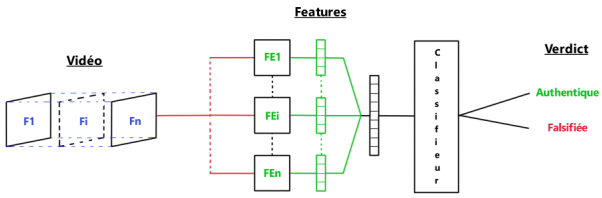


FIGURE 2 – *Architecture proposée*

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.

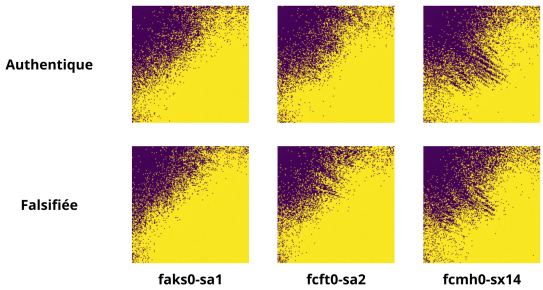


FIGURE 3 – *Visualisation spectre fréquentiel vidéos*

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.

Evaluation de la qualité sans référence des nuages de points basée sur les statistiques de co-occurrence 3D

S. Riache et M.-C. Larabi
CNRS, XLIM UMR 7252, Université de Poitiers
Poitiers, France

Résumé

L'évaluation de la qualité des nuages de points reste un défi majeur en raison de la complexité des applications associées et de la nature du contenu. Pour résoudre ce problème, cet article propose une nouvelle métrique d'évaluation de la qualité des nuages de points basée sur les statistiques de co-occurrence 3D. L'approche proposée implique une stratégie de voxelisation, où le concept de matrice de co-occurrence est étendu en 3D pour calculer l'occurrence dans les 26 directions possibles. Les attributs de Haralick sont ensuite calculées en fonction de l'espace de couleur sélectionné. Une étape de régression sert à mapper ces attributs à la vérité terrain (scores subjectifs) associés aux modèles de nuages de points. Les résultats expérimentaux montrent l'efficacité de l'utilisation des statistiques de co-occurrence 3D pour l'évaluation de la qualité des nuages de points (CO-PCQA). La métrique proposée présente de bonnes performance en comparaison avec la littérature.

Mots clefs

Nuages de points, Métrique de qualité sans référence, Matrice de co-occurences.

1 Introduction

Un nuage de points 3D (NP) est une collection de points dans l'espace 3D [1], chacun avec ses coordonnées géométriques $(X, Y, Z) \in \mathbf{Z}$ et des attributs associés tels que la couleur, la réflectance et la luminance. Les nuages de points sont de plus en plus populaires ces dernières années comme moyen de représenter, stocker et compresser du contenu 3D pour diverses applications telles que la navigation, l'industrie et la réalité virtuelle. Cependant, en raison du grand nombre de points et du potentiel de distorsions lors de l'acquisition, du stockage ou de la compression, il est important de mesurer l'impact sur la qualité visuelle. L'évaluation de cette dernière par des métriques, idéalement sans référence, semble être une alternative pratique. Pour surmonter ce défi, il est nécessaire d'extraire directement des attributs qui soient corrélées à la perception humaine à partir de nuages de points dégradés. La tâche est loin d'être aisée à cause de la nature même du nuage de points, constitué de points désordonnés dans l'espace 3D.

En tant que domaine émergent, différentes métriques avec référence ont été proposées pour les nuages de points. Deux métriques largement utilisées sont P2point et P2plane [2], qui comparent un nuage de points dégradé à un nuage de points de référence en mesurant leur distance géométrique. Cependant, ces métriques sont purement géométriques et ne tiennent pas compte des caractéristiques de couleur ou de texture, qui jouent souvent un rôle significatif dans la perception de la qualité. Pour pallier cette limitation, Gabriel et al. [3] ont proposé la métrique PC-MSDM, inspirée de SSIM [4], qui extrait et compare les caractéristiques visuelles du nuage de points dégradé et de sa référence. Cette métrique ne considère que des caractéristiques structurelles mais sa version étendue, PCQM [5], inclut des caractéristiques de couleur. Une autre approche consiste à utiliser les motifs binaires locaux (LBP de l'anglais Local Binary Patterns) [6], pour mesurer les différences de motifs texturaux et les traduire en terme de qualité visuelle [7]. D'autres chercheurs ont proposé de projeter les nuages de points 3D sur différents plans, comme c'est souvent le cas lors de tests subjectifs, pour obtenir des images 2D afin de les traiter avec des métriques 2D [8].

Plusieurs métriques sans référence ont été proposées pour évaluer la qualité des nuages de points. Une des approches consiste à extraire des attributs basés sur la géométrie et la couleur pour entraîner des modèles d'apprentissage automatique [9]. D'autres approches exploitent les réseaux de neurones convolutionnels et les blocs résiduels en tant qu'extracteurs d'attributs et appliquent la technique de projection multi-vue pour obtenir des projections 2D sous différents angles avant d'entraîner le modèle CNN [10]. Ces méthodes sont encore limitées par la dépendance aux méthodes de rendu, la perte d'informations due la projection et les différences entre les contenus 3D et 2D.

Dans cet article, nous proposons une nouvelle approche basée sur l'analyse de motifs de texture en utilisant une version étendue de la matrice de co-occurrences. Cela implique une voxelisation du nuage de points pour obtenir une représentation 3D régulière et le calcul de la matrice de co-occurrence 3D pour extraire les attributs de Haralick, qui sont utilisées pour entraîner des modèles d'apprentissage automatique pour prédire la qualité du nuage de points.

2 Approche proposée

La métrique proposée, baptisée CO-PCQA, suit le schéma donné par la figure 1 et est décrite ci-dessous :

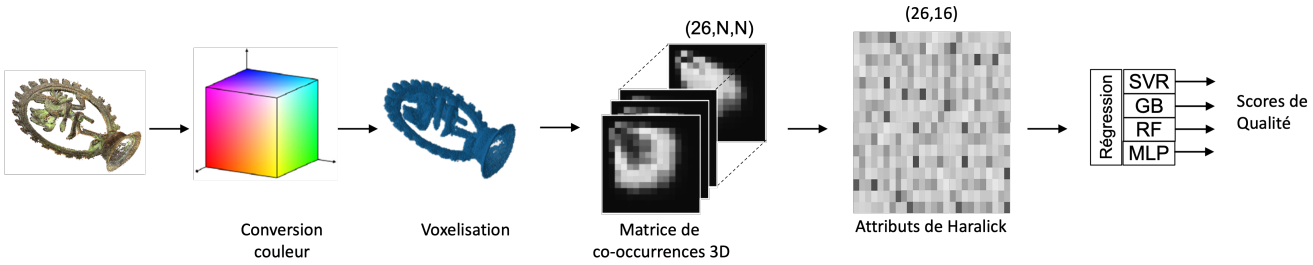


FIGURE 1 – *Diagramme de la métrique de qualité des nuages de points basée sur la matrice de co-occurrences (CO-PCQA). N représente le nombre de niveaux de gris.*

2.1 Conversion couleur

Dans cette étude, nous avons considéré 3 espaces colorimétriques différents effectuées avant l'étape de voxelisation. Tout d'abord, l'espace RVB utilisé dans diverses applications. Ensuite l'espace YUV aui permet de décorréler la luminance et la chrominance. Plusieurs modèles négligent la chrominance et se concentrent uniquement sur la luminance car elle a un impact plus important sur le Système Visuel Humain (SVH) [11]. Enfin, l'espace colorimétrique gaussien [12], considéré comme plus représentatif du fonctionnement du SVH. Ce dernier est formalisé comme suit :

$$\begin{bmatrix} \hat{E} \\ \hat{E}_\lambda \\ \hat{E}_{\lambda\lambda} \end{bmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

où $\hat{E}$ représente la luminance et $\hat{E}_\lambda$, $\hat{E}_{\lambda\lambda}$ deux canaux de chrominance.

2.2 Voxelisation

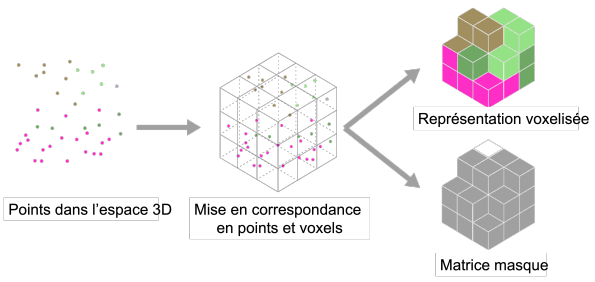


FIGURE 2 – *Illustration du processus de voxelisation, pour convertir la représentation d'objet 3D à partir de points distribués dans l'espace en une matrice régulière 3D (les cubes blancs représentent les trous).*

L'objectif principal de la voxelisation est de convertir un nuage de points, qui est un ensemble irrégulier de points, en un domaine cubique continu. Nous commençons par sélectionner la résolution (taille du voxel) pour déterminer la taille de la matrice résultante en fonction de la boîte englobante du nuage de points et de la résolution souhaitée. Le résultat de ce processus est une matrice tridimensionnelle, ainsi qu'un masque 3D qui indique si un voxel est vide, comme représenté dans la Figure 2.

2.3 Co-occurrence matrix

La matrice de co-occurrence est une matrice de dépendance spatiale des niveaux de gris qui caractérise la texture de l'image en quantifiant la fréquence d'apparition de paires de pixels ayant une valeur spécifique et une relation spatiale spécifiée, telle que la distance et l'angle [13]. C'est un extracteur de caractéristiques utile pour les problèmes basés sur la texture, y compris l'évaluation de la qualité.

Pour les images 2D, la matrice de co-occurrence est calculée comme suit : Pour une image en niveaux de gris I, la matrice de co-occurrence est une matrice carrée (C) de taille N, où N est le nombre de niveaux de gris dans l'image. La $(i,j)^{me}$ cellule de C représente le nombre de fois où un pixel P_1 avec une valeur d'intensité L_{P_1} est connecté à un pixel P_2 avec une valeur d'intensité L_{P_2} à une distance particulière k dans la direction θ.

Un pixel P peut avoir jusqu'à 8 pixels environnants, représentant 8 directions différentes dans une image 2D. Cependant, dans les nuages de points 3D, les points environnants ne correspondent pas nécessairement à des angles ou des directions spécifiques, mais plutôt à des points voisins qui se trouvent dans un certain rayon ou seuil de distance. Par conséquent, le nombre de directions possibles dans les nuages de points 3D dépend de la géométrie de la distribution de points et du seuil de distance choisi.

La voxelisation aide à pallier ce problème complexe en permettant de calculer la matrice de co-occurrence dans les 26 directions possibles en 3DCela donne lieu à 26 matrices de taille $(N \times N)$, où N représente le nombre de niveaux de gris. Chaque matrice 2D extraite de la matrice 3D correspond à l'une des 26 directions et est utilisée pour extraire les attributs de Haralick.

2.4 Attributs de Haralick et qualité

Les attributs de Haralick, telles que l'entropie, le contraste et le moment angulaire (ASM) de deuxième ordre, sont connus pour leur capacité à décrire les textures et leur corrélation avec la qualité visuelle. Cependant, dans ce cas particulier, les attributs sont calculés séparément pour chaque direction afin de permettre une plus grande précision, car chacune des 26 directions peut avoir un impact distinct sur la perception visuelle et potentiellement sur la qualité visuelle. Les attributs sont calculés à partir de la matrice de co-occurrence de niveaux de gris de chaque di-

rection, ce qui améliore encore la précision de l'analyse.
Dans notre étude, nous avons considéré 16 attributs pour chaque direction, ce qui donne un total de 416 pour chaque NP. Ces attributs sont utilisés comme entrée d'un régresseur pour les mapper à la qualité finale en utilisant des scores subjectifs.
La métrique de qualité proposée utilise les attributs précédemment extraits ainsi qu'une étape de régression. Nous avons adopté trois méthodes de régression différentes, à savoir la régression par forêt d'arbres décisionnels [14], la régression par gradient boosting [15] et la régression par vecteur de support [16]. De plus, nous avons mis en un perceptron multicouches, qui prend le vecteur d'attributs en entrée et estime le score de qualité final à l'aide des scores subjectifs fournis par les différents ensembles de données.

3 Expérimentations

3.1 Configuration expérimentale

Bases de données. Pour évaluer l'efficacité de CO-PCQA, nous avons utilisé deux bases de données largement reconnues comprenant des scores de qualité subjective. Ces bases de données sont décrites ci-dessous :

- **SJTU** [17] est basée sur 9 nuages de points MPEG dégradés avec 7 types de distorsions (bruit Gaussien, compression, etc.) à 6 niveaux différents.
- **WPC** contient 20 objets et 700 versions de chaque objet altérées par différentes distorsions (sous-échantillonnage, bruit Gaussien, compression, etc.)

Évaluation de la performance. Pour évaluer la performance de la prédiction, le score prédit est comparé à la vérité terrain en utilisant le coefficient de corrélation de Pearson (PLCC) pour la précision, et le coefficient de corrélation de Spearman (SRCC) pour la monotonie. La performance des métriques est évaluée en utilisant le principe de Pareto, avec 80% des données pour l'entraînement et les 20% restants pour le test. Cette procédure est répétée 1000 fois, et le résultat médian des itérations est considéré comme le résultat final et rapporté dans le Tableau 1.

3.2 Résultats

Pour analyser la performance de notre métrique proposée CO-PCQA, nous fournissons des scores notés CO-PCQA-* où -SVR, -RF, -GB et -MLP représentent l'utilisation de la régression par vecteurs de support, de la forêt d'arbres décisionnels, du gradient boosting et du perceptron multicouches, respectivement. Nous comparons nos scores à ceux obtenus à l'aide d'autres métriques de qualité avec référence et sans référence, comme indiqué dans le tableau 1. La comparaison est effectuée en utilisant l'espace de couleur YUV et une taille de voxel de 2.
Globalement, CO-PCQA donne les meilleurs résultats sur les bases SJTU et WPC avec un écart significatif entre le niveau de corrélation. Cette différence pourrait s'expliquer par le fait que la complexité et la diversité des nuages de points sont plus élevées dans WPC que dans SJTU.

TABLEAU 1 – *Performance de CO-PCQA comparée aux métriques FR et NR de l'état de l'art sur les bases de données SJTU-PCQA [17] et WPC [1]. Les meilleurs résultats sont en gras et les seconds meilleurs sont soulignés.*

Dataset /		SJTU-PCQA		WPC	
Approach	Metric	PLCC	SRCC	PLCC	SRCC
Avec référence	P2point	0.65	0.62	0.43	0.41
	P2plane	0.66	0.59	0.40	0.37
	GraphSIM	0.59	0.57	0.75	0.75
	PCQM	0.86	0.84	0.44	0.44
	LP-PCQM	<u>0.90</u>	0.88	0.71	0.72
Sans réference	ResSCNN	0.58	0.56	-	-
	IT-PCQA	-	-	0.55	0.54
	PQA-Net	0.85	0.82	0.70	0.69
	NR-PCQA	0.88	0.87	0.65	0.64
	CO-PCQA-SVR	0.82	0.78	0.69	0.75
	CO-PCQA-RF	0.89	0.79	0.71	0.74
	CO-PCQA-GB	<u>0.90</u>	**0.90**	**0.79**	**0.79**
	CO-PCQA-MLP	**0.91**	<u>0.89</u>	<u>0.78</u>	<u>0.76</u>

En se basant sur les métriques de performance présentées dans le tableau, la méthode CO-PCQA démontre une supériorité claire par rapport aux métriques sans référence testées, et fonctionne de manière comparable ou meilleure que les métriques avec référence. Les meilleures performances sont obtenues avec la méthode basée sur MLP sur la base de données SJTU, avec des valeurs PLCC et SRCC de 0,91 et 0,89, respectivement. Sur la base de données WPC, le gradient boosting obtient les meilleurs résultats avec des valeurs PLCC et SRCC de 0,79 chacune. Ces résultats indiquent que l'utilisation des statistiques de co-occurrences et des attributs de Haralick peut prédire efficacement la qualité des nuages de points. Parmi les métriques de pointe, NR-PCQA et PQA-Net fournissent de bonnes performances sur les deux bases de données, mais restent globalement en dessous de CO-PCQA-GB et CO-PCQA-MLP. Les méthodes basées sur des points tels que P2point et P2plane de MPEG sont moins efficaces car elles ne considèrent que la distribution de points et ne tiennent pas compte des distorsions de couleur ou de l'information structurelle. Enfin, ResSCNN et IT-PCQA montrent les pires performances parmi les métriques sans référence.
Comparée aux méthodes précédemment mentionnées, CO-PCQA offre plusieurs avantages. Tout d'abord, elle utilise efficacement les caractéristiques de Haralick pour décrire la structure texturale d'un objet, ce qui améliore la précision du modèle. De plus, CO-PCQA considère les 26 directions individuellement, plutôt que de se fier à des caractéristiques globales, ce qui donne un modèle plus solide avec une précision améliorée. Dans l'ensemble, ces caractéristiques font de CO-PCQA une méthode très efficace pour l'analyse texturale des nuages de points.

3.3 Ablation study

Pour mieux comprendre le comportement de la métrique CO-PCQA proposée, nous avons mené une étude par abla-

tion afin d'examiner l'impact des différents paramètres impliqués dans la métrique. A cause de la limitation de page, nous ne donnerons que les résultats les plus importants utilisant la version basée sur SVR.

Taille du voxel. C'est un paramètre critique dans CO-PCQA car il détermine la précision du format voxelisé. Des voxels de taille 2, 3, 5, et 10 ont été testé et performance inversement proportionnelle a été notée. Ce qui était prévisible au vu de la variation de la précision du nuage de points. Un compromis entre performance et complexité devra être trouvé en fonction de l'application ciblée.

Représentation de la couleur. L'espace couleur peut avoir un impact sur la performance des métriques de qualité. Ici, nous évaluons la performance de l'utilisation d'une luminance seule et de trois espaces couleur différents à savoir RVB, YUV et GCS. Les résultats ont montré une similarité des performances avec un léger avantage pour GCS et une nette supériorité par rapport à la luminance seule.

Direction des co-occurrences. Cela représente l'angle d'occurrence entre un voxel et ses voisins. Nous avons ainsi combiné les angles compatibles et mesuré les performances associées. Il s'avère que l'utilisation de certaines directions privilégiées de manière indépendante génère une baisse significative de performance par rapport aux résultats du Tableau 1. Cela indique que chaque direction apporte une force supplémentaire à l'approche globale.

4 Conclusion

Cet article présente une nouvelle métrique appelée CO-PCQA qui évalue la qualité des nuages de points sans référence, basée sur des matrices de co-occurrence. Nous avons adapté ces matrices 2D basées aux nuages de points 3D et calculé les attributs de Haralick en fonction des angles d'occurrence. Nous avons expérimenté plusieurs méthodes de régression, notamment SVR, Random Forest, Gradient Boosting et Multilayer Perceptron. Les résultats expérimentaux montrent que notre modèle surpasse la plupart des métriques de référence complète et celles sans référence. Les meilleurs résultats ont été obtenus avec GB et MLP pour SJTU et WPC, respectivement. Bien que le coût de calcul de cette méthode soit élevé en raison des matrices multidimensionnelles et des caractéristiques statistiques à calculer, en plus de l'étape de voxelisation, les résultats obtenus montrent que la métrique fournit des prédictions précises sans aucune référence. Un compromis doit être trouvé entre le coût de calcul et la précision de la performance, en fonction de l'application et/ou de l'appareil ciblé. Plusieurs aspects restent à explorer, tels que l'incorporation de la saillance lors du calcul des matrices de co-occurrence.

Références

[1] H.L. Su, Z. F. Duanmu, W. Liu, Q. Liu, et Z. Wang. Perceptual quality assessment of "3d" point clouds. Dans *IEEE ICIP*, pages 3182–3186, 2019.

[2] R.N. Mekuria, Z. Li, C. Tulvan, et P. Chou. Evaluation criteria for pcc (point cloud compression). 2016.

[3] G. Meynet, J. Digne, et G. Lavoué. Pc-msdm : A quality metric for 3d point clouds. Dans *IEEE QoMEX*, pages 1–3, 2019.

[4] A. Hore et D. Ziou. Image quality metrics : Psnr vs. ssim. Dans *IEEE ICPR*, pages 2366–2369, 2010.

[5] G. Meynet, Y. Nehmé, J. Digne, et G. Lavoué. Pcqm : A full-reference quality metric for colored 3d point clouds. Dans *IEEE QoMEX*, pages 1–6, 2020.

[6] R. Diniz, P. G. Freitas, et M. Farias. Towards a point cloud quality assessment model using local binary patterns. Dans *IEEE QoMEX*, pages 1–6, 2020.

[7] P. G. Freitas, S. Alamgeer, et M. Akamine, W.and Farias. Blind image quality assessment based on multiscale salient local binary patterns. Dans *ACM Multimedia Systems*, pages 52–63, 2018.

[8] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, et J. Sun. Predicting the perceptual quality of point cloud : A 3d-to-2d projection-based exploration. *IEEE Trans. on Multimedia*, 2020.

[9] Z. Zhang, W. Sun, X. Min, T. Wang, W. Lu, et G. Zhai. No-reference quality assessment for 3d colored point cloud and mesh models. *arXiv preprint arXiv :2107.02041*, 2021.

[10] Qi Liu, Hui Yuan, Honglei Su, Hao Liu, Yu Wang, Huan Yang, et Junhui Hou. Pqa-net : Deep no reference point cloud quality assessment via multi-view projection. *IEEE TCSVT*, 31(12) :4645–4660, 2021.

[11] M. Khosravy, N. Gupta, N. Marina, I. K Sethi, et M.R. Asharif. Perceptual adaptation of image based on chevreul–mach bands visual phenomenon. *IEEE Signal Processing Letters*, 24(5) :594–598, 2017.

[12] J.-M. Geusebroek, R. Van Den Boomgaard, A. WM Smeulders, et A. Dev. Color and scale : The spatial structure of color images. Dans *ECCV*, pages 331–341, 2000.

[13] L. Nanni, S. Brahnam, S. Ghidoni, E. Menegatti, et T. Barrier. Different approaches for extracting information from the co-occurrence matrix. *PloS one*, 8(12), 2013.

[14] L. Breiman. Random forests. *Machine learning*, 45 :5–32, 2001.

[15] T. Chen et C. Guestrin. Xgboost : A scalable tree boosting system. Dans *ACM int. Conf. on knowledge discovery and data mining*, pages 785–794, 2016.

[16] O.L. Mangasarian et D.R. Musicant. Robust linear and support vector regression. *IEEE TPAMI*, 22(9) :950–955, 2000.

[17] Y. Liu, Q. Yang, Y. Xu, et L. Yang. Point cloud quality assessment : Large-scale dataset construction and learning-based no-reference approach. *arXiv preprint arXiv :2012.11895*, 2020.

Transformer multimodal pour la détection du stress

Kevin Feghoul [1,2], *Deise Santana Maia* [2], *Mohamed Daoudi* [2,3], *Ali Amad* [1]
[1]Univ. Lille, Inserm, CHU Lille, UMR-S1172 - LilNCog, F-59000 Lille, France
[2]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[3]IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France
{kevin.feghoul, deise.santanamaia, mohamed.daoudi, ali.amad}@univ-lille.fr

Résumé

Le stress peut entraîner des conséquences nocives sur la santé mentale et physique des individus, ainsi que sur leur qualité de vie en général. Par conséquent, il est bénéfique de développer des outils automatisés pour aider à y faire face. Dans cette perspective, nous avons proposé différentes architectures Transformer multimodales sur l'ensemble de données WESAD afin de détecter le stress de manière automatique. Les résultats de notre étude démontrent l'adaptabilité des modèles proposés à cette tâche. En utilisant la méthode de fusion intermédiaire, nous avons dépassé l'état de l'art, avec une précision de $98,69\%$ *et un score F1 de* $98,73\%$*. Les résultats obtenus mettent en évidence l'efficacité de notre méthode et ouvrent des perspectives intéressantes pour le développement de techniques de reconnaissance des émotions basées sur des architectures Transformer multimodales.*

Mots clefs

Multimodal, Transformer, Stress, Données physiologiques

1 Introduction

Selon la *National Institutes of Health*, le stress se définit comme une réponse du corps à une pression physique, mentale ou émotionnelle. Les facteurs de stress peuvent inclure des pressions liées au travail, des difficultés financières, des problèmes relationnels, et bien plus encore [1]. Le stress peut être classé en deux catégories en fonction de sa durée : (1) le stress aigu, qui est de courte durée et qui est provoqué par un événement inhabituel ou une menace immédiate ; et (2) le stress chronique, qui est une réponse prolongée de l'organisme face à des facteurs de stress maintenus sur une longue période. Le stress chronique peut augmenter la susceptibilité à certains types de cancer [2], ralentir la guérison des plaies [3], et accroître la vulnérabilité aux infections [4].

Étant donné l'impact généralisé du stress sur les individus et la société dans son ensemble, il est de plus en plus important de développer des outils de détection automatique du stress. Ces outils permettront aux professionnels de santé de prendre des mesures préventives et d'offrir des traitements adaptés aux patients présentant des signes de stress. Pour développer de tels outils, il est important de comprendre que le stress est une réaction physiologique déclenchée par le système nerveux sympathique (SNS) en réponse à un stimulus, qui déclenche une réaction hormonale en cascade. Cette réaction hormonale implique la libération d'hormones telles que l'ACTH, le cortisol et l'adrénaline. À la suite de cette libération d'hormones, une accélération de la fréquence cardiaque et respiratoire, ainsi qu'une tension musculaire peuvent être observées.

Grâce aux récentes avancées dans les techniques d'apprentissage automatique, l'apprentissage profond est désormais largement utilisé pour le traitement des séries temporelles [5–7]. Les méthodes d'apprentissage profond offrent plusieurs avantages, notamment : (1) la capacité à capturer des caractéristiques complexes dans les séries temporelles, ce qui peut être difficile à détecter avec les méthodes traditionnelles ; (2) la possibilité de modéliser une large gamme de séries temporelles, telles que des données continues et discrètes, des séries multivariées et des séries à fréquence variable ; et (3) la capacité de faire des prévisions à long terme.

Comme le stress peut être détecté à l'aide de plusieurs types de capteurs, chacun possédant des propriétés différentes en termes de fréquence et de types de signaux enregistrés, nous avons traité la détection du stress comme un problème multimodal en combinant les signaux provenant de différents capteurs. Dans cette étude, nous nous sommes intéressés au modèle d'apprentissage profond Transformer [8] pour la tâche de détection automatique du stress, en adoptant une approche multimodale. Nous avons effectué nos expérimentations sur le jeu de données WESAD [9].

Les contributions de ce travail sont doubles et peuvent être résumées comme suit : (1) évaluation comparative des différentes méthodes de fusion de signaux physiologiques multimodaux en utilisant le modèle Transformer ; (2) amélioration significative des résultats par rapport à l'état de l'art.

2 Etat de l'art

Au cours des dernières années, de nombreuses études se sont intéressées à l'utilisation des données physiologiques pour la détection du stress. Le jeu de données WESAD [9] est couramment utilisé dans les travaux de recherche relatifs à cette problématique. Les auteurs de [9] ont proposé

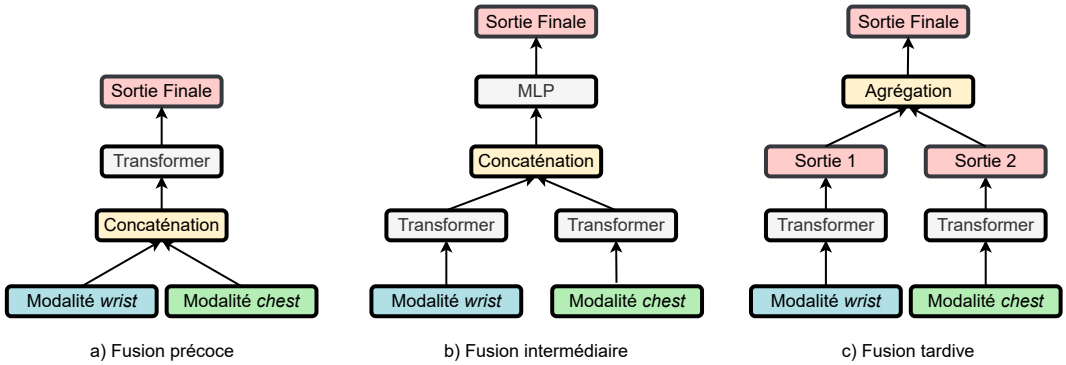


FIGURE 1 – *Illustration des trois principales stratégies de fusion multimodale utilisant le Transformer, à savoir la fusion précoce (a), la fusion intermédiaire (b) et la fusion tardive (c).*

un benchmark complet utilisant des statistiques provenant des domaines temporel et fréquentiel pour entraîner différents modèles d'apprentissage automatique traditionnels.

Plusieurs autres approches basées sur l'apprentissage profond ont été proposées en utilisant le jeu de données WESAD. Samyoun et al. [10] ont proposé d'utiliser des réseaux de neurones de type GAN, RNN et MLP pour traduire les signaux physiologiques du capteur du poignet en signaux provenant des capteurs placés au niveau de la poitrine. Les données traduites ont ensuite été utilisées pour détecter le stress à l'aide de méthodes d'apprentissage automatique. Gil-Martin et al. [11] ont proposé l'utilisation d'un réseau de neurones de type CNN pour la détection du stress, ainsi que l'analyse de plusieurs techniques de traitement du signal pour générer les entrées du modèle. Dans leur étude, Huynh et al. [12] ont proposé l'utilisation d'un schéma d'entraînement de réseau de neurones profonds optimisé basé sur des CNN, en utilisant la méthode de recherche d'architecture neuronale. L'étude menée par Lai et al. [13] a utilisé un réseau de neurones à convolution temporelle résiduelle pour traiter les différents signaux, et a proposé plusieurs stratégies de fusion. Wu et al. [14] ont étudié l'utilisation de matrices SPD pour fusionner efficacement des signaux physiologiques et comportementaux, permettant ainsi de capturer simultanément les informations de corrélation au sein et entre les différentes modalités. Les résultats de leur étude ont démontré l'impact positif de l'utilisation de plusieurs modalités sur les performances par rapport à l'utilisation d'une seule modalité.

Pour notre part, nous proposons l'utilisation d'un Transformer multimodal pour la tâche de détection du stress, où nous avons traité chaque ensemble de signaux provenant d'un capteur comme une modalité distincte.

3 Approche proposée

La présente section expose les différentes techniques de fusion multimodale que nous avons employées pour la détection du stress, en utilisant le Transformer. Ces techniques incluent la fusion précoce, la fusion intermédiaire et la fusion tardive, lesquelles sont illustrées dans la Figure 1.

3.1 Transformer

Le modèle Transformer est une architecture de réseau de neurones introduite par [8] qui est maintenant considérée comme la référence en matière de traitement de tâches liées au langage naturel. De plus, le Transformer a été étendu avec succès à d'autres domaines tels que le traitement des séries temporelles [15] et l'apprentissage multimodal [16]. Le Transformer repose sur le mécanisme d'auto-attention, qui permet au modèle de se concentrer sur différentes parties de la séquence d'entrée pour effectuer des prédictions. Ce mécanisme calcule une somme pondérée de la séquence d'entrée, où les poids sont appris pendant l'entraînement. Grâce à cela, le modèle est capable de capturer des dépendances à long terme et de faire des prédictions en se basant sur l'ensemble de la séquence d'entrée plutôt que sur des représentations passées limitées, contrairement aux réseaux de neurones récurrents.

En plus du mécanisme d'auto-attention, le Transformer utilise également des techniques telles que l'attention multi-tête et l'encodage de position pour améliorer sa performance en terme de prédiction. L'attention multi-tête permet au modèle de calculer plusieurs poids d'attention en parallèle pour différents sous-espaces de caractéristiques de l'entrée, ce qui permet au modèle de mieux capturer les relations entre les différentes parties de la séquence. L'encodage de position est une technique qui permet au modèle de prendre en compte l'ordre des éléments de la séquence, en ajoutant des informations sur leur position relative. Cela permet au modèle de mieux comprendre les relations entre les différentes parties de la séquence et de capturer les informations séquentielles.

3.2 Transformer multimodal

Nous avons adopté une approche multimodale basée sur le modèle Transformer pour détecter le stress. Cette méthode nous a permis d'intégrer de manière efficace différents signaux physiologiques provenant de divers capteurs.

L'un des principaux défis de l'apprentissage automatique multimodal est la fusion efficace de données provenant de

TABLEAU 1 – *Détection du stress : comparaison avec des méthodes de pointe.*

Methodes	Wrist		Chest	
	Acc	F1 score	Acc	F1 score
Schmidt et al. [9]	87.12	84.11	92.83	91.07
Samyoun et al. [10]	89.90	87.60	91.10	90.20
Gil-Martin et al. [11]	92.70	92.55	93.10	93.01
Huynh et al. [12]	93.14	-	-	-
Wu et al. [14]	94.65	93.99	95.54	94.76
Lai et al. [13]	94.16	93.62	**96.69**	**96.61**
Transformer	**95.74 ± 7.22**	**96.4 ± 5.34**	96.07± 4.81	95.93 ± 4.96

TABLEAU 2 – *Détection du stress (approches multimodales) : comparaison avec des méthodes de pointe.*

Methodes	Wrist + Chest	
	Acc	F1 score
Schmidt et al. [9]	92.28	90.74
Samyoun et al. [10]	94.70	93.40
Gil-Martin et al. [11]	96.62	96.63
Wu et al. [14]	96.88	96.44
Lai et al. [13]	97.75	97.74
MMT-early (ours)	98.13 ± 3.00	98.07 ± 3.09
MMT-inter (ours)	**98.69 ± 2.85**	**98.73 ± 2.62**
MMT-late (ours)	98.34 ± 3.31	98.33 ± 3.28

sources différentes. Les stratégies de fusion multimodale sont généralement classées en trois catégories : la fusion précoce, la fusion intermédiaire et la fusion tardive, chacune ayant ses avantages et ses inconvénients en fonction de la tâche à accomplir et des caractéristiques des données. En ce qui concerne la fusion précoce, les deux modalités d'entrée sont d'abord concaténées avant d'être traitées par un Transformer. Pour la fusion intermédiaire, en revanche, les deux modalités sont traitées de manière indépendante par un Transformer, permettant ainsi de découvrir les corrélations intra-modales avant de les fusionner pour découvrir les corrélations inter-modales. Les caractéristiques extraites sont ensuite concaténées et traitées par un réseau de neurones de type MLP pour la classification finale. Enfin, pour la fusion tardive, les deux modalités sont traitées par un Transformer jusqu'à la prédiction, suivie d'une fonction d'agrégation pour la prédiction finale. Cette approche s'avère utile lorsque les modalités d'entrée sont très différentes et ne peuvent pas être facilement combinées en une représentation conjointe. Nous désignerons respectivement ces méthodes sous les noms de MMT-early, MMT-inter et MMT-late (MMT pour *Multimodal Transformer*).

4 Résultats expérimentaux

4.1 Jeu de données

WESAD est un ensemble de données multimodal bien connu pour la détection du stress et de l'affect. Il contient des données physiologiques et de mouvement provenant de 15 sujets, qui ont été capturées à l'aide d'un bracelet Empatica E4 porté au poignet (*wrist*) et d'un dispositif RespiBAN placé au niveau de la poitrine (*chest*). Le bracelet Empatica enregistre l'activité électrodermale (EDA), le volume sanguin pulsé (BVP), la température corporelle (TEMP) et l'accélération sur trois axes (ACC) à des fréquences respectives de 4, 64, 4 et 32 Hz. En complément, le RespiBAN mesure l'électrocardiogramme (ECG), l'électromyographie (EMG), la respiration (RESP), la température de la peau (TEMP), l'EDA et l'ACC, échantillonnés à une fréquence de 700 Hz.
Le protocole d'étude a été conçu pour induire trois états émotionnels chez les sujets : neutre, stressé et amusé. En nous appuyant sur des travaux antérieurs [9–14], nous avons formulé un problème de détection de stress binaire (stress vs non-stress) en utilisant les séquences de stimuli catégorisé comme neutres et amusants pour constituer la classe "non-stress", conformément à la littérature.

4.2 Prétraitement

Nous avons tout d'abord appliqué un filtre passe-bas aux différents signaux physiologiques, afin de réduire le bruit et de conserver les fréquences d'intérêt. Ensuite, nous les avons sous-échantillonnés à une fréquence de 4 Hz. Les signaux ont ensuite été segmentés en fenêtres glissantes de 60 secondes, sans chevauchement entre les fenêtres successives. Chaque échantillon correspond à l'ensemble des signaux provenant des deux capteurs pendant une période de 60 secondes, ce qui donne un total de 240 points par échantillon.

4.3 Résultats

Conformément aux travaux antérieurs sur WESAD [9–14], nous avons utilisé la stratégie d'évaluation *Leave-One-Subject-Out Cross Validation* (LOSO-CV) pour valider nos différents modèles.
Le modèle Transformer a surpassé toutes les autres méthodes en termes de précision et de score F1 lorsqu'il a été entraîné avec les données provenant du bracelet, dépassant de près de $1,09\%$ et $2,41\%$ respectivement le modèle affichant la meilleure performance [14].
En ce qui concerne le capteur attaché au niveau de la poitrine, notre modèle se classe deuxième en termes de performance, avec une précision et un score F1 inférieurs de seulement $0,62\%$ et $0,68\%$ respectivement par rapport à [13].
Nous avons observé que les méthodes basées sur l'apprentissage profond [11–14] ont obtenu de meilleures performances que les méthodes basées sur l'apprentissage automatique [9, 10] pour les deux capteurs.
De plus, nous avons réalisé des expériences en utilisant à la fois les données provenant du bracelet et celles du capteur attaché à la poitrine. Les résultats de ces expériences sont présentés dans le tableau 2. Nous pouvons constater que les trois modèles que nous proposons, à savoir MMT-early, MMT-inter et MMT-late, affichent des performances

supérieures à toutes les autres méthodes. En particulier, le modèle MMT-inter a obtenu les meilleurs résultats pour les deux métriques d'évaluation, surpassant le modèle de [13] avec une amélioration de $0,94\%$ et $0,99\%$ en termes de précision et de score F1, respectivement. Ces résultats confirment notre hypothèse selon laquelle l'utilisation de modèles multimodaux est appropriée pour le traitement de groupes de signaux provenant de différents capteurs.

5 Conclusion

Dans cette étude, nous avons traité de la détection du stress en utilisant des architectures Transformer multimodales. À la suite de multiples expérimentations sur l'ensemble de données WESAD, nous avons établi un nouvel état de l'art, en atteignant des taux de précision et de score F1 respectifs de $98,69\%$ et $98,73\%$. Dans un futur travail, nous envisageons d'étendre nos recherches à l'utilisation de différents types de données tels que des données vidéo, audio et textuelles pour des tâches d'informatique affective.

Références

[1] Longfei Yang, Yinghao Zhao, Yicun Wang, Lei Liu, Xingyi Zhang, Bingjin Li, et Ranji Cui. The effects of psychological stress on depression. *Current neuropharmacology*, 13(4) :494–504, 2015.

[2] Alison N Saul, Tatiana M Oberyszyn, Christine Daugherty, Donna Kusewitt, Susie Jones, Scott Jewell, William B Malarkey, Amy Lehman, Stanley Lemeshow, et Firdaus S Dhabhar. Chronic stress and susceptibility to skin cancer. *Journal of the National Cancer Institute*, 97(23) :1760–1767, 2005.

[3] Ronald Glaser et Janice K Kiecolt-Glaser. Stress-induced immune dysfunction : implications for health. *Nature Reviews Immunology*, 5(3) :243–251, 2005.

[4] Sheldon Cohen, David AJ Tyrrell, et Andrew P Smith. Psychological stress and susceptibility to the common cold. *New England journal of medicine*, 325(9) :606–612, 1991.

[5] John Cristian Borges Gamboa. Deep learning for time-series analysis. *arXiv preprint arXiv :1701.01887*, 2017.

[6] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, et Pierre-Alain Muller. Deep learning for time series classification : a review. *Data mining and knowledge discovery*, 33(4) :917–963, 2019.

[7] Guansong Pang, Chunhua Shen, Longbing Cao, et Anton Van Den Hengel. Deep learning for anomaly detection : A review. *ACM computing surveys (CSUR)*, 54(2) :1–38, 2021.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, et Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[9] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, et Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. Dans *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[10] Sirat Samyoun, Abu Sayeed Mondol, et John A Stankovic. Stress detection via sensor translation. Dans *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 19–26. IEEE, 2020.

[11] Manuel Gil-Martin, Ruben San-Segundo, Ana Mateos, et Javier Ferreiros-Lopez. Human stress detection with wearable sensors using convolutional neural networks. *IEEE Aerospace and Electronic Systems Magazine*, 37(1) :60–70, 2022.

[12] Lam Huynh, Tri Nguyen, Thu Nguyen, Susanna Pirttikangas, et Pekka Siirtola. Stressnas : Affect state and stress detection using neural architecture search. Dans *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 121–125, 2021.

[13] Kenneth Lai, Svetlana N Yanushkevich, et Vlad P Shmerko. Intelligent stress monitoring assistant for first responders. *IEEE Access*, 9 :25314–25329, 2021.

[14] Yujin WU, Mohamed Daoudi, Ali Amad, Laurent Sparrow, et Fabien D'Hondt. Fusion of physiological and behavioural signals on spd manifolds with application to stress and pain detection. *arXiv preprint arXiv :2207.08811*, 2022.

[15] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, et Wancai Zhang. Informer : Beyond efficient transformer for long sequence time-series forecasting. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.

[16] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, et Boqing Gong. Vatt : Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34 :24206–24221, 2021.

X-RCRNet: an explainable deep learning network for COVID-19 detection using ECG beat signals

Marc Junior Nkengue [1,2], Xianyi Zeng[1], Ludovic Koehl[1], Xuyuan Tao[1]
[1]Univ. Lille, ENSAIT, Laboratoire Génie et Matériaux Textile (GEMTEX), F-59000, Lille, France
[2]Univ. Lille, Ecole Centrale Lille, F-59000, Lille, France
marc-junior.nkengue@ensait.fr

Abstract

Wearable systems measuring human physiological indicators with integrated sensors and supervised learning-based medical image analysis (e.g. ECG, X-ray, CT or ultrasound images for lung or the chest) have been considered relevant tools for COVID-19 monitoring and diagnosis. However, these two technical roadmaps have their respective advantages and drawbacks. The current wearable systems enable to realize real-time monitoring of COVID-19 but are limited to its basic symptoms only, neither allowing to distinguish it from other diseases nor performing deep analysis. Current medical image analysis can provide accurate decision support for diagnosis but rarely deals with real-time data processing. In this context, we propose a new wearable system by combining the advantages of these two technical roadmaps. Considering that electrocardiogram (ECG) has been proved relevant to evolution of COVID-19 symptoms, the proposed wearable system will integrate an explainable Deep Neural Network to realize online monitoring of COVID-19 gravity by using ECG beat signal. This paper will focus on the Deep Neural Network model named X-RCRNet. The network is based on ResNet18 but with few enhancements: 1) LSTM Layers for regenerating the block memory and further extracting the involved time-varying features; 2) LeakyReLU for increasing the performances of the model. With an accuracy of 99.24 % after experiments, our model has not only outperformed the existing methods in terms of accuracy and robustness, but also originally identify the ST interval of the ECG pattern, as the most prominent key features affected by the virus.

Mots clefs

Machine learning, data augmentation, Deep Learning, Multiclass Classification, COVID-19, Signal Processing.

1 Introduction

Three years after its emergence in late 2019, severe acute respiratory syndrome coronavirus 2 (Sars-CoV-2) or COVID-19 infected more than 630 million people, causing more than 6 million deaths [1]. However, the symptoms of COVID-19 patients differ from one variant to another in this long duration. For all reported variants and all periods, the most serious symptoms are shortness of breath (blood oxygen level<92%) and heart failure (heart rate>90 bpm),[2-6]. Although the intensity of infection and symptoms have attenuated thanks to the vaccination and follow-up of barrier gestures, we are still far from the termination of the pandemic. This is mainly due to the high infection rate, the proliferation of its variants that can escape from vaccination coverage, and the inability to detect the virus in real-time and thus, control its proliferation. This situation promotes the emergence of remote monitoring and diagnosis tools using the IoT (Internet of Things), including wearable systems, and supervised learning with medical images (ECG images, X-ray, CT or ultrasound images for lung or the chest). Currently, these two categories of tools have effectively reduced the pressure of medical resources (e.g., medical doctors, healthcare staff, devices, materials, etc.). However, they are unable to perform reliable real-time monitoring and analysis for supporting quick decisions of medical professionals in complex diagnosis scenarios. The current wearable systems mainly realize real-time detection for the basic symptoms of COVID-19 from skin temperature, blood oxygen saturation (SPO_2) level, and heart rate. Since these symptoms are common with other diseases (e.g. Flu), it is impossible to distinguish COVID-19 from others without further investigation [7]. Medical image analysis with supervised learning is highly accurate and capable of providing reliable decision support to medical professionals. However, real-time processing with these images has rarely been involved in the existing research work due to its heavy computational load and complex conditions for supervised learning data acquisition (e.g. radiation exposure, requirement for qualified technicians and experts for image interpretation). Furthermore, the prediction results delivered by a supervised learning model are usually inexplainable from a medical viewpoint, which cannot provide relevant decision support for medical diagnosis. A new approach combining the advantages of real-time processing with wearable systems and accurate analysis of medical images with supervised learning models and high interpretation capacities will be significant for providing more efficient tools combatting COVID-19.

Currently, we consider that heart failure is one of the most prominent symptoms of COVID-19 [5]. Heart failure can be described using the Electrocardiogram (ECG) signal, especially beat signal (beat signal contains all information about cardiac condition). Different from ECG images, X-ray or CT images requiring huge devices for data acquisition, ECG signals can be measured and processed in

real-time by using portable tiny sensors. The analysis of ECG beat signals using an AI-based supervised learning model will be significant for COVID-19 monitoring. Our idea is to design and implement a wearable system coupling with an explainable supervised learning model named X-RCRNet (Explainable Residual Convolutional and Recurrent Network), to monitor the patient's symptoms from his/her ECG beat signals and make appropriate decision support in real-time.

The main contributions of this paper are summarized below.

- At the supervised learning level, the performance of accuracy and robustness will be improved related to the existing work by introducing LeakyReLU (Leaky version of a Rectified Linear Unit) to avoid vanishing gradient and LSTM (Long-Short-Term Memory) units to extract temporal features.
- At the data acquisition level, a new open database of ECG beat signals for COVID-19 patients will be set up, enabling the collect more relevant data on infected patients.
- At the result interpretation level, the proposed super vised learning model will enable to accurately explain the results of classification for ECG beat signals.

2 The proposed model

2.1 Data-prepocessing

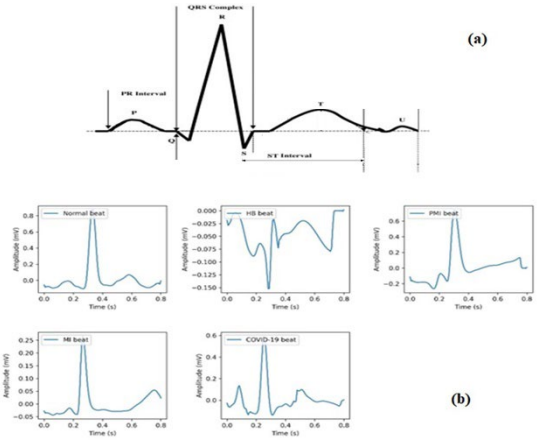


Figure 1 – A Normal Beat Signal (a disease-free regular ECG heartbeat signal and its decomposition); (b) beat signal samples per class

The original dataset used in this study is an ECG image dataset of cardiac and COVID-19 patients [8]. It consists of 1937 distinct patient records, with five distinct classes: Normal, COVID-19, myocardial infarction (MI), abnormal heartbeat (HB) and history of myocardial infarction (PMI). The data were collected using the ECG device **EDAN SERIES-3**. The device collected 12 leads ECG trace images, sampling at 500 Hz. Each lead has a 2.5 second duration and the total duration length is 10 s on the 12 lead ECG images. The current dataset is not suitable for our wearable system due to data mismatch (Our single lead wearable ECG sensor measure 1d signal, X-RCRNet need to be train with 1d signal from a single lead instead of images). The first contribution of this paper will be a new single lead ECG beat signal dataset by converting single lead ECG trace images to 1d signals and extract heartbeat for each signal. Lead II ECG images was chosen for the image-signal conversion for two reasons: 1) The wearable ECG sensor of our system is a single lead ECG sensor; 2) Lead II is the lead that describes the best all ECG beat waves. The next lines describe the dataset creation process.

The pre-processing was done by following three steps:

- **Image segmentation:** Image cropping near the lead II area has been performed. is a black-and-white image with the ECG signal as the foreground has been obtained by using Otsu Method.
- **Image to signal conversion:** The signal was extracted by detecting the foreground pixel. The resulting signal has been down sampled to 125 Hz (Suitable for both time-domain analysis [34] and wearable systems [32]) . The signal is smoothed by convoluting the signal with a Hanning Windows [35]. The Hanning window was chosen for two reasons :1) Good frequency resolution; 2) Spectral leakage reduction, especially for non-linear signals.

- **Heartbeat Extraction and data augmentation**: The heartbeats were extracted using the Neurokit2 toolbox[9]. The toolbox use Discrete Wavelength Transform (DWT) [10] to detect the P-waves, the R-peaks, and the T-waves Each heartbeat is centered around an R-peak. The interval from the P wave onset (The beginning of the P-wave) to the T-wave offset (The end of the T-wave) represents the beat duration. As show by Figure 1 the main features of an ECG beat are the PR interval, the QRS interval and the ST interval. The data augmentation has been done by applying different operations, like jittering, scaling, permutation, magnitude and time warping, resampling [11] We were able to generate up to 316 368 signals per class. To our best knowledge, this is the first public ECG beat signal dataset for COVID-19 patients. After the dataset creation, we proceed to the model training and evaluation.

2.2 Proposed framework

Illustrated by Figure 2,the model architecture can be divided into two parts:

- **Features extraction block:** The purpose is to

extract the signal spatial and temporal features using residual blocks. Inspired by ResNet, our model residual block possesses significant differences: 1) LeakyReLU is applied instead of ReLU. Indeed, negative values of the ECG beat signal will be replaced by zero. LeakyRELU the range of ReLU and allows negative values to be process; 2) LSTM Layer is applied before the pooling operation, to extract temporal features. Applying the LSTM Layer regenerate the residual block memory to retain the backpropagation error between the time step and the level. The operation preserves the learning state in multiple time steps and improve the ability to extract temporal features.

- **Classification block:** After flattening the feature extraction block output, the fully connected layer transformed the data into a vector of numerical values corresponding to the outputs for each class.

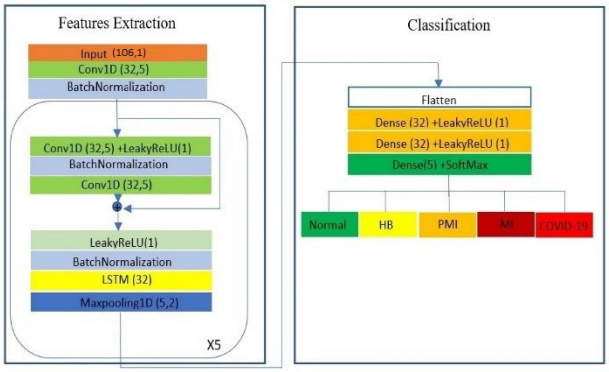


Figure 2 ─ XRCNet architecture

3 Results

Figure 3 shows a convergence of the training and validation losses curves, with a final value of 0.0937 and confirms our model robustness. These results show the ability of our model to clearly identify COVID-19 patient heartbeat.

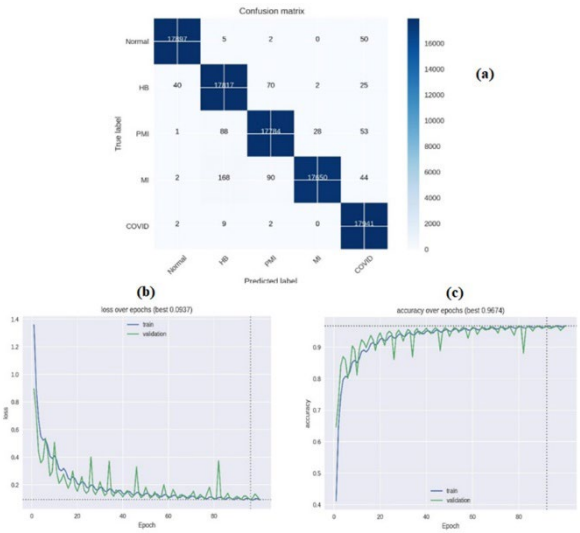


Figure 3 : (a) Confusion Matrix ; (b) Loss curve ; (c) Accuracy curve

We compared our model with related studies with a five classes classification. Table 1 Show our model outperforms the existing ECG images classification tools for COVID-19. It also shows the possible use of ECG beat signal to monitor COVID-19 patients.

Reference	Sensitivity	Precision	Specificity	Accuracy
Proposed framework	99.24%	99.24%	99.24%	99.80%
[12]	96.00 %	90.58%	90.00%	93.00%
[13]	91.70 %	91.90%	95.90%	91.73%
[14]	90.80%	91.90%	92.80%	90.79%

Table 1 : Comparison with related studies

At local level, the model explainability was performed using an approach called Gradient Weighted Class Activation Mapping ++ (Grad-CAM ++), is proposed in [15] . Grad-CAM visualize the gradients of the final layer of the Network in a heatmap. The heatmap can be used for analyzing factors that influence the classification result, and thus, help visualize where the network is looking. the main issue of Grad-CAM++ is the inability to give a dataset-level explanation. By using Grad-CAM++ we propose to compute, for each key feature, his frequency as the most prominent feature for each sample.

Table 2 represents each key feature occurrence frequency,as the most important feature across the dataset. The results shows that X-RCRNet identifies the S-T interval as the most relevant feature for COVID-19 classification at the dataset level. Other researchers [2, 4, 5] confirm that the manifestation of a severe case of COVID-19 patients occurs in the ST interval. To our best knowledge, X-RCRNet is the first Deep Learning model to confirm ST interval of ECG pattern as the feature affected by the COVID-19 at a dataset-level.

Keys features	P-R interval	QRS interval	ST interval
key feature occurrence frequency, as the most important feature (%)	4.75	7.90	87.35

Table 2:Feature importance at dataset-level. The key feature with the highest occurrence is the most prominent key feature at the dataset level

4 Conclusion

This paper presents X-RCRNet, a novel explainable deep neural network, based in ResNet18 for COVID-

19 patient symptoms monitoring using an ECG beat signal. The first known open database of ECG beat signals for COVID-19 patients has been created to conduct our experiments. The results from the five-class classification demonstrated that X-RCRNet can be efficiently used to identify COVID-19 patients using an ECG beat signal. The experimental results and the benchmarking confirmed that the model is superior to the state-of-the-art models. In addition to the overall performance, X-RCRNet is the first known Deep Neural Network to confirm the ST interval as

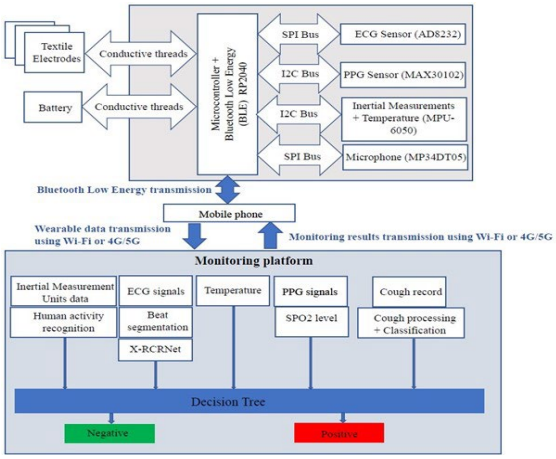


Figure 4: Proposed overall system

the most prominent feature affect by the COVID-19. Despite its great performances, there are some limitations: 1) Since our experiments were conducted using synthetic data, another cross-validation with experts is recommended to further confirm of results; 2) While relevant, using the ECG beat signal alone is not sufficient to monitor the patient symptoms. The introduction of other relevant parameters (Body temperature, SPO2 level, cough, and human activity recognition) will be performed, to improve the monitoring ability of the overall proposed system, describes

References

1. *WHO Coronavirus (COVID-19) Dashboard*. 2022; Available from: https://covid19.who.int/.
2. Romero, J., et al., *T-wave inversion as a manifestation of COVID-19 infection: a case series.* Journal of Interventional Cardiac Electrophysiology, 2020. **59**(3): p. 485-493.
3. *Coronavirus disease (COVID-19) - Symptoms*. 2022; Available from: https://www.who.int/health-topics/coronavirus#tab=tab_3.
4. Long, B., et al., *Electrocardiographic manifestations of COVID-19.* The American Journal of Emergency Medicine, 2021. **41**: p. 96-103.
5. Barman, H.A., et al., *The effect of the severity COVID-19 infection on electrocardiography.* The American Journal of Emergency Medicine, 2021. **46**: p. 317-322.
6. Dhadge, A. and G. Tilekar. *Severity Monitoring Device for COVID-19 Positive Patients*. in *2020 3rd International Conference on Control and Robots (ICCR)*. 2020.
7. Karimian, P. and M.A. Delavar, *Comparative Study of Clinical Symptoms, Laboratory Results and Imaging Features of Coronavirus and Influenza Virus, Including Similarities and Differences of Their Pathogenesis.* Pakistan Journal of Medical & Health Sciences, 2020. **14**(3): p. 1405-1411.
8. Khan, A.H., M. Hussain, and M.K. Malik, *ECG Images dataset of Cardiac and COVID-19 Patients.* Data Brief, 2021. **34**: p. 106762.
9. Makowski, D., et al., *NeuroKit2: A Python toolbox for neurophysiological signal processing.* Behavior research methods, 2021. **53**(4): p. 1689-1696.
10. P, S. and D. Wahidabanu, *Robust R Peak and QRS detection in Electrocardiogram using Wavelet Transform.* International Journal of Advanced Computer Sciences and Applications, 2011. **1**.
11. Um, T.T., et al. *Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks*. in *Proceedings of the 19th ACM international conference on multimodal interaction*. 2017.
12. Ozdemir, M.A., G.D. Ozdemir, and O. Guren, *Classification of COVID-19 electrocardiograms by using hexaxial feature mapping and deep learning.* BMC Medical Informatics and Decision Making, 2021. **21**(1): p. 1-20.
13. Attallah, O., *ECG-BiCoNet: An ECG-based pipeline for COVID-19 diagnosis using Bi-Layers of deep features integration.* Computers in Biology and Medicine, 2022. **142**: p. 105210.
14. Rahman, T., et al., *COV-ECGNET: COVID-19 detection using ECG trace images with deep convolutional neural network.* Health Information Science and Systems, 2022. **10**(1): p. 1-16.
15. Chattopadhay, A., et al. *Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks*. in *2018 IEEE winter conference on applications of computer vision (WACV)*. 2018. IEEE.

Liste des auteurs

Karima ALIOUA
Ali AMAD
Anis AMZIANE
Marc ANTONINI
Bouzid AREZKI
Patrick BAS
Mathis BAUBRIAUD
Ioan Marius BILASCO
Adrian BORS
Samuel BRAU
Pascal BRUNIAUX
Jan BUTORA
Marco CAGNAZZO
Rene CHALON
Christophe CHARRIER
Marc CHAUMONT
Patrick CORLAY
François-Xavier COUDOUX
Stephane COULOMBE
Mohamed DAOUDI
Stephane DERRODE
Thu Ha DO
Touradj EBRAHIMI
Mireille EL-ASSAL
Jorge ENCINAS RAMOS
Kevin FEGHOUL
Fangchen FENG
Sid Ahmed FEZZA
Christophe FIORIO
Guillaume FOURRET
Matthieu GENDRIN
Emmanuel GIGUET
Wassim HAMIDOUCHE
Félix HENRY
Minh Chau HUYNH
Bianca JANSEN VAN RENSBURG
Hugo JEAN
Marc JUNIOR NKENGUE
Hind KANJ
K. KERNN
Michel KIEFFER
Ludovic KOEHL
Marc LAMBERT
Mohamed-Chaker LARABI
Davi LAZZAROTTO
Trung Hieu LE
Smitha LINGADAHALLI RAVI
Olivier LOSSON
Ludovic MACAIRE
Benjamin MATHON
Anissa MOKRAOUI
Luce MORIN
Guillaume PICAUD
Mathieu PONT
William PUECH
Pauline PUTEAUX
Erwan REINDERS
Souheib RIACHE
Christophe ROSENBERGER
Deise SANTANA MAIA
Gérard SUBSOL
Xuyuan TAO
Ahmed TELILI
Luc TEOT
Paul TESSE
Michela TESTOLINA
Julien TIERNY
Pierre TIRILLY
Kim Phuc TRA
Anthony TRIOUX
Jules VIDAL
Xianyi ZENG
Yujing ZHANG